Modeling Information-Seeker Satisfaction in Community Question Answering

EUGENE AGICHTEIN and YANDONG LIU
Emory University
and
JIANG BIAN
Gerogia Institute of Technology

Question Answering Communities such as Naver, Baidu Knows, and Yahoo! Answers have emerged as popular, and often effective, means of information seeking on the web. By posting questions for other participants to answer, information seekers can obtain specific answers to their questions. Users of CQA portals have already contributed millions of questions, and received hundreds of millions of answers from other participants. However, CQA is not always effective: in some cases, a user may obtain a perfect answer within minutes, and in others it may require hours—and sometimes days—until a satisfactory answer is contributed. We investigate the problem of predicting information seeker satisfaction in collaborative question answering communities, where we attempt to predict whether a question author will be satisfied with the answers submitted by the community participants. We present a general prediction model, and develop a variety of content, structure, and community-focused features for this task. Our experimental results, obtained from a large-scale evaluation over thousands of real questions and user ratings, demonstrate the feasibility of modeling and predicting asker satisfaction. We complement our results with a thorough investigation of the interactions and information seeking patterns in question answering communities that correlate with information seeker satisfaction. We also explore *personalized* models of asker satisfaction, and show that when sufficient interaction history exists, personalization can significantly improve prediction accuracy over a "one-size-fits-all" model. Our models and predictions could be useful for a variety of applications, such as user intent inference, answer ranking, interface design, and query suggestion and routing.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms: Algorithms, Design, Experimentation, Evaluation

E. Agichtein and Y. Liu were partially supported by the Emory College Seed Fund.
Authors' addresses: E. Agichtein and Y. Liu, Emory University, Mathematics and Computer Science Department, 400 Dowman Drive, Suite W401, Atlanta, GA 30322; email: {eugene,yliu49}@mathcs.emory.edu. J. Bian, Georgia Institute of Technology College of Computing, Klaus Advanced Computing Building, 266 Ferst Dr., Atlanta, GA 30332; email: jbian3@mail.gatech.edu.

DOI 10.1145/1514888.1514893 http://doi.acm.org/10.1145/1514888.1514893

Additional Key Words and Phrases: Community question answering, information seeker satisfaction

ACM Reference Format:

Agichtein, E., Liu, Y., and Bian, J. 2009. Modeling information seeker satisfaction in community question answering. ACM Trans. Knowl. Discov. Data. 3, 2, Article 10 (April 2009), 27 pages. DOI = 10.1145/1514888.1514893 http://doi.acm.org/10.1145/1514888.1514893

1. INTRODUCTION

Community question answering (CQA) emerged as a popular alternative to finding information online. It has attracted millions of users who post millions of questions and hundreds of millions of answers, producing a huge knowledge repository of all kinds of topics, so many potential applications can be possibly made on top of it. For example, automatic question answering systems, which try to find the information to questions directly instead of giving a list of related documents, might use CQA repositories as a useful information source. In addition, instead of using general-purpose web search engines, information seekers now have an option to post their questions (often complex and specific) on Community QA sites such as Naver or Yahoo! Answers, and have their questions answered by other users. These sites are growing rapidly. However, it is not clear what information needs these CQA portals serve, and how these communities are evolving. Understanding the reason for the growth, the characteristics of the information needs that are met by such communities, and the benefits and drawbacks of community QA over other means of finding information, are all crucial questions for understanding this phenomenon.

In this article we pose one such fundamental question: can we predict if an asker in CQA will be *satisfied* with the answers contributed by the community? Our goal is to begin to unravel the many factors that go into success of a CQA portal, and ultimately to apply our insights to better design of social media applications. In particular, community question answering allows us to directly study search satisfaction from the information seeker perspective. This is in contrast to the more traditional relevance-based assessment that is often done by judges different from the original information seeker, which may result in ratings that do not agree with the target user. While the idea of relevance being inherently subjective has been pointed out in the past (e.g., see Zobel [1998] and more recently Ruthven et al. [2007]), nowhere does the problem of subjective relevance arise more prominently than within Community QA, where many of the questions are inherently subjective, complex, ill-formed, or often all of the above. The problem of complex and subjective QA has only recently started to be addressed in the question answering community, most recently as the first opinion QA track in TREC [Dang et al. 2007]. We review related work in more detail in Section 8.

In addition to studying asker satisfaction to expand our understanding of information seeking, there are significant practical benefits to predict satisfaction in CQA. Potential applications include user intent inference, answer ranking, and query suggestion and routing. For example, we could notify the

information seeker when an appropriate answer has been posted (which we call the "offline" setting), or predicting at the time of posting whether the asker is likely to get a satisfactory answer to this question (the "online" setting). As we will show, human assessors have a difficult time predicting asker satisfaction, thereby requiring novel prediction techniques and evaluation methodology that we begin to develop in this paper.

Not surprisingly, user's previous interactions such as questions asked and ratings submitted are a significant factor for predicting satisfaction. We hypothesized that asker's satisfaction with contributed answers is largely determined by the asker expectations, prior knowledge and previous experience with using the CQA site. We report on our exploration of how to *personalize* satisfaction prediction—that is, to attempt to predict whether a *specific* information seeker will be satisfied with any of the contributed answers. Our aim is to provide a "personalized" recommendation to the user that they have answers that satisfy their information need.

This article provides the following contributions:

—Investigation of the problem of predicting asker satisfaction in QA communities (Section 3).
—Describes a general prediction framework that can work in both offline and online settings (Section 4).
—Reports on a thorough evaluation of both automatic and manual asker satisfaction predictions over thousands of real users' questions (Section 6).
—Investigates which features and methods are most effective for predicting asker satisfaction (Section 6).
—Reports on the feasibility of personalization for predicting satisfaction (Section 7).

This article expands upon and merges our previous work [Liu et al. 2008; Liu and Agichtein 2008a; Liu and Agichtein 2008b]. We describe the relationship of the current article to previous work in Section 8.

2. COMMUNITY QUESTION ANSWERING

Community Question Answering is an environment where users can post questions and provide answers to seek specific information need. It has emerged as a popular alternative to finding information online and has attracted millions of users who post millions of questions and hundreds of millions of answers, producing a huge knowledge repository of all kinds of topics. Figure 1 shows a typical question thread from Yahoo! Answers Website. As the figure illustrates, questions can often attract excellent, thorough, and well-researched answers that integrate a variety of information sources—the result of which cannot be matched by automatic search technologies. Of course, this is not always the case, as we explore in the rest of the article. But first, let us examine in more detail the process by which many CQA portals operate. For concreteness, we will focus on the Yahoo! Answers CQA portal, but the observations and procedures are quite similar in other portals such as Naver and Baidu Knows.

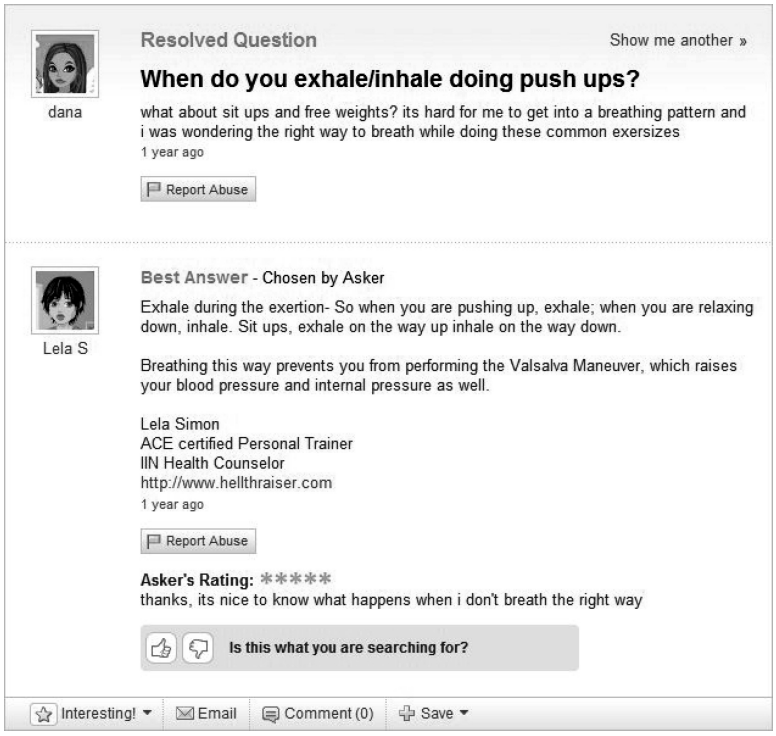

Fig. 1. An example of a Yahoo! Answers question thread.

