Finding Answerers on Yahoo! Answers

Venu Gopal Kasturi
Computer Science Department
Stanford University
venuk@stanford.edu

C.V.Krishnakumar Iyer
Computer Science Department
Stanford University
cvkkumar@stanford.edu

ABSTRACT

People use online forums extensively to seek information which cannot be easily found using search engines or which are subjective. Some of the categories of questions are very specialized and some of them suffer with the Free-rider problem where users ask questions but do not actively look for questions to answer. Finding the appropriate users to answer a question is one of the main challenges of Question-Answer communities and propagate the questions to a group of answerers is very important. In this work, we analyze the Yahoo! Answers media and study some interesting properties that exist in an apparently non-network structure. Further, we use K-means clustering algorithm to find out the most probable group of users who might be able to answer a question posed by a questioner using textual, structural and auxiliary information.

Keywords

Yahoo Answers, QA Systems

1. INTRODUCTION

Often people seek information/answers to non-objective queries that cannot be easily obtained by searching over Google. This information is based on the experience and thought of individuals, most of which is not transferred to the Web and is hence largely inaccessible. To overcome this problem, people resort to Online forums and Question-Answer communities like Yahoo! Answers, Sun Java Forum, Aardvark etc. Though the latter ones are more specialized forums, Yahoo! Answers is very diverse in the categories of the information posted and we can observe a diverse set of interaction patterns among the users too. Some of the challenges facing these communities is the Free-rider problem as well as the delay in getting the answers. A possible solution to these issues is to find the set of users who would be able to answer the question and propagate it to them. The kind of responses and the interaction patterns of users in Yahoo! Answers is diverse and hence we need to take these patterns into account when we are predicting the answerers. We intend to take the textual information of the question along with the attributes of the questioner obtained from the structural properties of the user's interaction graph and also auxiliary information like the quality of the answers he has given and cluster them by a k-means clustering algorithm. For a new question posed by a user, we can then route the question to answerers for the question in an obtained cluster.

2. LITERATURE SURVEY

Questions such as finding an expert in a particular topic are extremely commonplace. In recent years, the advent of the social network has resulted in the free sharing of information without any boundaries. Question-Answering, too has been a part of the human life since times immemorial. Among other things, the combination of these two factors has led to the growth of Question-Answering forums on the web. Since they are a natural mode of communication for humans, they have never quite lost their popularity. And in today's age, forums like Yahoo Answers enable us to throw questions around and wait for an answer. There are also commercial companies like Aardvark that route questions to people, based on their interest and ensures that a personal answer is received by the question asker. In fact [2] states that in some markets, this form of information seeking is dominating the web search on account of it being more natural and personal in nature.

This topic of knowledge sharing using an online media has been the subject of many researchers, though many focus on different aspects of the system. [2] deals with the task of identifying 'high quality content in community driven question/answering sites'. It also explores the exploitation of auxiliary sources of information in order to identify the best quality of content. In particular, [2] proposes a *user-question-answer* tripartite graph to model the interactions between the users. It also uses the texual structure to estimate the value of an answer to the user based on other auxiliary information.

[1] presents a detailed analysis of the characteristics of Yahoo! Answers. It defines the Poster-Replier graph to understand the interactions among the users. and study the categories by clustering them into factual,semi-factual, and discussion forums; depending on whether the answers are factual (Programming), semi-factual and discussion forums (sports, politics) etc. It also introduces the Poster-Replier Graph - a directed graph with edges from the answerer to the questioner. [1] also describes the pattern of finding signatures of the Yahoo! *communities*, something that we found

to be extremely powerful and use in our work. [1] tries to solve the problem of predicting the best answer for a question using a classification approach, based on features that are network properties.

In [4], the authors compare various heuristics to find the expertise of users in a forum and developed a modified version of the Page Rank algorithm as the Expertise Rank algorithm. They develop mathematical models to represent the network using the expertise of the users.

3. OBJECTIVE

The objective of this work is to be able to identify the set of users who will be able to answer a question posed by a user, i.e given a new question we must be able to generate a small pool of answerers who can probably answer that question. This could be used to propagate the question to those users who would potentially answer it.

