Booming Up the Long Tails: Discovering Potentially Contributive Users in Community-Based Question Answering Services

Juyup Sung and **Jae-Gil Lee*** and **Uichin Lee**
Department of Knowledge Service Engineering
Korea Advanced Institute of Science and Technology (KAIST)
291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea
{ juyup.sung, jaegil, uclee }@kaist.ac.kr

Abstract

Community-based question answering (CQA) services such as Yahoo! Answers have been widely used by Internet users to get the answers for their inquiries. The CQA services totally rely on the contributions by the users. However, it is known that newcomers are prone to lose their interests and leave the communities. Thus, finding expert users in an early phase when they are still active is essential to improve the chances of motivating them to contribute to the communities further. In this paper, we propose a novel approach to discovering "potentially" contributive users from *recently-joined users* in CQA services. The likelihood of becoming a contributive user is defined by the user's *expertise* as well as *availability*, which we call the *answer affordance*. The main technical difficulty lies in the fact that such recently-joined users do not have abundant information accumulated for many years. We utilize a user's productive vocabulary to mitigate the lack of available information since the vocabulary is the most fundamental element that reveals his/her knowledge. Extensive experiments were conducted with a huge data set of Naver Knowledge-In (KiN), which is the dominating CQA service in Korea. We demonstrate that the top rankers selected by the answer affordance outperformed those by KiN in terms of the amount of answering activity.

Introduction

As the amount of information on the Web has grown dramatically over the years, users are often frustrated with the vast quantity of the results returned by web search engines. Even worse, these results may contain many irrelevant and/or poorly written documents (Suryanto et al. 2009). Users typically have to browse a long list of the search results to look for a satisfactory answer. Thus, question answering (QA) services emerged to solve this information overload problem. QA services such as IBM Watson and Apple Siri aim at returning precise answers to natural language questions in an automated way. However, in the technical point of view, understanding such natural language questions completely is still far from perfection.

Community-based question answering (CQA) services have been widely used for bypassing the technical barriers to fully automated services. Instead, CQA services rely on the participation of many people. Through CQA services, people ask questions and obtain answers either by waiting other users to provide the answers or by searching for already answered similar questions. A knowledge base is built up as the question-answer pairs are accumulated through user contribution. Compared with search engines, CQA services can offer personalized information, recommendations, and advice from humans (Budalakoti and Barber 2010). Moreover, a questioner can expect newly updated information directly from humans. Many web portals are now offering CQA services—for example, Yahoo! Answers (YA) and Naver[1] Knowledge-In (KiN).

Despite the above-mentioned advantages, CQA services also have disadvantages. After asking a question, the questioner must wait for an answer. Also, since the question is open to all users, there is no guarantee that a potential answerer is well-qualified to answer the question. To make matters worse, a large proportion of questions remain unanswered—from 5% to 53% depending on the type of QA services (Dearman and Truong 2010). The main reason for unanswered questions is a high level of dependence on *heavy users* (Twedt 1964) who have written answers much more than the average. Most contributions are made by a small number of heavy users, and some questions are possibly not responded when they are unavailable. Because the heavy users have engaged in CQA services for a long time, they are well-qualified to and familiar with the services.

In contrast, *light users*, who recently joined and have written answers less than the average, are prone to leave CQA services because they are *not* deeply tied to the services. As a result, light users are mostly newcomers who are not acquainted with the services. Nam et al. (Nam, Ackerman, and Adamic 2009) found that the answering activity of sampled users significantly dropped after the first week. Low usability and time-consuming factors were the two main reasons (Brandtzæg and Heim 2008). Furthermore, people will not stay motivated unless they think that they are necessary for the communities (Kraut and Resnick 2008). Thus, *motivating light users* to stay in the communities—e.g., routing proper questions to them so that they can easily contribute—is of prime importance towards the success of the services.

*Jae-Gil Lee is the corresponding author.

[1]Naver (http://www.naver.com/) is the most popular web portal in Korea.

Here, it is more cost-efficient and effective to select those light users who will be more likely to contribute to the communities in the future if they are taken care of.

In this paper, we propose a novel approach to discovering "potentially" contributive users *among light users* in CQA services. We contend that the likelihood of becoming a contributive (i.e., heavy) user is dependent on the user's *expertise* and *availability*. The more expert a user is in an area and the more active the user is recently, the more questions that user can answer. Consequently, the first goal is to precisely measure the expertise of a light user in a given area. This problem is technically challenging *owing to the lack of available data* to judge the expertise of a light user. Table 1 defines four types of answerers depending on the activity period and expertise. Previous studies on expert finding and question routing are shown to successfully find the type C but has limitations in finding the type D, whereas our study focuses on the type D. That is, the main difference between previous studies and ours is the amount of information available. Therefore, our problem can be considered as a *cold-start problem*.

Table 1: Four types of answerers' answering patterns.

Type	Activity Period	Expertise
A	Long	Low
B	Short	Low
C	Long	High
D	Short	High

We attempt to measure the expertise of each user by looking into his/her *vocabulary* for the following two reasons. First, a person's productive vocabulary reveals his/her knowledge (Marjorie and Sima 1996). For example, for a question asking the advantages of a multi-core processor, the experts are likely to mention "parallel processing" whereas non-experts just say "fast." That is, some difficult or specialized words are only used by experts, though the experts use common or trivial words as well. Second, vocabulary has sharable characteristics so that domain-specific words are repeatedly used by expert answerers (White, Dumais, and Teevan 2009). Although answers were made at different times, these answers share some domain-specific words if the questions address a similar issue.