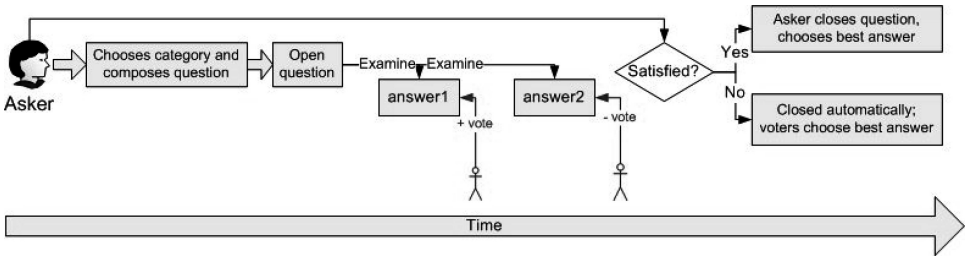

Fig. 2. A simplified lifecycle of a question in a typical CQA site.

2.1 Lifecycle of a Question in Community QA

The process of posting and obtaining answers to a question in CQA is outlined in Figure 2. A user posts a question by selecting a category, and then enters the question subject (title) and, optionally, detail (description). For conciseness, we will refer to this user as the *asker* for the context of the question, even though the same user is likely to also answer other questions or participate in other roles for other questions. Note that to prevent abuse, the community rules typically forbid the asker from answering their own questions or vote on answers. After a short delay (which may include checking for abuse, and other processing) the question appears in the respective category list of *open* questions, normally listed from the most recent down. At the point, other users can *answer* the

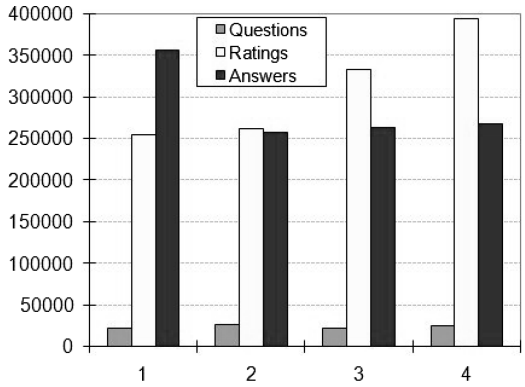

Fig. 3. Questions, answers, and votes contributed during period 1: August 2006–November 2006, 2: December 2006–March 2007, 3: April 2007–July 2007, and 4: August 2007–November 2007 .

question, *vote*[1] on other users' answers, or *comment* on the question (e.g., to ask for clarification or provide other, nonanswer feedback), or provide various metadata for the question (e.g., give questions *stars* for quality). Depending on the site, many more interactions may be available.

2.2 Case Study: Growth of the Yahoo! Answers Community

Community Question Answering sites have been growing rapidly, which may be changing the dynamics of information seeking. To understand whether our results are likely to apply in the future, and to gain better understanding of the CQA setting, we report the statics on the growth of the Yahoo! Answers site during the years of 2006 and 2007. For this experiment we collected a large sample of the Yahoo! Answers site, resulting in an archive of 96,000 questions and 1,150,000 answers, covering about 125 categories. We divided all the questions and answers equally by posting time, resulting in four time periods: August 2006–November 2006; December 2006–March 2007; April 2007–July 2007; and August 2007–November 2007.

With this large sample, we can examine the growth of the Yahoo! Answers community (Figure 3). We report the numbers of newly posted questions, answers, and the number of votes (either positive or negative) contributed during each time period. The number of newly produced questions and answers during each time period remained steady, though their speed of growth decreases. This is a counterintuitive finding, suggesting that users rarely post multiple questions: if existing users were continuing to post questions, the number of questions would increase as new users join the site. It is also worth noting that the number of ratings votes increases faster than the content. This implies that instead of contributing more answers, Yahoo! Answers participants are more actively rating the contributions of other users.

[1]Yahoo! Answers and many other CQA sites distinguish a "vote" for the best answer (if not closed by asker in a timely manner), from the "thumbs up" and "thumbs down" ratings. We use the term "votes" to include the "thumbs up" or "thumbs down" ratings to avoid confusion with the asker rating.

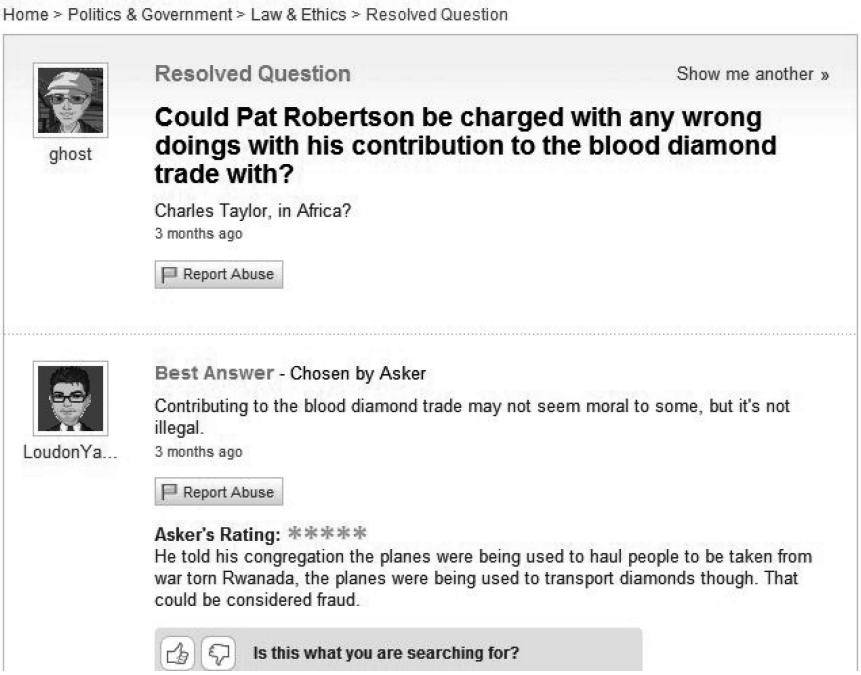

Fig. 4. A Yahoo! Answers "satisfied" question thread.

Now that we have introduced and explored the Community Question Answering setting, we proceed to the core of our paper, that is, to define and solve the *information seeker satisfaction problem*.

3. THE ASKER SATISFACTION PROBLEM

Recall, that in CQA a user posts a question by selecting a category, and then enters the question. After a short delay (which may include checking for abuse, and other processing) the question appears in the respective category list of *open* questions. At the point, other users can *answer* the question, *vote* on other users' answers, or *comment* on the question. If the asker is satisfied with any of the answers, she can choose it as *best*, and provide feedback ranging from assigning *stars* or rating for the best answer, and possibly textual feedback. We believe that in such cases, the asker is likely *satisfied* with at least one of the responses, usually the one she chooses as the best answer. An example of such "satisfactory" interaction is shown in Figure 4.

But in many cases the asker never closes the answer personally, and instead, after some fixed period of time, the question is *closed automatically*. So, while it is possible that the best answer chosen automatically is of high quality, it is unknown if the asker's information need was satisfied. There may be many reasons why the asker never closed a question by choosing a best answer. Based on our exploration we believe that the main reasons are either a) user loses interest in the information and b) none of the answers are satisfactory. In both cases, the QA community has "failed" to provide satisfactory answers in a timely

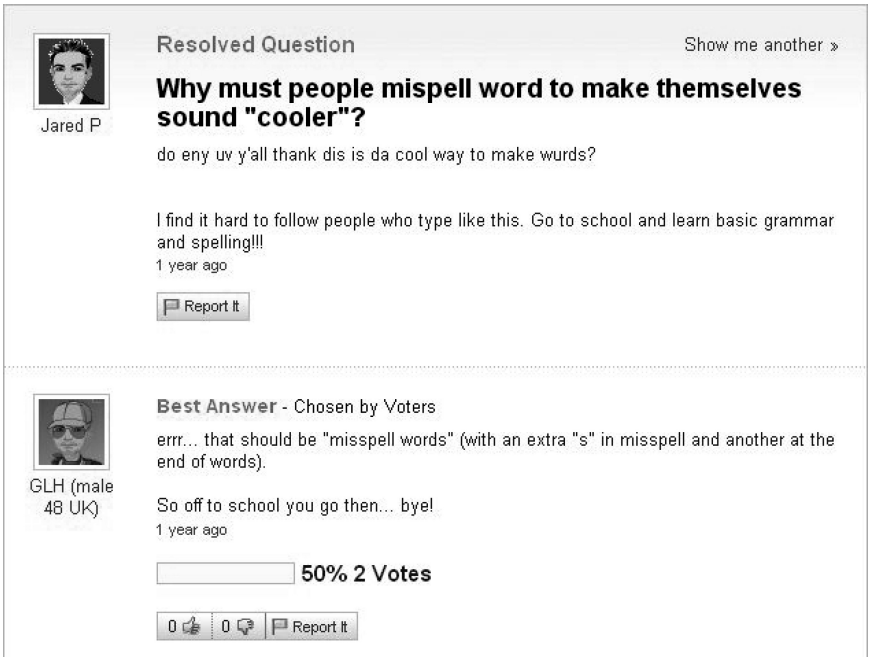

Fig. 5. A Yahoo! Answers "unsatisfied" question thread.

manner and "lost" the asker's interest. While the true reasons are not known, for simplicity, to contrast with the "satisfied" outcome above, we consider this outcome to be "unsatisfied." An example of such interaction is shown in Figure 5.