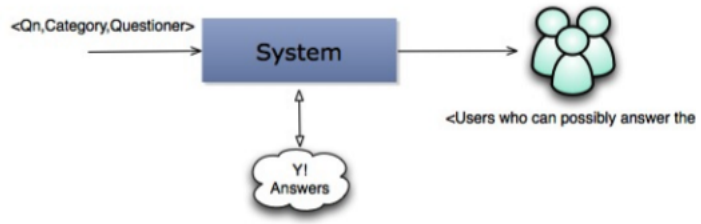

Figure 1: Objective of the System

4. DATA COLLECTION

The first step in the process of understanding of the dynamics behind the Yahoo! Answers was the collection of the data. We used the Yahoo! Answers API to collect all the features of the questions. The features that were collected are summarized below:

```
<
Qid,UserId,Category,Qn_Time,numAnswers,Subject,
Content,ChosenAnswererId,
<AnswerList>,
<AnswererTime>
<AnswererIDs>
>
```

The statistics for the data are summarized in the table below:

Number of Questions	318690
Number of Answers	2005529
Number of Question Askers	156498
Number of Answer Gives	196048

5. CHARACTERIZATION OF YAHOO! ANSWERS

Yahoo! Answers, in its primitive form is a question-answering forum, and as such does not have any explicit network structure like *Facebook* or *LinkedIn*. However, in Yahoo Answers, we have users who ask questions, say the *Questioners*, those who answer (say *answerers*) and one answer is denoted as a *Post*. The entire collection of the questions with all of its answers is called as a *Thread*.

In this section, we investigate some properties of the Yahoo! Answers network that would enable us to formulate our method better. When we studied the properties of the Yahoo! Answers, it was surprising to know that power law and the law of preferential attachment operated there. For instance, in Fig 3, we see that the curve follows a near power law distribution as is shown by Fig 4

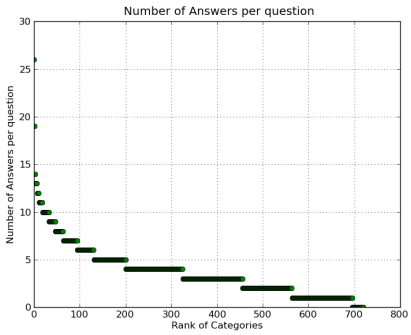

Figure 2: Answers Per Qn By Category

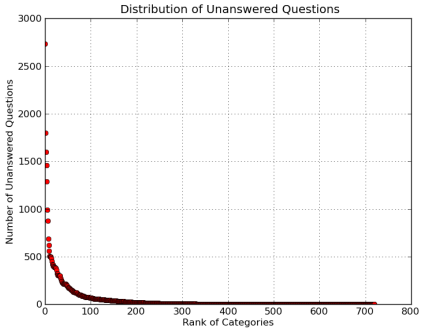

Figure 3: Unanswered Qns wrt Category

Moreover, Fig 5 and Fig 6 show that it does sometime take an huge amount of time before a question can be answered and this is usually due to the fact that the potential answerers to those questions have not been able to see them.

In [1], Lada Adamic et al, give the signature of a category to be its Thread Length, its Post Length and its Asker Replier Overlap. We plotted the categories that we had crawled on their <Thread Len , Post Len > axis. Fig 8 shows the result of a K-means clustering on this plot. Here, the categories on the lower green cluster (esp on the left) are categories where there are clear notions of experts, and they provide answers that saturates the discussion. *Programming* is an example of this category.

On the other hand, there are categories like *Politics* where people who ask questions in the category are also the people who answer, thereby giving a high asker replier overlap. These occur towards the right end of Fig 9

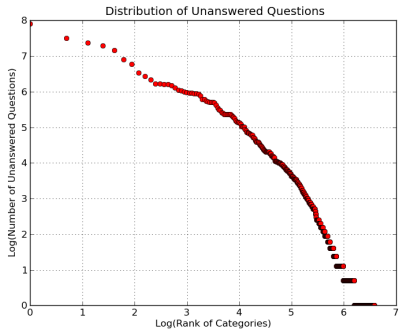

Figure 4: Unanswered Questions By Category Log Log

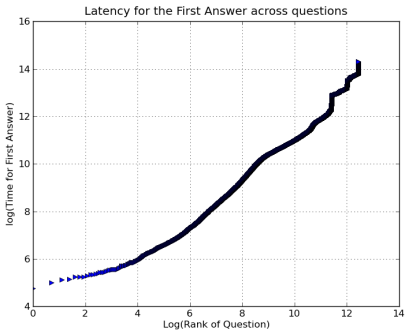

Figure 5: Delay Before the First Answer (log scale in ms)

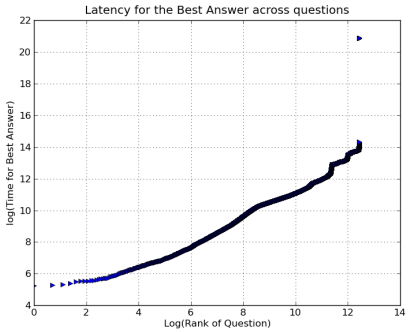

Figure 6: Delay before the Best Answer (log scale in ms)

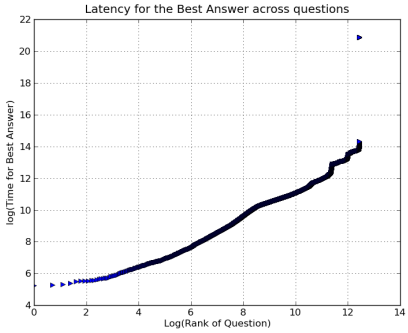

Figure 7: Delay before the Best Answer

6. ANSWERER POOL DETECTION ALGORITHM

The idea behind this algorithm is to use just the features of the question, the category and the attributes of the questioner and be able to give back a pool of eligible users who would possibly answer a new question.