Considering these two properties, our key idea is to bridge the gap between the abundant data of heavy users and the insufficient data of light users. Our approach consists of four major steps. First, the expert score of a heavy user is calculated using the abundant data. Second, the level of a word is determined by the expert scores of the heavy users who used the word before. Third, these word levels are propagated to a set of words used by a light user in his/her answers. Fourth, the expert score of the light user is reversely calculated based on his/her vocabulary. Our approach decomposing an answer into words is reliable even for a small number of answers, because each answer typically has quite a few words. In this way, we exploit the vocabulary to compensate for the lack of a light user's data.

In summary, the contributions of this paper are as follows.

- We propose a novel approach to measuring the expertise of a light user based on his/her vocabulary. Our approach is shown to precisely predict the expertise of a user when there is no sufficient amount of data.
- Using the approach for expertise prediction, we develop a methodology of discovering "potentially" contributive (i.e., heavy) users among light users. It is observed that the users selected by our methodology dominated those selected by the service provider in terms of the amount of answering activity.
- Our methodology was verified by extensive experiments using huge amounts of question-answer data for ten years (2002∼2012).

The rest of this paper is organized as follows. The 2nd section explains the preliminaries of CQA services. The 3rd section defines our problem and gives an overview of our methodology. The 4th section explains our methodology in more detail. The 5th section describes the evaluation results. The 6th section summarizes related work. Finally, the 7th section concludes this study.

Preliminaries

In this section, we introduce the common features of CQA services, which are used in this paper.

CQA services maintain a hierarchy of categories so that users can easily ask questions and find answers. Figure 1 shows the top-level categories of Yahoo! Answers. Our methodology defines the expertise of a user on a top-level category. The more concentrated a user's answers are on a category, the more expert the user is on the category (Adamic et al. 2008). We call the top-level category in consideration as the *target category*.

Figure 1: The top-level categories of Yahoo! Answers.

CQA services maintain the statistics of each user as shown in Figure 2. Our methodology considers also these statistics in measuring the expertise of a user. The features used throughout this paper are summarized as follows.

- The *selection count* of a user is the count of his/her answers selected as the best answer by the questioner, i.e., A in Figure 2.
- The *selection ratio* of a user is the ratio of his/her selection count to the total number of his/her answers, i.e., $B = A/D$.
- The *recommendation count* of a user is the count of the recommendations made by other users when they feel his/her question or answer is interesting, i.e., C.

Figure 2: The statistics of a user in Yahoo! Answers.

Overview

Problem Definition

In this paper, we develop a methodology of measuring the likelihood of a light user becoming a contributive (i.e., heavy) user in the future in CQA services. Because there is not enough information for light users, our main research interest is to study how to exploit the abundant data accumulated by heavy users for a long time to predict the expertise (as a part of the likelihood) of light users. Our solution is to use the vocabulary to bridge a gap between heavy users and light users. Hereafter, $\mathcal{U}_\mathcal{H}$ denotes a set of heavy users, and $\mathcal{U}_\mathcal{L}$ a set of light users.

The input of our methodology is composed of the three pieces of information as shown below. Here, using the first and second pieces of information, we predict the expertise of each light user in the third piece of information.

1. The statistics of $\mathcal{U}_\mathcal{H}$: the statistics consist of the selection count, the selection ratio, and the recommendation count of a heavy user.
2. The sets of the answers written by $\mathcal{U}_\mathcal{H}$: the information of an answer includes (i) the time when the answer was written, (ii) the category where the corresponding question was posted, and (iii) the full text of the answer.
3. The sets of the answers written by $\mathcal{U}_\mathcal{L}$: each answer carries the time, category, and full text as above.

As the output, our methodology produces the likelihood of a light user becoming a heavy user in the future. We call this likelihood as the *answer affordance*. More specifically, for a set $\mathcal{U}_\mathcal{L}$ of light users and the target category c_{target}, the answer affordance, which is denoted by $Affordance(u_l)$, is calculated for $\forall u_l \in \mathcal{U}_\mathcal{L}$. We typically rank all u_l's in the decreasing order of $Affordance(u_l)$ and return the k most promising users. The service provider can encourage these users to answer more questions by routing proper questions to them.

The heavy and light users are informally defined in this paper. The heavy users are those whose answers are very abundant enough to judge their expertise. In contrast, the light users are those whose answers are insufficient. There is no clear cut point between the heavy and light users, and it really depends on the application and data set.

Overall Procedure

In this subsection, we present the overall procedure with the rationale behind our design.

The answer affordance of a user takes *expertise* as well as *availability* into consideration and should have the properties in Property 1. Experts, of course, will be able to answer more questions than non-experts, and giving an encouragement will be more effective to currently active users than inactive users.

Property 1. The *answer affordance* of a user becomes higher as the user is more expert on the target category and as the user's activity gets closer to the present.

Measurement of Expertise The first component, *expertise*, is measured through the four major steps in Figure 3.