The problem of whether an asker in QA community is satisfied is a special instance of the general problem of predicting if an information need of a searcher is satisfied. Question Answering communities are an important application by itself, and also provide unprecedented opportunity to study feedback from the asker. Furthermore, asker satisfaction plays crucial role in the growth or decay of a question answering community. If many of the askers in CQA are not satisfied with their experience, they will not post new questions and will rely on other means of finding information. Furthermore, by modeling asker satisfaction, we could provide better ranking of questions, or notify an asker if they are likely to be satisfied with the answers to their questions. Hence, predicting, understanding and monitoring asker satisfaction is at the core of maintaining an active and healthy QA community.

It is important to note the differences of our task from traditional question answering and ad-hoc information retrieval: we want to predict what is essentially a *subjective* notion of satisfaction, which requires to model the intent of the asker, the expectation of what comprises a satisfying answer, and to some extent providing a "recommendation" to the asker on the expected satisfaction with the answers. Furthermore, the information needs of askers in CQA are typically more *complex* and *subjective* compared to the traditional TREC benchmarks. Often, the intent of the asker is not obvious to either annotator or community participants, as we explore in Section 6. In summary, we believe

that asker satisfaction, as studied in the context of CQA, can provide both new insights into information-seeking behavior and spur the development of new techniques for user modeling and information finding.

3.1 Problem Statement

We now state more formally what we mean by asker satisfaction:

Definition 3.1. An asker in a QA community is considered *satisfied* iff: the asker personally has closed the question, selected the best answer, and provided a rating of at least 3 "stars" for the best answer quality. Otherwise, we define the asker to be *unsatisfied*.

We believe that this definition captures key aspects of asker satisfaction, namely that we can reliably identify when asker is satisfied but not the converse. Similarly, we do not attempt yet to analyze the distinction between possibly satisfied and completely unsatisfied, or otherwise dissect the case where the asker is not satisfied. We now state our problem more formally:

The Asker Satisfaction Problem: Given a question submitted by an asker in CQA, *predict* whether the user will be *satisfied* with the answers contributed by the community.

There are two important special cases of this problem: the *offline* setting, where the posted question has already obtained some answers; and the *online* setting, where we attempt to predict immediately whether a user will be satisfied with the answers at some intermediate point in the process (e.g., while answers are still arriving), or even before any answers arrive. We will attempt to solve the general version of this problem by adapting machine learning techniques, and, as our results in Section 6 show, our techniques are feasible for both the offline and online variants of the problem.

4. PREDICTING ASKER SATISFACTION

Our approach to predicting asker satisfaction is to apply supervised machine learning algorithms, trained on past user interactions. Specifically, we developed our Asker Satisfaction Prediction system (ASP) that learns to classify whether the question asker is satisfied with the obtained answers. Given a *question thread* posted by an asker, we derive features to represent the associated information (e.g., question text, text of the answers, user feedback) to predict whether the asker would be satisfied. Naturally, the features used are crucial (which we describe next). We then briefly describe the specific classification algorithms used for the experiments of Section 6.

4.1 Features

Our features are organized around the basic entities in a question answering community: questions, answers, question-answer pairs, users, and categories. We now review the features we used to represent our problem. Table I includes

Table I. Sample Features:
Question (Q), Question-Answer Relationship (QA), Asker User History (UH), Answerer User History (AH), and Category Features (CA). Textual features are not listed. The complete set of features is listed in Appendix A.

Feature (72 total)	Description	Known at Question?
Question (from 23 total)		
Q: Subj_Len	Number of words in question subject	yes
Q: Post_Time	Time(in hours) of the day when the question was posted	yes
Q: Num_Ans	Number of answers received for this question	no
Q: Stars	Number of stars received for this question	no
Q: Wh-type	Wh-word introducing the question title (e.g., "what", "where", etc.)	yes
Q: Num_Comm	Number of comments added by other participants	no
Q: KLDivergence	KL-Divergence with Wikipedia/TREC/ language model	yes
Q: Visual_Quality	Question puctuation/typo/space density	yes
Question-Answer Relationship (from 18 total)		
QA: Total_Pos/Neg_Votes	Total number of positive and negative votes	no
QA: Avg_Pos/Neg_Votes	Average number of positive/negative votes	no
QA: Top_10_Ans_KLDiv	KL-Divergence of top ten answers' content with Wikipedia/TREC/ language model	
QA: Top_10_Ans_Visual	Top ten answers content puctuation/typo/space density	no
Asker User History (6 total)		
UH: Ques_Resolved	Number of questions resolved in the past	yes
UH: Total_Ans_Received	Total past number of answers this user obtained	yes
UH: Member_Since	How long since the user's last registration	yes
UH: Ans/Ques_Ratio	Ratio of answers to number of questions posted	yes
UH: Avg_Past_Rating	Average rating of past closed questions	yes
UH: Most_Recent_Rating	Most recent rating given for last question	yes
Answerer Reputation (from 21 total)		
AH: Sum_Points_Ans	Combined reputation points of all answerers	no
AH: Sum_Past_Best_Ans	Total number of best answers, combined for all answerers	no
AH: Max_Member_Since	Length of longest registration among all answerers	no
AH: Avg_Best_Ans_Ratio	Average best answer ratio, combined over all answerers	no
AH: Max_Pts_Ans	Highest reputation points among answerers	no
Category Features (6 total)		
CA: Avg_Close_Time	Average interval to close a question in category	yes
CA: Avg_Ans_Count	Average number of answers per question for category	yes
CA: Avg_Rating	Average rating by asker when closing questions in category	yes
CA: Avg_Ques_Arrival	Average number of questions posted per hour in category	yes
CA: Avg_Ans_Arrival	Average number of answers posted per hour in category	yes

some samples features used in our experiments. The complete list is reported in Appendix A.

Question. This group includes traditional question answering features such as the words and 2-word phrases in the question, the wh-type (e.g., "what" or "where"), and the length of the subject (title) and detail (description) of the question. As a more specific feature to communities we also include posting time, as well as any user feedback received for the question (e.g., "stars" in Yahoo! Answers community).

Question-Answer Relationship. This group describes the relationship between the question and the answer. We include standard features such as overlap between question and answer, answer length, and number of candidate answers. We also use specialized features such as the number of positive votes ("thumbs up" in Yahoo! Answers), negative votes ("thumbs down"), and various vote-related statistics such as the maximum of positive or negative votes received for any one answer (e.g., to detect cases of brilliant/popular answers or, conversely, blatant abuse).

Asker User History. This group is unique to question answering communities, and particularly important for our task. Since user satisfaction is, to a large extent, subjective, we posit that it relies largely on past user activity history – in particular, how the asker was satisfied with responses to previous questions. Care was taken not to "cheat"—only information available about the asker *prior* to posting the question was used.

Answerer Reputation. Similarly to the *Asker User History*, we develop features to describe the history of the users providing the answers, such as the number of questions resolved, number of answers provided, and number of answers rated as *best*. Since a question may draw multiple answers, we include three "surrogate" answerer features: the average of the answerer history, the features for the answerer with the highest CQA reputation score, and the answerer that attracted the most positive votes for this question.

Category Features. We hypothesized that user behavior (and asker satisfaction) varies by topical question category, as recently shown in reference [**?**]. Therefore we model the *prior* of asker satisfaction for the category, such as the average asker rating (satisfaction) with answers contributed to all previous questions in the category.

Textual Features. Additionally, we derive word n-gram (unigram and bigram) features from the text of the question, and the text of the answers (separate features spaces are used to represent the question and answer terms). As a simple feature selection method, only the most frequent 1000 features are included.

4.2 Classification Algorithms

We explored three families of classification algorithms: Support Vector Machines (SVM), Decision trees, Boosting and Naive Bayes, all using the implementations in the Weka [Witten and Frank 2005] framework.