We represent each question in the current set by a feature vector that is:

```
<Question Based>
--Word Length
--Number of sentences
<Category Based >
--Thread Length
--Post Length
--QuestionerâĂŞAnswerer Cosine overlap
<Questioner based>
--No. of qns asked by the user
--No. of qns asked by the user in the category
--No. of qns that were answered to him
--No. of qns that were answered to him in the category
--No. of ans given by the user
--No. of ans given by the user in the current category
--No. of best answers
--No. of best answers in the category
--Average post length
--Average question length
```

The model is built by clustering the feature vectors corresponding to the existing questions using k-means clustering algorithm and storing the list of answerers corresponding to each cluster.

As depicted in Fig 10, When a new question μ comes in to be answered, we find out the representative cluster for the feature vector and return the set of answerers associated with that cluster.

However, the number of results returned this way might be quite high. So, we adopt a method of Cascaded clustering, wherein we perform clustering inside the current cluster to zero in on the most promising set. The number of cascades can e application specific.

7. EVALUATION AND RESULTS

In this section, we show the results of the evaluation on two sets of data - one (A) having a 1000 new questions and 20K existing questions and the other (B) having a 5000 new questions and 50K existing questions.

Then the results are captured in the Fig 11 and Fig.12. Here the blue line represents the result for (A) and the red line for (B).

We observe that as we increase the cluster size the recall decreases slightly while the precision increases dramatically. Also, by adopting the method of cascaded clustering, the recall does not fall much while the precision is increased.

8. CONCLUSION AND FUTURE SCOPE

Currently, our approach of cascaded clustering gives good

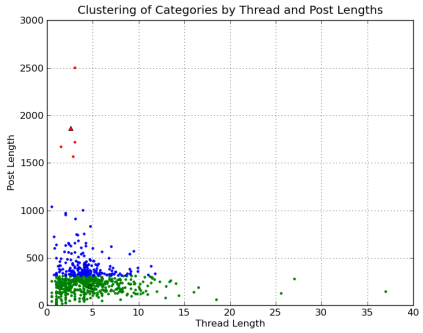

Figure 8: Clustering based on Thread Length and Post Length

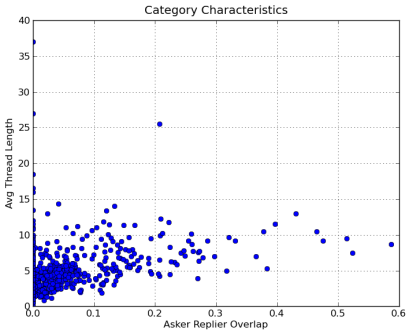

Figure 9: Distribution of Categories

recall (85 - 92%) with reasonable cluster sizes (50). However, the precision was still found to be $< 10\%$. We hope to incorporate another step that would overcome this drawback. To this effect, we plan to use a supervised learning approach that would enhance the precision.

Thus, in this paper, we have explored the characteristics of the Yahoo Answers Dataset. We also propose a way of suggesting a pool of potential answerers by using the properties of the question, the category and the question asker. With this we get a high recall value but low precision value, and so to overcome this we suggested using the process of cascaded clustering. Alternatively, we could could train a classifier Q such that Q <Category,$\text{Qn}_A sker, User >$ $returns 1 if user is the answerer; 0 otherwise.$

9. REFERENCES

[1] Lada Adamic et al. *Knowledge Sharing and Yahoo Answers: Everyone knows Something*, WWW 2008

[2] Eugene Agichtein,Carlos Castillo et al. *Finding High Quality Content in Social Media* WSDM 2008

[3] Jun Zhang, Mark S Ackerman Lada Adamic *Expertise Networks in Online Communities:Structure and Algorithms* WWW 2007

[4] Lada Adamic *Expertise Sharing Dynamics in Online Forums*

[5] Pawel Jurczyk, Eugene Agichtein *Discovering Authorities in Question Answer Communities by Using Link Analysis* CIKM 2007