Step 1: For a set $\mathcal{U}_\mathcal{H}$ of heavy users, the *expertise*, denoted by $Expertise(u_h)$, is calculated for $\forall u_h \in \mathcal{U}_\mathcal{H}$. The expertise should have the properties in Property 2. As for the first property, Adamic et al. (Adamic et al. 2008) found that higher concentration (lower entropy) correlates with receiving higher answer ratings. In addition, the second property reflects the evaluations by other users.

Property 2. The *expertise* of a user becomes higher (i) as the user's answers are more concentrated on the target category and (ii) as the user has higher selection count, selection ratio, and recommendation count.

Please note that a technique of measuring the expertise of heavy users is orthogonal to our methodology. That is, we can adopt any technique that quantifies expertise. The most important contribution of this paper is the concept of *propagating the expertise* from heavy users to light users through the vocabulary.

Step 2: For a set of $\mathcal{W}$ of usable words, the *word level*, denoted by $WordLevel(w_i)$, is calculated for $\forall w_i \in \mathcal{W}$. The word level should have the property in Property 3. It implies that a user's knowledge is reflected in his/her vocabulary (Marjorie and Sima 1996).

Property 3. The *word level* of a word becomes higher as the word is used by more expert users and more frequently.

Step 3: The word levels derived from $\mathcal{U}_\mathcal{H}$ are propagated to $\mathcal{U}_\mathcal{L}$. This step, in fact, is fitting abundant historical data into the present data in considering that $\mathcal{U}_\mathcal{H}$ and $\mathcal{U}_\mathcal{L}$ are divided on the timeline. Heavy users usually have been members for a long time, whereas light users are mostly newcomers. This step is supported by the observation that the vocabulary of an expert stays mostly unchanged despite a temporal gap (White, Dumais, and Teevan 2009).

Step 4: For a set $\mathcal{U}_\mathcal{L}$ of light users, their expertise is reversely calculated for $\forall u_l \in \mathcal{U}_\mathcal{L}$ using the word levels. Please note that this value is an *estimation* of the value in Step 1 since the expertise of heavy users permeated the word level. This value for light users is denoted by $EstimatedExpertise(u_l)$ to distinguish the expertise of light users from that of heavy users.

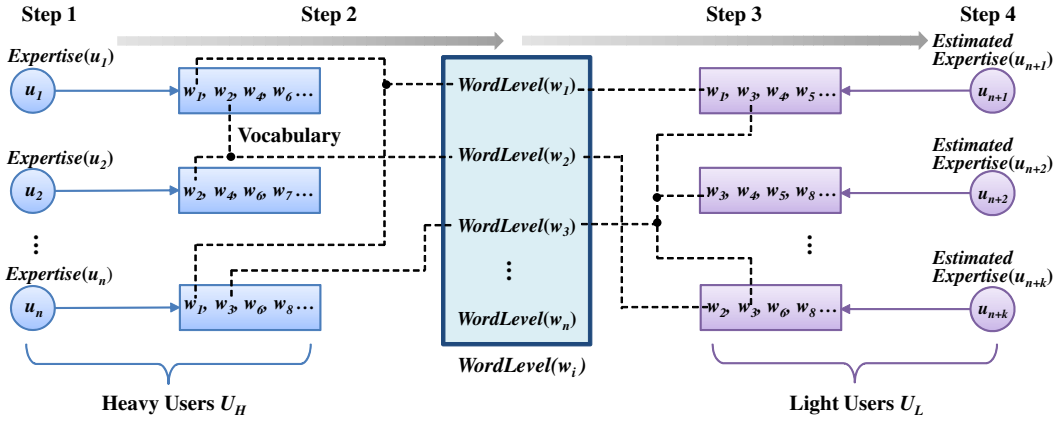


Figure 3: The overall procedure of predicting the expertise of a light user.

Measurement of Availability The second component, *availability*, is simply the number of a user's answers with their importance proportional to their recency. Figure 4 intuitively describes this concept: each bar represents an answer, and the height indicates its importance. Thus, the availability of a light user u_l, denoted by $Availability(u_l)$, should have the property in Property 4.

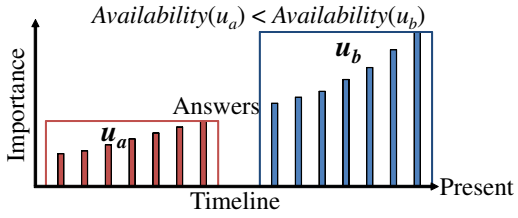


Figure 4: An intuitive illustration of availability.

Property 4. The *availability* of a user becomes higher as the user wrote more answers at the time closer to the present.

Detailed Methodology

In this section, we formalize our methodology informally presented in the previous section. Table 2 lists the measures that will be formally defined in this section.

Table 2: The notation used throughout this paper.

Notation	Description
$Affordance(u_l)$	the answer affordance of a light user
$Expertise(u_h)$	the expertise of a heavy user
$EstimatedExpertise(u_l)$	the estimation of the expertise of a light user
$WordLevel(w_i)$	the word level of a usable word
$Availability(u_l)$	the availability of a light user

Answer Affordance

The answer affordance of a light user u_l is defined as Eq. (1), reflecting Property 1. Here, w_a $(0 \leq w_a \leq 1)$ is a balancing factor, and its suitable value depends on the context and data set. $EstimatedExpertise(u_l)$ is defined in the subsection "Expertise," and $Availability(u_l)$ in the subsection "Availability."

$$Affordance(u_l) = w_a \cdot EstimatedExpertise(u_l) + (1 - w_a) \cdot Availability(u_l) \quad (1)$$

Expertise

As shown in Figure 3, $EstimatedExpertise(u_l)$ is obtained through four steps. All measures defined here will be made to have the range $[0, 1]$ by min-max normalization unless a measure inherently has that range.