Decision Trees. We use two implementations of the decision tree [Quinlan 1996]: C4.5 and RandomForest. A benefit of decision tree is interpretability of the models and results. By using a decision tree classifier, we expect to get high precision on the target class, with the potential drawback of overfitting. To account for this, we use the Random Forrest classifier (that avoids overfitting by selecting feature subsets), as well as explicit feature selection.

SVM. Support vector machines are considered the classifier of choice for many tasks, due to robustness in the presence of noise, and high reported accuracy. Specifically, we use the Weka implementation of SMO [Platt 1998].

Boosting. Additionally, we use metalearning as an alternative to SVM for the noisy features (and labels) in our domain. AdaBoost [Freund and Schapire 1996] has been shown quite effective for many text-classification applications, and we apply the Weka implementation of AdaBoost.

Naive Bayes. Last, we use Naive Bayes classifier, which is a very simple and fast, yet often surprisingly effective method to quickly investigate the success of our approach.

The methods above are representative of the state of the art in classification, so we expect the experimental results described in Section 6 to be generalizable to other variants of classification algorithms.

5. EXPERIMENTAL SETUP

We now describe the metrics used for the evaluation, the datasets, and methods compared in the experimental results of Section 6.

5.1 Evaluation Metrics

Even though ours is formally a two-class classification problem, we primarily focus on the *satisfied* or positive class. The reason for this is that we have higher certainty about the true positive likelihood of our *satisfied* labels compared to the *unsatisfied*—more properly to be stated as *unknown* cases. Specifically, we measure the *Precision*, *Recall*, and *F1* for the *satisfied* class, and, where appropriate, the overall *Accuracy* for both classes.

—*Precision.* the fraction of the predicted *satisfied* asker information needs that were indeed rated satisfactory by the asker.

—*Recall.* the fraction of all rated *satisfied* questions that were correctly identified by the system.

—*F1.* the geometric mean of Precision and Recall measures, computed as $\frac{2PR}{P+R}$.

—*Accuracy.* the overall fraction of instances classified correctly into the proper class. Often, accuracy is not the right metric when the class distribution is skewed; however, for completeness, we will also report Accuracy in some of our experiments.

In the experiments that follow we will primarily focus on predicting the *satisfied* class, hence we will rely more on the Precision, Recall, and F1 rather than the overall Accuracy.

5.2 Human Judgments

Our problem is inherently subjective. Hence, as the gold standard we use the asker rating for the best answer (if chosen) as a measure of satisfaction. Note that in many cases askers do not even bother to choose the best answer, indicating a degree of dissatisfaction that we plan to quantify in future work. For this study, however, we simply consider the asker ratings as the "truth," interpreted as defined in Section 3.1.

Table II. Ratings for 130 Questions (54 satisfied/76 unsatisfied)

Rater Group	Redundancy	Agreement
Experts	3	0.82
Mechanical Turk Workers	5	0.9

Table III. Statistics of the Complete Data Crawled from the Yahoo! Answers Site

Questions	Answers	Askers	Categories	Satisfied (%)
216,170	1,963,615	158,515	100	50.7

To complement the asker ratings we also obtained human judgements from Amazon's paid rater service, the Mechanical Turk.[2] The raters are provided a "HIT" (Human Intelligence Task), and for a small fee the workers submit their responses. For our task we obtained five independent ratings for each question, and used a majority to identify and resolve ambiguous cases. In total, 130 questions were manually rated by Mechanical Turk workers. Finally, we obtained a number of "expert" ratings—provided by researchers to calibrate the asker satisfaction and the Mechanical Turk (henceforth, MTurk) ratings. Interestingly, as we will show in Section 6, MTurk ratings have higher correlation with the asker satisfaction than the (more strict) expert ratings. The rated dataset is summarized in Table II.

5.3 Datasets

Our data is based on a snapshot of Yahoo! Answers (`http://answers.yahoo.com`), a popular CQA site, crawled in the early 2008. The initial broad categories to start the crawl were "Health," "Education & Reference," "Sports," "Science and Mathematics," and the "Arts." The resulting snapshot is our universe of 216,170 questions, summarized in Table III and Table IV.

In order to focus on a realistic asker satisfaction prediction task (that is, reflective of the *current* state of Yahoo! Answers), we selected a random subset of **5,000** questions from the *most recent* 10,000 questions in the snapshot above. We will use this sample of 5,000 questions for all of the experiments. To allow other researchers to replicate our results, all the datasets used in this article are available online.[3]

The details of our dataset are reported in Table V and VI. The total of 90 categories are represented, and we report detailed statistics for the top 10 most frequent categories. As we can see, questions in these categories comprise almost 51% of all questions in the dataset (this skewed distribution is representative of our complete crawl snapshot). In particular, the *Mathematics* category is the most popular, containing 13% of the questions and drawing on 3.6 answers for each question on average. Interestingly, *Chemistry*, while also a popular category, draws only about 2 answers per question, while *Football (American)* attracts more than 11 answers for each question. The asker satisfaction varies widely with the category. While more than 70% of askers are satisfied with the answers provided in the *Mental Health* category, only 34% are satisfied with the

[2]`http://www.mturk.com/mturk/welcome`

[3]`http://ir.mathcs.emory.edu/shared/sigir2008`

Table IV. Distribution of Questions, Answers and Askers

#Questions per Asker	# Questions	# Answers	# Users
1	132,279	1,197,089	132,279
2	31,692	287,681	15,846
3–4	23,296	213,507	7,048
5–9	15,811	143,483	2,568
10–14	5,554	54,781	481
15–19	2,304	21,835	137
20–29	2,226	23,729	93
30–49	1,866	16,982	49
50–100	842	4,528	14
Total:	216,170	1,963,615	158,515

Table V.
Selected statistics for the top 10 most popular categories in our dataset (together comprising 51% of questions in dataset).

Category	Questions	Answers per Question	Answers	Freq.	Satisfied
Mathematics	651	2,329	3.58	13.0%	44.5%
Diet & Fitness	450	2,436	5.41	9.0%	68.4%
Women's Health	277	1,824	6.58	5.5%	62.8%
Chemistry	236	508	**2.15**	4.7%	37.3%
Biology	176	589	3.35	3.5%	**34.1%**
Books & Authors	161	645	4.01	3.2%	42.2%
Football (American)	152	1,722	**11.33**	3.0%	55.3%
Mental Health	151	1,159	7.68	3.0%	**70.9%**
Physics	149	428	2.87	3.0%	48.3%
General Health	135	737	5.46	2.7%	70.4%
Cumulative (10 Cat.)	2,538	12,377	4.88	50.8%	53.4%
Overall (90 Cat.)	5,000	25,063	5.01	100%	50.7%

Table VI.
Selected statistics for the top 10 most popular categories in our dataset (together comprising 51% of questions in dataset).

		Closed by Asker		Closed by Voters	
Category	Time to Close	Asker Rating	Time to Close	Voter Rating	Time to Close
Mathematics	3 days 20 hours	4.48	33 minutes	1.76	6 days
Diet & Fitness	2 days 17 hours	4.30	1.5 days	**4.46**	6 days
Women's Health	2 days 23 hours	4.28	35 minutes	1.98	6 days
Chemistry	4 days 7 hours	4.39	1 day 13 hours	**1.19**	6 days
Biology	4 days 5 hours	4.06	28 minutes	1.33	6 days
Books & Authors	4 days 6 hours	4.35	1 day 20 hours	2.13	6 days
Football (American)	3 days 11 hours	4.29	1 day 13 hours	2.05	6 days
Mental Health	2 days 16 hours	4.30	1 day 13 hours	1.32	6 days
Physics	3 days 13 hours	4.29	35 minutes	1.48	6 days
General Health	2 days 17 hours	4.49	1 day 13 hours	1.31	6 days
Cumulative (10 Cat.)		4.32		1.90	
Overall (90 Cat.)	3 days 15 hours	4.32	1 day 12 hours	1.87	6 days

answers contributed for *Biology* questions, and similar low satisfaction holds for other sciences. Not surprisingly, questions that are closed by the asker are usually closed within a day (and often within 1 hour). Also, when the asker closes the question personally, the asker rating is usually high, averaging 4.3 "stars" out of 5 possible, with low variance across categories. However, when a question is closed by community voters, the average number of votes awarded to the best answer varies widely by category. For example, Voters in the *Chemistry* category on average award only 1.2 votes to the best answer (despite the high popularity of the *Chemistry* category). In contrast, voters in the *Diet & Fitness* category on average award about 4.5 votes to the best answer, which indicates higher overall satisfaction of the community with the contributed answers. In summary, asker satisfaction and other statistics of the questions vary widely by the topical category, and the corresponding user community, supporting our decision to develop a number of category-normalized features (Section 3).