[6] Pawel Jurczyk, Eugene Agichtein *HITS on Question*

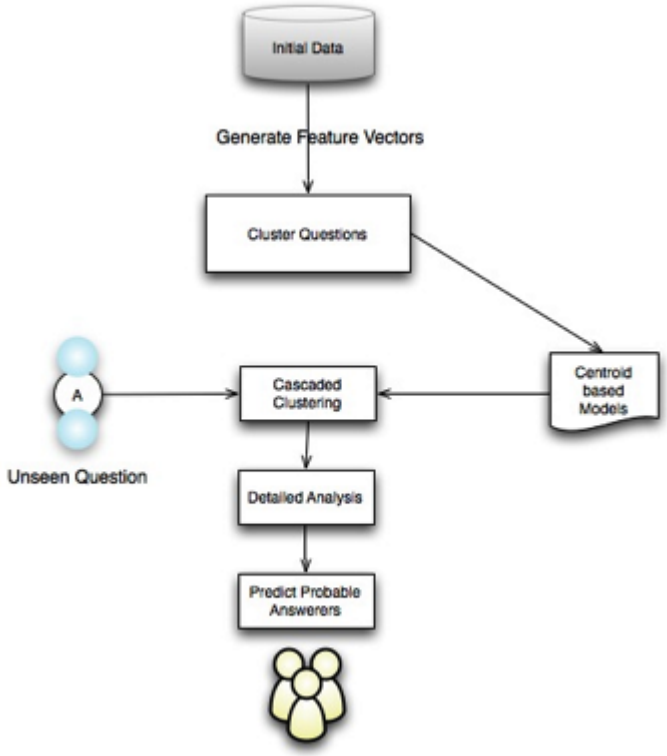

Figure 10: Answer Pool Detection by Cascaded Clustering

Answer Portals : Exploration of Link Analysis for Author Ranking SIGIR 2007

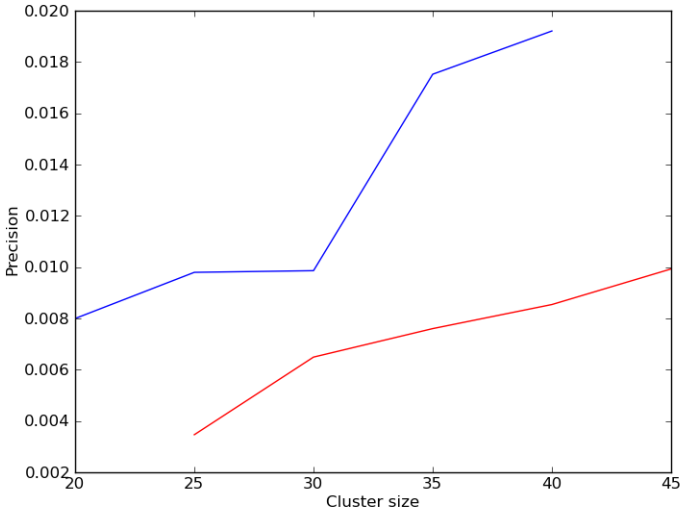

Figure 11: Result : Variation of Precision

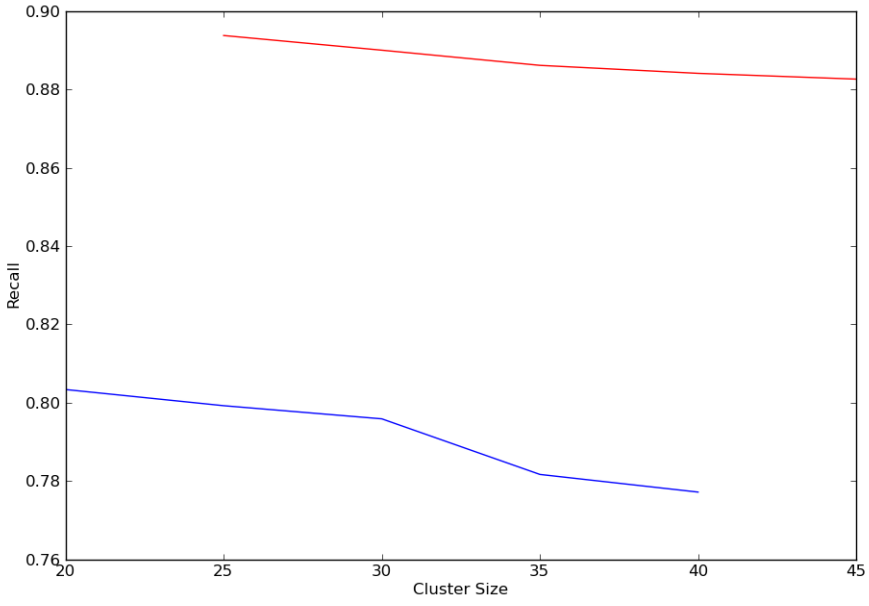

Figure 12: Result :Variation of Recall