Step 1: The expertise of a heavy user u_h is defined through Eqs. (2)∼(4), reflecting Property 2. Here, w_e $(0 \leq w_e \leq 1)$ is a balancing factor.

$$Expertise(u_h) = w_e \cdot Entropy(u_h) + (1 - w_e) \cdot Rating(u_h) \quad (2)$$

First, $Entropy(u_h)$ is Eq. (3), which is basically the entropy of a user's interests. The more concentrated a user's answers are, the lower the entropy is. Here, we assume that there are only two categories—the target category and one including all others—since we are not interested in other than the target category. The entropy $E(u_h)$ starts to decrease at $P(c|u_h) = 0.5$. Hence, the entropy curve for $P(c|u_h) > 0.5$ is flipped vertically such that $Entropy(u_h)$ monotonically increases as $P(c|u_h)$ increases.

$$Entropy(u_h) = \begin{cases} \dfrac{(1 + (1 - E(u_h)))}{2} & \text{if } P(c_{target}|u_h) > 0.5 \\ \dfrac{E(u_h)}{2} & \text{otherwise} \end{cases} \quad (3)$$

where

$$E(u_h) = \sum_{c \in \{c_{target}, c_{others}\}} -P(c|u_h) \cdot log_2 P(c|u_h)$$

and

$$P(c|u_h) = \frac{\text{the number of } u_h\text{'s answers in the category } c}{\text{the total number of } u_h\text{'s answers}}$$

Second, $Rating(u_h)$ is Eq. (4), which represents the evaluations by other users as well as the *absolute* amount of user activity. The entropy alone is very sensitive to the number of answers. For example, if a user responded to only one question and it is under the target category, his/her entropy will be the lowest. This is the reason that we need to include a kind of absolute values.

We consider the selection count and the recommendation count equally important. These two features follow the power-law distribution. That is, too high values of one feature may override other features of expertise. Besides, the gap between the majority of small values is negligible. Thus, we use a sigmoid function $S(t)$ to smooth the high values of the two features. Here, α is a scaling factor.

$$Rating(u_h) = \frac{S(SelCount(u_h)) + S(RecommCount(u_h))}{2} \tag{4}$$

where

$$S(t) = \frac{1}{1+e^{-t\cdot\alpha}},$$

$$SelCount(u_h) = \text{the selection count of } u_h,$$

and

$$RecommCount(u_h) = \text{the recommendation count of } u_h$$

Step 2: $WordLevel(w_i)$ is Eq. (5), reflecting Property 3. It is the weighted sum of $Expertise()$'s of the heavy users who used the word w_i. The weight is given by the tf-idf (term frequency-inverse document frequency) score of a word, which is very popular in information retrieval. In calculating a tf-idf score, an answer corresponds to a document, and all answers of a heavy user to a collection of documents.

$$WordLevel(w_i) = \frac{\sum_{u_h \in \mathcal{U}_\mathcal{H}} \sum_{a_j \in A_{u_h}} Expertise(u_h)\, \delta(w_i, a_j)\, tfidf(w_i, a_j, A_{u_h})}{\sum_{u_h \in \mathcal{U}_\mathcal{H}} \sum_{a_j \in A_{u_h}} \delta(w_i, a_j)\, tfidf(w_i, a_j, A_{u_h})} \tag{5}$$

where

$$A_{u_h} = \text{the set of the answers from } u_h,$$

$$\delta(w_i, a_j) = \begin{cases} 1 & \text{if } w_i \text{ is used in } a_j \in A_{u_h} \\ 0 & \text{otherwise} \end{cases},$$

and

$$tfidf(w_i, a_j, A_{u_h}) = tf(w_i, a_j) \times idf(w_i, A_{u_h})$$

Steps 3 & 4: $EstimatedExpertise(u_l)$ is Eq. (6), which is basically the average of $WordLevel()$'s of the words used by the light user u_l in his/her answers. Here, we ignore insignificant words whose word level is less than 0.1 to avoid his/her expertise being blurred by them.

$$EstimatedExpertise(u_l) = \frac{\sum_{w_i \in W_{u_l}} WordLevel(w_i) \cdot \rho(w_i)}{\sum_{w_i \in W_{u_l}} \rho(w_i)} \tag{6}$$

where

$$W_{u_l} = \bigcup_{a_j \in A_{u_l}} \{\text{the usable words in an answer } a_j\}$$

and

$$\rho(w_i) = \begin{cases} 1 & \text{if } WordLevel(w_i) \geq 0.1 \\ 0 & otherwise \end{cases}$$

Availability

$Availability(u_l)$ is defined by Eq. (7), reflecting Property 4. $Recency(u_l)$ is inversely proportional to the ages of the answers posted by the light user u_l. For example, an answer posted today adds $\frac{1}{2}$, and one posted yesterday adds $\frac{1}{3}$. A sigmoid function is applied to $Recency(u_l)$ to smooth high values. Min-max normalization will be applied to $Availability(u_l)$ to make its range $[0, 1]$.

$$Availability(u_l) = S(Recency(u_l)) \tag{7}$$

where

$$Recency(u_l) = \sum_{a_j \in A_{u_l}} \frac{1}{Age(a_j) + 2}$$

and

$$Age(a_j) = \text{the number of days since } a_j \text{ was posted}$$

Evaluation

In this section, we describe the evaluation results of our methodology through two subsections: first for the *expertise* in the subsection "Accuracy of Expertise Prediction" and then for the *answer affordance* in the subsection "Effectiveness of Answer Affordance."