5.4 Methods Compared

We now describe the baselines and our specific methods for predicting asker satisfaction. Note that the "truth" labels are provided by the asker and hence are difficult to predict even for human judges. The predictions we compare include:

—*Human*. As the human raters we report the prediction of Amazon's Mechanical Turk workers: a question is predicted as *satisfied* if the majority of raters label the best answer as satisfactory. The specific threshold for a majority will be fixed in our calibration experiments in the next section.

—*Heuristic*. Intuitively, if a question receives many answers, at least one of them should be satisfactory. Therefore, our heuristic baseline predicts the label *satisfied* if a question received many answers. The exact threshold on the number of answers is set using a decision tree (C4.5 in our experiments).

—*Baseline*. Random baseline, that simply predicts the majority class (which is usually *satisfied*).

—*ASP_SVM*. Our system implementation using the SVM classifier (Section 4.2).

—*ASP_RandomForest*. Our system implementing a decision tree classifier using the random forest.

—*ASP_C4.5*. Our system implementing a decision tree using the C4.5 algorithm (Section 4.2).

—*ASP_Boosting*. Our system implementing the AdaBoost algorithm combining weak learners (Section 4.2).

—*ASP_NB*. Our system implementing the Naive Bayes classifier (Section 4.2).

We now turn to the experimental evaluation of the asker satisfaction prediction methods.

6. EXPERIMENTAL RESULTS

First, we present some intuitions into the problem itself. In Section 6.1 we report the main classification results of the paper, which we subsequently will

Table VII. Comparing casual human raters (Mechanical Turk Workers) with expert raters (130 randomly sampled questions)

Rater Group	Precision	Recall	F1	Accuracy
Expert (strict)	0.36	0.68	0.47	0.45
Casual (majority=3/5)	0.43	1.0	0.60	0.47
Casual (majority=4/5)	**0.44**	**1.0**	**0.61**	**0.48**
Casual (majority=5/5)	0.41	0.75	0.53	0.46

Table VIII.
Accuracy of ASP_SVM, ASP_C4.5, ASP_RandomForest, ASP_Boosting, and ASP_NB for varying parameters (5-fold cross validation).

	With Text		Without Text		Selected Features	
Classifier	F1	Accuracy	F1	Accuracy	F1	Accuracy
ASP_SVM	0.69	0.70	0.72	0.73	0.62	0.70
ASP_C4.5	0.75	0.74	0.76	0.75	**0.77**	0.77
ASP_RandomForest	0.70	0.67	0.74	0.73	0.68	0.68
ASP_Boosting	0.67	0.72	0.67	0.72	0.67	0.72
ASP_NB	0.61	0.63	0.65	0.68	0.58	0.67
Human	0.61	0.48				
Baseline	0.66	0.51				

study in depth in the remainder of the section. In particular, we show that our ASP system is able to take advantage of the context (i.e., asker user history) to make better predictions than human raters. We conclude this section with feature analysis and analysis of the results (Section 6.2).

Before we present our experiments, it is important to understand the difficulty of the problem of predicting asker satisfaction (Table VII). For example, the labels of expert judges at best had weak correlation with asker satisfaction, and with the most favorable thresholding only achieved the precision of 0.36 and recall of 0.68 when trying to predict satisfaction. Similarly, Mechanical Turk workers (whom we call "casual labelers"), had better success with precision of 0.44 and recall of 1 (i.e., they were overly optimistic about satisfaction). Interestingly, the best precision and recall were achieved not where all the raters agreed, but rather when at least 4 out of 5 raters predicted asker satisfaction. Based on these results, we will use the Mechanical Turk labels as the strongest manual baseline, using the majority threshold of 4 for all subsequent experiments.

6.1 Predicting Asker Satisfaction

Table VIII reports prediction accuracy for the different implementations of ASP, in particular comparing the choice in classifier algorithm and feature sets (namely, whether to use the textual features, and whether to use feature selection). Surprisingly, ASP_C4.5 results in the best performance of all the classification variants, with F1 on the *satisfied* class of 0.77 when selecting only the top 15 features, chosen by Information Gain. In contrast, the human raters only achiever the F1 of 0.61, which is in fact lower than the naive baseline that always guesses the "satisfied" class, and lower than the heuristic baseline that achieves the best F1 of 0.64.

Table IX. Top 15 Features with Highest Information Gain (IG)

IG	Feature
0.14219	UH: Most_Recent_Rating
0.13965	UH: Avg_Past_Rating
0.10237	UH: Member_Since
0.04878	UH: Avg_Ans_Received
0.04878	UH: Ques_Resolved
0.04381	CA: Avg_Rating
0.04306	UH: Total_Ans_Received
0.03274	CA: Avg_Ratings
0.03159	Q: Post_Time
0.02840	CA: Avg_Ans_Count
0.02633	AH: Max_Member_Since
0.02080	AH: Max_Best_Ans_Ratio
0.02046	AH: Avg_Best_Ans_Ratio
0.01747	CA: Avg_Answer_Arrival
0.01531	QA: Top_10_Ans_KLDiv_Wikipedia

Feature Selection. The top 15 features selected are reported in Table IX. Note that all the four asker history features are included. Interestingly, the most salient feature is the previous rating by the asker (when available). We can view it as the prior on the asker which may relate to the self-selecting nature of CQA (i.e., askers who recently were successful return to submit new questions). Similarly, the amount of experience with CQA (the "member since" features) is an important factor. Another interesting result is the presence of several category features, which confirms our intuition about the importance of the category as the prior on question satisfaction independent of the asker. Also note that the reputation of the answerers submitting the responses is not as important as many other features, suggesting that authority or expertise of answer contributors is only important for some, but not all, information needs.

We next report the precision, recall, and F1 for varying training set sizes in Figures 6, 7, and 8 respectively. We report the average of three experiments, each with a randomly chosen test set of 1,000 questions, held fixed for varying amounts of training data. Our ASP system outperforms all other predictors, including human raters. In particular, 2,000 questions in training is sufficient to achieve F1 of 0.75, and additional training data is not as helpful, nevertheless improving performance of ASP to achieve F1 of 0.77, substantially outperforming all other methods. In fact, as few as 500 training questions are sufficient to achieve F1 of 0.7, which may be practical enough even for the less popular question categories.

6.2 Analysis and Discussion

Online vs. Offline Prediction. Previously, we discussed results of predicting satisfaction in the *off-line* setting—that is, after some answers have been contributed, allowing us to exploit features such as the number of answers, answer content length, and feedback from other users (votes). We now consider a more difficult task of predicting asker satisfaction in the *online* setting—that

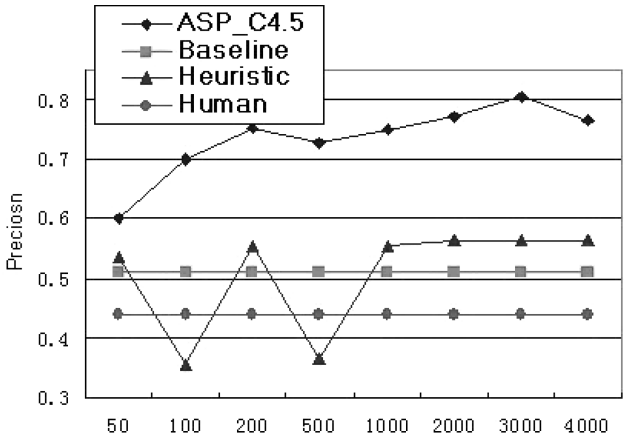

Fig. 6. Precision of ASP, Human, Baseline, and Heuristic for varying amount of training data.

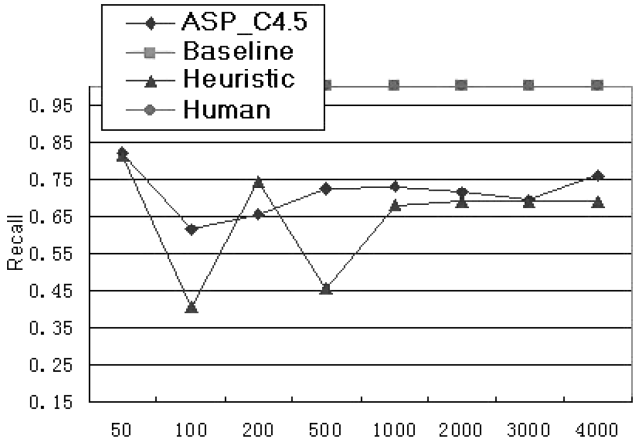

Fig. 7. Recall of ASP, Human, Baseline, and Heuristic for varying amount of training data.

is, before any answers to the question are contributed. Table X reports the comparison between offline and online settings. As we can see, there is a noticeable degradation in accuracy (0.74 F1 online vs. 0.77 F1 offline), nevertheless that performance is significantly higher than the various baselines—suggesting that ASP is practical even for online prediction.

Feature Ablation. To gain a better understanding of the important features for this domain, we perform ablation study on our feature set. For this, remove each of the feature categories listed in Section 4.1. Table XI reports the accuracy of ASP with each of the feature categories removed. Without question features or asker user history, the prediction F1 score drops drastically. In contrast, Question-Answer relationship, and Answerer User History, appear to have less of an effect—or perhaps are redundant given the presence of the other feature categories. Nevertheless, it should be noted that, surprisingly, answerer reputation does not appear to be important for asker satisfaction. We conjecture that

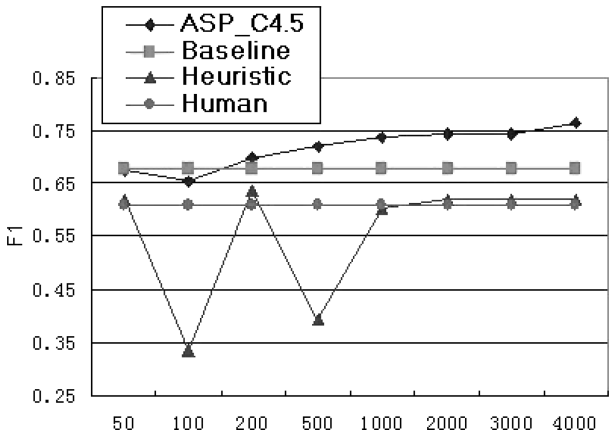

Fig. 8. F1 of ASP, Human, Baseline, and Heuristic for varying amount of training data.

Table X. Online vs. offline Prediction of Satisfaction

	Precision	Recall	F1
Online	0.78	0.70	0.74
Offline	0.78	0.76	**0.77**

Table XI. Prediction Accuracy with Feature Ablation

	Precision	Recall	F1
Selected features	0.80	0.73	**0.77**
No question features	0.68	0.72	0.70
No question-answer features	0.76	0.74	0.75
No asker features	0.72	0.69	0.71
No answerer features	0.76	0.75	0.75
No category features	0.75	0.76	0.75

this is due to increasingly subjective nature of many questions in CQA, where the accuracy of the provided answer is less important than other, more subjective characteristics of the answer, for example, whether the answer appears as caring or supportive for Health-related questions.

Textual Features. We also explore which textual features are most helpful for this task, using the Information Gain metric. From Table XII, it appears that most of the textual features suggest the predominance of subjective questions, which may in fact correlate with asker satisfaction (and requires further investigation).

Asker Satisfaction Varying with Past Experience. The importance of previous asker history features suggests that prediction accuracy should vary significantly with the amount of history available for the asker. To explore this hypothesis we test our model on groups of askers with varying number of previous questions posted. For this experiment, we train our ASP_C4.5 system as described before, but instead of averaging accuracy over all the questions in the test set, we compute Precision, Recall, and F1 separately for each group

Table XII. Textual Features with High Information Gain (IG)

IG	Feature
0.003734	”i don’t” in question
0.003335	”i was” in question
0.003147	”i have” in question
0.002595	”you are” in answer
0.002581	”to your” in answer
0.002543	”to get” in question
0.002536	”that i” in question
0.002532	”and i” in question
0.00238	”a few” in answer
0.002342	”but i” in question

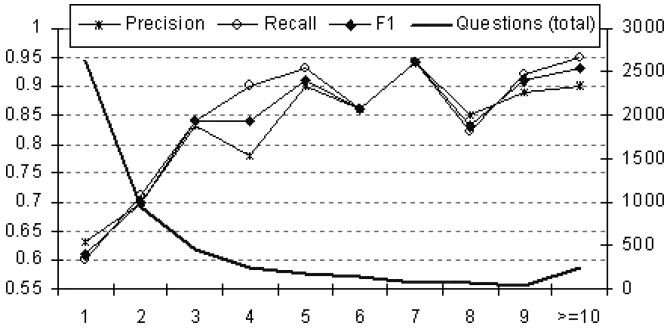

Fig. 9. Satisfaction prediction accuracy for groups of askers with varying number of posted questions, and the corresponding number of questions posted by askers in each group.

of askers. In particular, we group together questions from askers with just 1 question (that is, no prior questions posted), 2 questions (i.e., only 1 previous question posted), etc. The results are reported in Figure 9. Not surprisingly, the accuracy of prediction increases dramatically for askers with at least one previous question, reaching F1 of 0.9 for askers with at least three previous questions resolved in the past.

One explanation for the varying prediction accuracy with the asker’s length of membership, is “self-selection”: askers who have been satisfied with the CQA experience in the past are more likely to submit future questions. Figure 10(a) reports the average best answer rating (used as proxy for asker satisfaction), for varying number of days that an asker was a member of Yahoo! Answers. The first two to four days are crucial: those askers that are registered for less than three days tend to have far lower best answer ratings (0.98–1.0) than those who remain active members for three days or longer (1.98). Similarly, askers who submit more than two questions tend to be satisfied with the answers significantly more (average 2.3 best answer rating) than those with fewer previous question attempts (1.5), validating our intuition that askers who do not obtain a satisfactory answer to their first one or two questions, or within their first one or two days of memberships, are not likely to be satisfied in the future, or even attempt to submit questions again. Hence, personalizing satisfaction prediction would be highly effective, as we describe next.

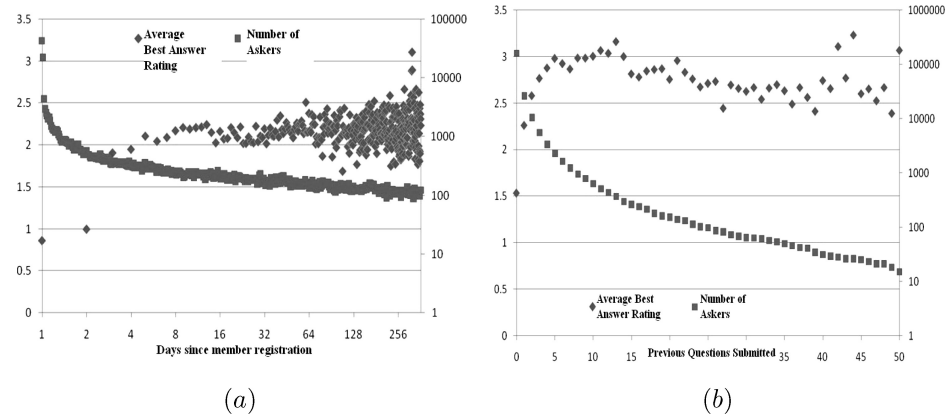

Fig. 10. Prediction accuracy and number of askers, for varying time of membership (a) and number of previous questions posted (b).

7. PERSONALIZED SATISFACTION PREDICTION

As the experimental results above indicated, personalizing satisfaction prediction may provide significant accuracy improvements when sufficient prior asker history exists. Our aim is to provide a "personalized" recommendation to the user that they've got answers that satisfy their information need. The following experiments are based on the general prediction framework we described in Section 3, using different training approaches. The personalization methods we have explored include:

—*ASP_Pers+Text*. We first consider the naive personalization approach where we train a separate classifier for each user. That is, to predict a particular asker's satisfaction with the provided answers, we apply the individual classifier trained solely on the questions (and satisfaction labels) provided in the past by that user.

—*ASP_Group*. A more robust approach is to train a classifier on the questions from the group of users *similar* to each other. Our current grouping was done simply by the number of questions posted, essentially grouping users with similar levels of "activity." As we will show below, text features only help for users with at least 20 previous questions. So, we only include text features for groups of users with at least 20 questions.

—*ASP*. Baseline method: a "one-size-fits-all" satisfaction predictor that is trained on 10,000 randomly sampled questions with only non-textual features

Certainly, more sophisticated personalization models and user clustering methods could be devised. However, as we show next, even the simple models described above prove surprisingly effective.