Experiment Setting

Data Description The data set for the experiments was collected from Naver Knowledge-In (KiN)[2]. The URL of its home page is `http://kin.naver.com`. The period of the data set crawled is from September 2002 to August 2012, which comprises a period of **ten** years.

We conducted two sets of experiments by choosing different target categories: *computers* and *travel*. Please note that the computers category typically deals with factual information whereas the travel category has subjective opinions. This difference is significant in our methodology since the expertise is based on the entropy of a user. As Adamic et al. (Adamic et al. 2008) pointed out, answer quality is higher as the entropy is lower, especially for the categories where factual expertise is primarily sought after.

[2]KiN started its service in early 2002, and the number of registered questions in KiN is over 0.1 billion as of October 2012.

We crawled only the question-answer pairs whose answer was selected by the questioner. Unselected answers often have poor quality or may include spam, advertisements, or jokes from abusers. There were about 11 million registered questions in the computers category as of October 2012, and we obtained 3,926,794 question-answer pairs with a selected answer. For the travel category, we gathered 585,316 question-answer pairs from 1.4 million questions.

Since our methodology exploits the vocabulary, it is required to refine the set of words used in the answers. We corrected spelling as well as spacing using publicly available software. In addition, we removed the words that appeared only once in the data set since those words are useless in our methodology. The total number of usable words is 422,400; the number of usable words is 191,502 in the computers category and 232,076 in the travel category.

During the period of ten years, the number of the users who answered at least once is 465,711 in the computers category and 106,194 in the travel category. However, we did not use all of them since the users with few answers could bring the possibility of distorted results owing to the characteristics of the entropy. Thus, we filtered out the users who answered less than five times in each category. As a result, we obtained 228,369 users in the computers category and 44,866 users in the travel category.

Table 3 shows the statistics of the data set for the two categories *after preprocessing*.

Table 3: The statistics of the data set.

	Computers	Travel
# of answers	3,926,794	585,316
# of words	191,502	232,076
# of users	228,369	44,866

Period Division We divided the entire period into three periods—the *resource*, *training*, and *test* periods—as shown in Figure 5 to separate the users. In the experiments, the set $\mathcal{U}_{\mathcal{H}}$ of heavy users were those who joined during the resource period. Because the resource period is very long (almost seven years), many users in that period responded to many questions, though some users did not. On the other hand, the set $\mathcal{U}_{\mathcal{L}}$ of light users were those who joined during the training period. Because the training period is short (only one year), many users in that period did not have enough answers for us to judge their expertise.

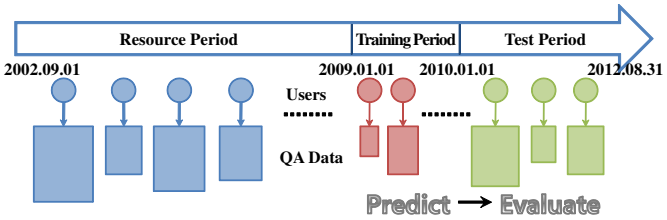


Figure 5: Division of the entire period into three periods.

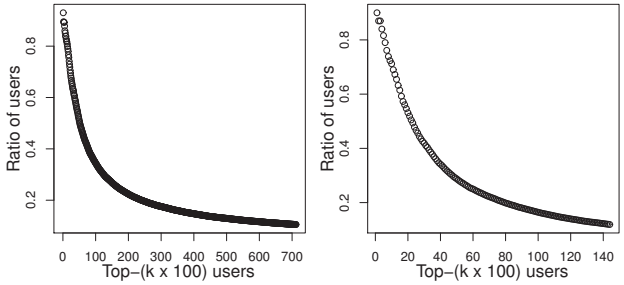
Please note that we assume that the present time is the end of the training period. The test period is assumed to be the future and used for evaluating the effectiveness of our methodology.

Parameter Configuration Our methodology has two balancing factors w_a in Eq. (1) and w_e in Eq. (2) as well as a scaling factor α in the sigmoid function. The proper values for these parameters depend on the context and data set. Thus, these parameter values were configured empirically. w_a and w_e were set to be 0.6 and 0.3, respectively. α was set to be 0.01 in Eq. (4) and 0.1 in Eq. (7).

Accuracy of Expertise Prediction

In this subsection, we show that $EstimatedExpertise()$ precisely predicts the expertise of a light user.

Preliminary Test In KiN, the users are allowed to designate their main interest on their own in order to expose their expertise on a specific subject. Using this information, we examined the ratio of the *light users* who selected the target category as their main interest. Overall, the ratio was 10.5% in the computers category and 24.5% in the travel category. That is, the ratio was low. However, the ratio increases significantly if we confine this test to expert users. To this end, we sorted the light users in the decreasing order of $EstimatedExpertise()$ for each category. Figure 6 shows the ratio of such *self-declared* experts in the top-k users. When k was set to be 100, 93% of the light users in the computers category set their main interest to the target category, and 90% of those in the travel category did. The top users (i.e., experts) more tend to choose the target category as their main interest. This result implies that our approach accurately measures the expertise of a light user.



(a) Computers category. (b) Travel category.

Figure 6: The ratio of users who expressed their interests.

Evaluation Method We sorted the light users in the decreasing order of $EstimatedExpertise()$ and obtained a set $E_{training}$ of the top-k users. In addition, we obtained another set E_{test} of the top-k users *from the test period* by KiN's ranking scheme and regarded E_{test} as the ground truth. We compared the users in $E_{training}$ with those in E_{test}. This comparison shows how many "potential" experts selected by our approach will become "real" experts in the near future.

There is no definite way to obtain the ground truth for the expertise in CQA services. We decided to imitate KiN's ranking scheme for this purpose since Naver must have

elaborated the scheme. KiN's ranking scheme is using a weighted sum of the *selection count* and the *selection ratio*. The detailed explanation is omitted here because of the lack of space. We believe that KiN's scheme should work well only if we have a sufficient amount of information. One might think that we could use KiN's scheme also for ranking the light users in the training period. However, the user profiles such as the selection count and the selection ratio used in KiN's scheme are not very reliable when there is no sufficient information, as will be shown later. The test period is long enough (almost three years) to apply KiN's scheme to the users in that period.

$EstimatedExpertise()$ is the approach used in our methodology for generating the set $E_{training}$. In addition to $EstimatedExpertise()$, we generated the sets of potential experts by using three different alternatives. That is, we sorted the light users in the training period in the decreasing order of the following values. We normalized all these values by using the number of the answers of the corresponding user during the test period.

- $Expertise()$: the way of ranking heavy users rather than light users in our methodology
- $SelCount()$: the selection count
- $RecommCount()$: the recommendation count

We used two metrics, P@k and R-precision, to measure the precision of prediction. The metrics are popular in information retrieval and natural language processing. Precision is more important in our evaluation than recall since we are considering only light users.

- *P@k*: the precision evaluated at a given cut-off rank, considering only the topmost k results
- *R-precision*: the precision at R-th position in the ranking of results for a query that has R relevant documents

Evaluation Results Tables 4 and 5 show the results of the precision performance. $EstimatedExpertise()$ is what we propose in our methodology. The number of experts selected was 200 in the computers category and 100 in the travel category. In Table 4, $EstimatedExpertise()$ is shown to outperform other alternatives significantly (by up to about 3 times) for the computers category, meaning that the other alternatives based on user profiles are *not* robust when there is no sufficient information.

In contrast, in Table 5, our approach is not superior to the other alternatives, though its performance is reasonably good. The main reason is that our approach exploits the entropy of a user, and this result conforms to the findings by Adamic et al. (Adamic et al. 2008). If we used another technique for calculating $Expertise()$, we would be able to make our approach outperform others for subjective opinions. We leave it as a topic of future research.

Test with Very Small Data Although we chose a one-year period to obtain a sufficient number of light users, it is worthwhile to further reduce the amount of available information. With this motivation in mind, we selected the users who answered to more than 30 questions in the computers category (1,043 users) and 15 questions in the travel category (313 users). Then, we increased the size of the question-answer set for each user from 5 to 30 (or 15), which was used for estimating his/her expertise. The correlation coefficient between the estimated expert rankings and the ground truth is shown in Figure 7. We found out that the correlation coefficient was sufficiently high (about 0.7) even when only five answers were used. This result implies that our approach indeed mitigates the cold-start problem.

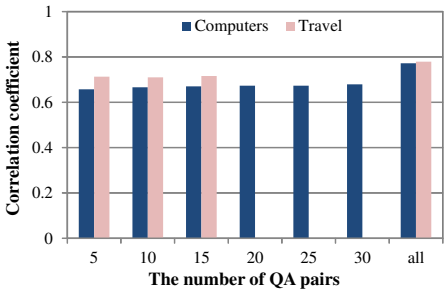


Figure 7: The performance with very small data.

Effectiveness of Answer Affordance

Finally, in this subsection, we show that the light users selected by $Affordance()$ are likely to become contributive users in the future, proving the primary goal of this paper.

Evaluation Method We sorted the light users in the decreasing order of $Affordance()$ and obtained a set E_{ours} of the top-k users. Because KiN manages the rankings of expert users for each category, we could obtain another set E_{KiN} of the top-k users based on the rankings provided by KiN. Since KiN provides only the *current* rankings, we need to make the time frame comparable. Thus, we calculated $Affordance()$ using the data for the past one year from the time of doing this experiment, i.e., from July 2011 to July 2012. k was set to be 200 in the travel category and doubled for the computers category owing to its larger size.

Definitions 1 and 2 introduce the metrics used for this experiment. These two metrics were captured for the two sets E_{ours} and E_{KiN} during a four-week period.

Definition 1. The *user availability* of a set $\mathcal{U}_{\mathcal{E}}$ of users on a given day is defined as the ratio of the number of the users in $\mathcal{U}_{\mathcal{E}}$ who appeared on the day to the total number of users who appeared on that day. A user is regarded to appear on the day if he/she left at least one answer.