7.1 Experimental Results

Figures 11, 12, and 13 report the Precision, Recall, and F1 for ASP, ASP_Text, ASP_Pers+Text, and ASP_Group for groups of askers with varying number of

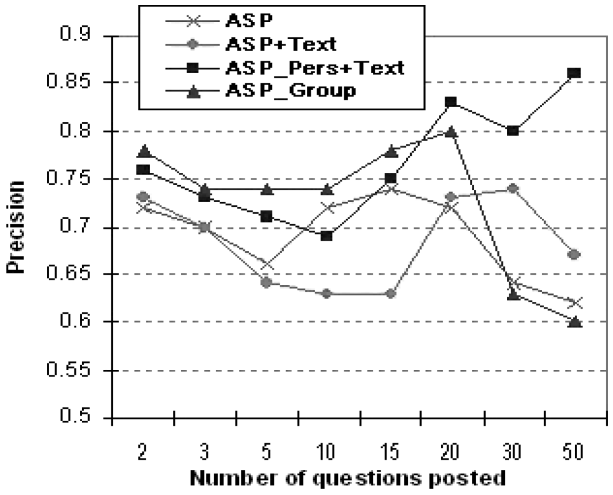

Fig. 11. Precision of ASP, ASP_Text, ASP_Pers+Text, and ASP_Group for predicting satisfaction of askers with varying number of questions.

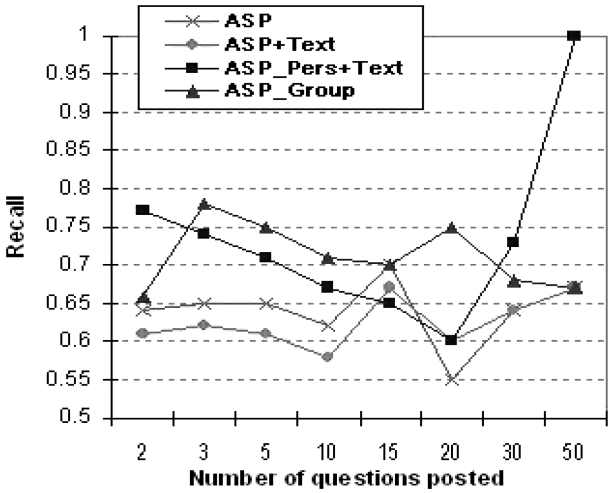

Fig. 12. Recall of ASP, ASP_Text, ASP_Pers+Text, and ASP_Group for predicting satisfaction of askers with varying number of questions.

previous questions posted. Surprisingly, for ASP_Text, textual features only become helpful for users with more than 20 or 30 previous questions posted and degrade performance otherwise. Also note that baseline ASP classifier is not able to achieve higher accuracy even for users with large amount of past history. In contrast, the ASP_Pers+Text classifier, trained only on the past question(s) of each user, achieves surprisingly good accuracy—often significantly outperforming the ASP and ASP_Text classifiers. The improvement is especially dramatic for users with at least 20 previous questions. Interestingly, the simple strategy of grouping users by number of previous questions (ASP_Group) is even more effective, resulting in accuracy higher than both other methods for users with

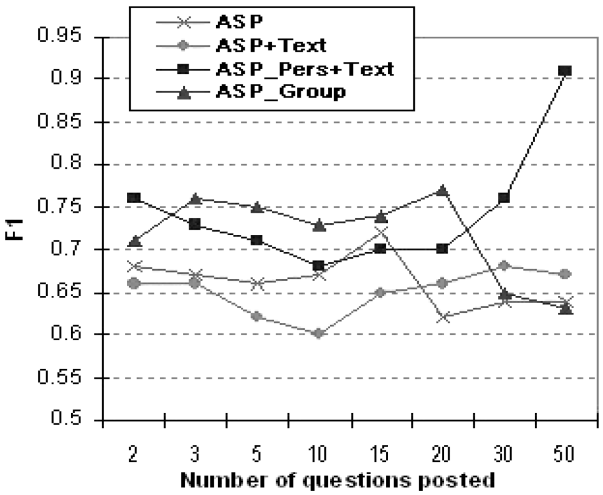

Fig. 13. F1 of ASP, ASP_Text, ASP_Pers+Text, and ASP_Group for predicting satisfaction of askers with varying number of questions.

moderate amount of history. Finally, for users with only two questions total (that is, only one previous question posted) the performance of ASP_Pers+Text is surprisingly high. We found that the classifier simply "memorizes" the outcome of the only available previous question, and uses it to predict the rating of the current question.

To better understand the improvement of personalized models, we report the most significant features, sorted by Information Gain (IG), for three representative ASP_Pers+Text models (Table XIII). Interestingly, whereas for Pers 1 and Pers 2, textual features such as "good luck" in the answer are significant, for Pers 3 nontextual features are most significant.

We also report the top 10 features with the highest information gain for the ASP and ASP_Group models (Table XIV). Interestingly, while asker's average previous rating is the top feature for ASP, the length of membership of the asker is the most important feature for ASP_Group, perhaps allowing the classifier to distinguish more expert users from the active newbies. In summary, we have demonstrated promising preliminary results on personalizing satisfaction prediction even with relatively simple personalization models.

8. RELATED WORK

Community Question Answering sites, such as Yahoo! Answers and Naver, have been gaining increasing popularity among many online users. Unlike in automatic question answering, the goal is not to develop a better algorithm for retrieving and extracting answers, but instead to enable the exchange of high-quality, relevant information between community participants. Finding such quality information, where in QA communities quality varies significantly [Su et al. 2007], provides a unique challenge, which recently has been addressed [Agarwal et al. 2008; Agichtein et al. 2008; Jeon et al. 2006, 2005].

Table XIII. Top 10 Features by Information Gain for Three Sample ASP_Pers+Text Models

Pers 1 (97 questions)	Pers 2 (49 questions)	Pers 3 (25 questions)
UH: Total_Ans_Received	Q: Avg_Pos_Votes	Q: Cont_KL_Div_TREC
UH: Ques_Resolved	”would” in answer	Q: Cont_KL_Div_Wikipedia
”good luck” in answer	”answer” in question	UH: Total_Ans_Received
”is an” in answer	”just” in answer	UH: Ques_Resolved
”want to” in answer	”me” in answer	Q: Cont_KL_Div_All_Cate
”we” in answer	”be” in answer	UH: Avg_Past_Rating
”want in” answer	”in the” in question	CA: Avg_Rating
”adenocarcinoma” in question	CA: History	“anybody” in question
”was” in question	”who is” in question	Q: Cont_Typo_Density
”live” in answer	”those” in answer	Q: Cont_Len

Table XIV. Top 10 features by Information Gain for ASP (trained for all askers) and ASP_Group (trained for the group of askers with 20 to 29 questions)

IG	ASP	IG	ASP_Group
0.104117	UH: Avg_Past_Rating	0.30981	UH: Member_Since
0.102117	UH: Most_Recent_Rating	0.25541	UH: Avg_Past_Rating
0.047222	QA: Avg_Pos_Vote	0.22556	UH: Most_Recent_Rating
0.041773	QA: Sum_Pos_Vote	0.15237	CA: Avg_Votes
0.041076	Q: Max_Pos_Vote	0.14466	CA: Avg_Close_Time
0.03535	QA: Most_Vote_Ans_Time_Diff	0.13489	CA: Avg_Rating
0.032261	UH: Member_Since	0.13175	CA: Avg_Ans_Arrival
0.031812	CA: Avg_Rating	0.12437	CA: Avg_Ques_Arrival
0.03001	CA: Avg_Ans_Count	0.09314	QA: Avg_Pos_Vote
0.029858	CA: Avg_Ans_Arrival	0.08572	CA: Avg_Ans_Count

Community question answering builds on the rich history in automatic question answering [Voorhees 2003] and web question answering [Brill et al. 2002]. However, a significant difference includes the large amount of metadata available to find relevant and high-quality content [Agichtein et al. 2008]. Additionally, while previous work focused on how to retrieve high quality answers from the CQA content, the question of information seeker satisfaction was not explored. In contrast, we present a general prediction model to investigate the ability of a QA community to provide satisfactory answers from the asker's perspective.

Our work is related to, but distinct from interactive Question Answering [Dang et al. 2007]. In particular, we can directly study the satisfaction from information seeker perspective. Nowhere does the problem of subjective relevance arise more prominently than in community QA, where many of the questions are inherently subjective, complex, ill-formed, or all of the above. While automatic complex QA has been an active area of research, ranging from simple modification to factoid QA technique [Soricut and Brill 2004] to knowledge intensive approaches for specific domains [Demner-Fushman and Lin 2007], the technology does not yet exist to automatically answer open domain, complex and subjective question. A corresponding problem is complex QA evaluation. Recent efforts at automatic evaluation show that even for well-defined, objective, complex questions, evaluation is extremely labor-intensive and has many

challenges [Lin and Demner-Fushman 2006; Lin and Zhang 2007]. The problem of subjective QA has only recently started to be addressed in the question answering community, most recently as the first opinion QA track in TREC [Dang et al. 2007]. We believe that this work can contribute to both the understanding of complex QA satisfaction, and explores important evaluation issues in a new setting. To our knowledge, this paper is the first large-scale study of real user satisfaction with obtaining information for complex and/or subjective information needs.