Definition 2. The *answer possession* of a set $\mathcal{U}_{\mathcal{E}}$ of users on a given day is defined as the ratio of the number of the answers posted by the users in $\mathcal{U}_{\mathcal{E}}$ on the day to the total number of answers posted on that day.

The answering activity of the users selected by our methodology diminishes quickly unless they are encouraged to stay (Nam, Ackerman, and Adamic 2009). However, we are not able to encourage such users since we are not the service provider. That is, the effectiveness of the answer affordance cannot be verified exactly as what we suggested. Instead, we decided to update the set E_{ours} every week to *simulate* motivation and encouragement. Thus, we recalculated

Table 4: The precision performance for the computers category.

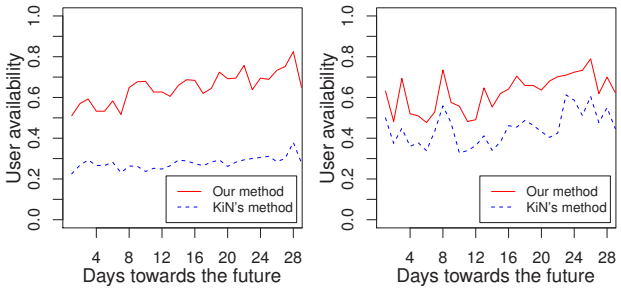
Target	Method	P@20	P@40	P@80	P@120	P@160	R-Precision
Only Light Users (# of Experts: 200)	$EstimatedExpertise()$	0.833	**0.833**	**0.748**	**0.600**	**0.413**	**0.600**
	$Expertise()$	**0.870**	0.606	0.396	0.229	0.129	0.390
	$SelCount()$	0.800	0.714	0.625	0.558	0.366	0.570
	$RecommCount()$	0.741	0.690	0.556	0.386	0.205	0.475

Table 5: The precision performance for the travel category.

Target	Method	P@10	P@20	P@40	P@60	P@80	R-Precision
Only Light Users (# of Experts: 100)	$EstimatedExpertise()$	0.833	0.870	**0.889**	0.789	0.702	0.730
	$Expertise()$	**0.909**	0.833	0.833	0.674	0.398	0.620
	$SelCount()$	0.833	**0.909**	**0.889**	**0.800**	**0.741**	**0.760**
	$RecommCount()$	**0.909**	0.870	0.816	0.698	0.530	0.650

$Affordance()$ with the one-year period shifted towards the present by one week. Meanwhile, the rankings managed by KiN respect the users who have shown steady activity for many years and built their valuable identity. For a fair comparison, KiN's rankings were also updated whenever the answer affordance was recalculated.

Evaluation Results Figure 8 shows the user availability of E_{ours} (our method) and E_{KiN} (KiN's method) from July 28, 2012 to August 24, 2012. The user availability was consistently much higher in our method than in KiN's method. Since KiN's method is the only one published to date, this comparison is reasonable although it does not seriously consider availability. In Figure 8(a), the average was 0.650 in our method and 0.278 in KiN's method. In Figure 8(b), the average was 0.624 in our method and 0.447 in KiN's method. Our method outperformed KiN's method by a larger margin in the computers category than in the travel category.

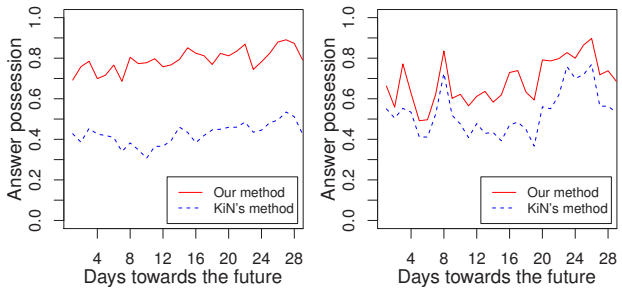


(a) Computers category. (b) Travel category.

Figure 8: The result of the user availability.

Figure 9 shows the answer possession for the same period. We observed the same trend as in Figure 8. In Figure 9(a), the average was 0.792 in our method and 0.426 in KiN's method. In Figure 9(b), the average was 0.686 in our method and 0.533 in KiN's method. The gap between the two methods was also larger in the computers category.

Overall, these results show that the light users selected by our methodology contribute to the services more than the top rankers selected by the service provider. These light users



(a) Computers category. (b) Travel category.

Figure 9: The result of the answer possession.

are expert in an area as well as eager to answer many questions at the beginning. Therefore, they are likely to become contributive users if they get motivated properly.

Related Work

Expert Finding and Question Routing

The work on expert finding and question routing is the closest to our work. Many algorithms have been proposed in the literature: for example, graph-based algorithms (Dom et al. 2003; Zhang, Ackerman, and Adamic 2007) and language-based algorithms (Liu, Croft, and Koll 2005; Zhou et al. 2009). These algorithms were shown to successfully find the experts when there is an abundant amount of information. However, most of them did not pay much attention on the cold-start problem. Pal et al. (Pal and Konstan 2010) tackled the cold-start problem with an observation that experts have a biased pattern when selecting a question to answer. Their model simply distinguishes experts from normal users, whereas ours provides a sophisticated measure indicating the degree of expertise.

In addition, many studies along this direction have neglected the availability of users as opposed to our methodology. Nevertheless, the availability is important since questions should be routed to the users who have spare time. As far as we know, the QuME system (Zhang et al. 2007) first considered the recency of answerers in computing an

expert score. Afterwards, Aardvark (Horowitz and Kamvar 2010) was developed for social search with focusing on three factors—relevance, connectedness, and availability.

Characteristics of Vocabulary Knowledge

It is widely recognized that a productive vocabulary implies the knowledge level (Henriksen 1999). That is, if one comfortably uses a domain-specific word, this means that he/she is knowledgeable on the field. The relation between vocabulary and domain expertise has been well-studied in the research on web search behaviors. Several studies addressed that domain experts typically use specialized and standard words. White et al. (White, Dumais, and Teevan 2009) found that more than 50% of words used by domain experts at search engines were included in the lexicon defined by the authors. Besides, Zhang et al. (Zhang, Anghelescu, and Yuan 2005) showed that experts not only used longer queries but also shared more thesaurus terms than the non-experts. We catch these characteristics of vocabulary in order to define the light users' expertise.

Conclusions

In this paper, we developed a novel methodology of discovering "potentially" contributive users among light users in CQA services. The notion of the answer affordance was defined to measure the likelihood of becoming a contributive user. The answer affordance considers both expertise and availability. Our key idea of measuring the expertise of a light user is to look into his/her productive vocabulary to compensate for the lack of available information. Since the vocabulary reveals his/her knowledge and has sharable characteristics, our approach is shown to be quite robust to a small amount of information.

We performed extensive experiments using large-scale real data for ten years. Our approach predicted the expertise of light users more accurately than other alternatives. More importantly, the top-k users by the answer affordance responded to more questions than those by the service provider. These results show that such light users have high potential for becoming contributive users.

Overall, we believe that we provided an interesting insight into whom the service provider should really care about. This insight can be exploited to reduce the number of unanswered questions and improve the quality of answers.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012012954).

References

Adamic, L. A.; Zhang, J.; Bakshy, E.; and Ackerman, M. S. 2008. Knowledge sharing and Yahoo Answers: Everyone knows something. In *Proc. of WWW*, 665–674.

Brandtzæg, P. B., and Heim, J. 2008. User loyalty and online communities: Why members of online communities are not faithful. In *Proc. of INTETAIN*, 1–10.

Budalakoti, S., and Barber, K. S. 2010. Authority vs affinity: Modeling user intent in expert finding. In *Proc. of IEEE CPSRT*, 371–378.

Dearman, D., and Truong, K. N. 2010. Why users of Yahoo! Answers do not answer questions. In *Proc. of ACM CHI*, 329–332.

Dom, B.; Eiron, I.; Cozzi, A.; and Zhang, Y. 2003. Graph-based ranking algorithms for e-mail expertise analysis. In *Proc. of DMKD*, 42–48.

Henriksen, B. 1999. Three dimensions of vocabulary development. *Studies in Second Language Acquisition* 21(2):303–317.

Horowitz, D., and Kamvar, S. D. 2010. The anatomy of a large-scale social search engine. In *Proc. of WWW*, 431–440.

Kraut, R. E., and Resnick, P. 2008. Encouraging contribution to online communities. In Kraut, R. E., and Resnick, P., eds., *Evidence-based social design: Mining the social sciences to build successful online communities*. MIT Press.

Liu, X.; Croft, W. B.; and Koll, M. 2005. Finding experts in community-based question-answering services. In *Proc. of ACM CIKM*, 315–316.

Marjorie, W., and Sima, P. T. 1996. Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review* 53(1):13–40.

Nam, K. K.; Ackerman, M. S.; and Adamic, L. A. 2009. Questions in, Knowledge iN?: A study of Naver's question answering community. In *Proc. of ACM CHI*, 779–788.

Pal, A., and Konstan, J. A. 2010. Expert identification in community question answering: Exploring question selection bias. In *Proc. of CIKM*, 1505–1508.

Suryanto, M. A.; Sun, A.; Lim, E.-P.; and Chiang, R. H. 2009. Quality-aware collaborative question answering: Methods and evaluation. In *Proc. of ACM WSDM*, 142–151.

Twedt, D. W. 1964. How important to marketing strategy is the "heavy user"? *Journal of Marketing* 28:71–72.

White, R. W.; Dumais, S. T.; and Teevan, J. 2009. Characterizing the influence of domain expertise on web search behavior. In *Proc. of ACM WSDM*, 132–141.

Zhang, J.; Ackerman, M. S.; and Adamic, L. 2007. Expertise networks in online communities: Structure and algorithms. In *Proc. of WWW*, 221–230.

Zhang, X.; Anghelescu, H. G.; and Yuan, X. 2005. Domain knowledge, search behaviour, and search effectiveness of engineering and science students: An exploratory study. *Information Research* 10(2):217.

Zhang, J.; Ackerman, M. S.; Adamic, L.; and Nam, K. K. 2007. QuME: A mechanism to support expertise finding in online help-seeking communities. In *Proc. of UIST*, 111–114.

Zhou, Y.; Cong, G.; Cui, B.; Jensen, C. S.; and Yao, J. 2009. Routing questions to the right users in online communities. In *Proc. of IEEE ICDE*, 700–711.