There is a rich tradition of relevance-based assessment of IR and QA (see Voorhees [2001]) for an overview). While the idea of relevance being inherently subjective has been pointed out by many researchers (e.g., see Zobel [1998] and more recently Ruthven et al. [2007]), we note that in community QA a large fraction of the questions are subjective, compounding the problem of both relevance assessment (which is no longer meaningful). Information seeker satisfaction has been studied in ad-hoc IR context in Harter and Hert [1997] (refer to Kobayashi and Takeda [2000] for an overview), but studies have been limited by lack of realistic user feedback on whole-result satisfaction and instead worked primarily within the Cranfield evaluation model.

Our work is also related to user modeling for Web search, where the goal is to predict which results will be relevant [Agichtein et al. 2006; White and Drucker 2007; White et al. 2007; Downey et al. 2007]; other uses include classifying user intent into a particular category [Rose and Levinson 2004]. This work builds on the influential user model introduced by Belkin [Belkin et al. 1982], [Belkin 1997]. Recently, eye tracking has started to emerge as a useful technology for understanding some of the mechanisms behind user behavior (e.g., Joachims et al. [2007], and Cutrell and Guan [2007], which may provide additional insight into user satisfaction with web search results. In contrast, we deal with complex information needs and community-provided answers (with explicit, noisy, "relevance" ratings from other users). Furthermore, we deal with subjective ratings provided by users themselves, instead of other assessors.

In order to predict asker satisfaction, we exploit standard classification techniques. Many models and techniques have been proposed for classification problem, including support vector machines, decision tree based techniques [Quinlan 1996] and boosting-based techniques [Freund and Schapire 1996]. We use these techniques to build our prediction models by using Weka [Witten and Frank 2005], a popular library of machine learning methods. In particular, we use the Weka's implementation of SMO [Platt 1998], AdaBoost [Freund and Schapire 1996], and an implementation of the C4.5 decision tree [Quinlan 1996].

To the best of our knowledge, ours is the first exploration of predicting user satisfaction in complex and subjective information seeking environments. While information seeker satisfaction has been studied in ad-hoc IR context (see Kobayashi and Takeda [2000] for an overview), previous studies have been limited by the lack of realistic user feedback. In contrast, we deal with complex information needs and community-provided answers, trying to predict subjective ratings provided by users themselves. Furthermore, while automatic complex QA has been an active area of research, ranging from simple

modification to factoid QA technique [Soricut and Brill 2004] to knowledge intensive approaches for specialized domains, the technology does not yet exist to automatically answer open domain, complex, and subjective questions. Hence, this article contributes to both the understanding of complex question answering, and explores evaluation issues in a new setting.

This paper combines and extends the work presented in Liu et al. [2008] and Liu and Agichtein [2008b]. The contribution of this paper is the added perspective of combining the results together, additional experimental results, and the expanded description and analysis of the experiments and the methods and features used.

9. CONCLUSIONS

Community Question Answering is rapidly growing popularity. However, the quality of answers, and the user satisfaction with the CQA experience, varies greatly. This article describes our work on predicting information seeker satisfaction in question answering communities. We introduced and formalized the problem of asker satisfaction prediction, and explored state-of-the-art classification techniques, and sophisticated lexical, semantic, and statistical features to implement our models. We have shown the importance of asker history to this highly personal, difficult, and subjective task, and demonstrated that our system can outperform human assessors, who do not benefit from knowing the prior asker history. We also reported results on personalizing satisfaction prediction, demonstrating significant accuracy improvements over a "one-size-fits-all" satisfaction prediction model, and identified unique challenges inherent to the CQA domain.

In summary, this paper outlines a promising area in the general field of modeling user intent, expectations, and satisfaction, and can potentially result in practical improvements to the effectiveness and design of question answering communities.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

ACKNOWLEDGMENTS

We thank the Yahoo! Answers team for allowing us extended use of the API, the Emory College Seed Fund for partially supporting this research.

REFERENCES

AGARWAL, N., LIU, H., TANG, L., AND YU, P. S. 2008. Identifying the influential bloggers in a community. In *Proceedings of 1st ACM International Conference on Web Search and Data Mining (WSDM)*. 207–218.

AGICHTEIN, E., BRILL, E., DUMAIS, S., AND RAGNO, R. 2006. Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

AGICHTEIN, E., CASTILLO, C., DONATO, D., GIONIS, A., AND MISHNE, G. 2008. Finding high-quality content in social media. In *Proceedings of 1st ACM International Conference on Web Search and Data Mining (WSDM)*. 183–194.

BELKIN, N., ODDY, R. N., AND BROOKS, H. M. 1982. Information retrieval: Part ii. results of a design study. *J. Documen. 38,* 3, 145–164.

BELKIN, N. J. 1997. User modeling in information retrieval. In *Proceedings of the 6th International Conference on User Modelling (UM'97)*.

BRILL, E., DUMAIS, S., AND BANKO, M. 2002. An analysis of the askmsr question-answering system. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

CUTRELL, E. AND GUAN, Z. 2007. Eye tracking in msn search: Investigating snippet length, target position and task types. *MSR-TR-2007-01*.

DANG, H. T., KELLY, D., AND LIN, J. 2007. Overview of the TREC 2007 question answering track. In *Proceedings of the 16th Text Retrieval Conference (TREC)*.

DEMNER-FUSHMAN, D. AND LIN, J. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Comput. Linguist. 33,* 1, 63–103.

DOWNEY, D., DUMAIS, S. T., AND HORVITZ, E. 2007. Models of searching and browsing: Languages, studies, and applications. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*.

FREUND, Y. AND SCHAPIRE, R. 1996. Experiments with a new boosting algorithm. In *Proceedings of the 13th International Conference on Machine Learning (ICML)*.

HARTER, S. P. AND HERT, C. A. 1997. Evaluation of information retrieval systems: Approaches, issues, and methods. *Annual Review of Information Science and Technology (ARIST) 32*, 3–94.

JEON, J., CROFT, W., AND LEE, J. 2005. Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM)*.

JEON, J., CROFT, W., LEE, J., AND PARK, S. 2006. A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

JOACHIMS, T., GRANKA, L., PAN, B., HEMBROOKE, H., RADLINSKI, F., AND GAY, G. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst. 25,* 2.

KOBAYASHI, M. AND TAKEDA, K. 2000. Information retrieval on the web. *ACM Compu. Surv. 32,* 2, 144–173.

LIN, J. AND DEMNER-FUSHMAN, D. 2006. Methods for automatically evaluating answers to complex questions. *Inform. Retrie. 9,* 5, 565–587.

LIN, J. AND ZHANG, P. 2007. Deconstructing nuggets: the stability and reliability of complex question answering evaluation. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 327–334.

LIU, Y. AND AGICHTEIN, E. 2008a. On the evolution of the yahoo! answers qa community. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 737–738.

LIU, Y. AND AGICHTEIN, E. 2008b. You've got answers: Towards personalized models for predicting success in community question answering. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*. 97–100.

LIU, Y., BIAN, J., AND AGICHTEIN, E. 2008. Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

PLATT, J. C. 1998. Fast training of support vector machines using sequential minimal optimization. *Advances in Kernal Methods—Support Vector Learning*, 185–208.

QUINLAN, J. 1996. Improved use of continuous attributes in c4.5. In *J. Artif. Intell. Resear.*

ROSE, D. E. AND LEVINSON, D. 2004. Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web (WWW)*.

RUTHVEN, I., GLASGOW, L. A., BAILLIE, M., BIERIG, R., NICOL, E., SWEENEY, S., AND YAKICI, M. 2007. Intra-assessor consistency in question answering. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 727–728.

SORICUT, R. AND BRILL, E. 2004. Automatic question answering: Beyond the factoid. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

SU, Q., PAVLOV, D., CHOW, J., AND BAKER, W. 2007. Internet-scale collection of human-reviewed data. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*.

VOORHEES, E. M. 2001. The philosophy of information retrieval evaluation. In *Proceedings of the Workshop of the Cross-Language Evaluation Forum (CLEF)*.

VOORHEES, E. M. 2003. Overview of the TREC 2003 question answering track. In *Proceedings of the 12th Text Retrieval Conference (TREC)*.

WHITE, R., BILENKO, M., AND CUCERZAN, S. 2007. Studying the use of popular destinations to enhance web search interaction. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

WHITE, R. W. AND DRUCKER, S. M. 2007. Investigating behavioral variability in web search. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*.

WITTEN, I. AND FRANK, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Morgan Kaufman.

ZOBEL, J. 1998. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 307–314.

Received September 2008; revised January 2009; accepted January 2009