Automatic Identification of Best Answers in Online Enquiry Communities

Grégoire Burel, Yulan He, and Harith Alani

Knowledge Media Institute,
The Open University, Milton Keynes, UK
{g.burel,y.he,h.alani}@open.ac.uk

Abstract. Online communities are prime sources of information. The Web is rich with forums and Question Answering (Q&A) communities where people go to seek answers to all kinds of questions. Most systems employ manual answer-rating procedures to encourage people to provide quality answers and to help users locate the best answers in a given thread. However, in the datasets we collected from three online communities, we found that half their threads lacked best answer markings. This stresses the need for methods to assess the quality of available answers to: 1) provide automated ratings to fill in for, or support, manually assigned ones, and; 2) to assist users when browsing such answers by filtering in potential best answers. In this paper, we collected data from three online communities and converted it to RDF based on the SIOC ontology. We then explored an approach for predicting best answers using a combination of content, user, and thread features. We show how the influence of such features on predicting best answers differs across communities. Further we demonstrate how certain features unique to some of our community systems can boost predictability of best answers.

Keywords: Social Semantic Web, Community Question Answering, Content Quality, Online Communities.

1 Introduction

Nowadays, online enquiry platforms and Question Answering (Q&A) websites represent an important source of knowledge for information seekers. According to Alexa,[1] 14% of Yahoo!'s traffic goes to its Q&A website whereas *Stack Exchange*[2] (SE) Q&A network boast an average of 3.7 million visits per day.

It is very common for popular Q&A websites to generate many replies for each posted question. In our datasets, we found that on average each question thread received 9 replies, with some questions attracting more than 100 answers. With such mass of content, it becomes vital for online community platforms to put in place efficient policies and procedures to allow the discovery of best answers. This allows community members to quickly find prime answers, and

[1] Alexa, http://www.alexa.com
[2] Stack Exchange, http://stackexchange.com

E. Simperl et al. (Eds.): ESWC 2012, LNCS 7295, pp. 514–529, 2012.

to reward those who provide quality content. The process adopted by Q&A systems for rating best answers range from restricting answer ratings to the author of the question (e.g. the *SAP Community Network*[3] (SCN) forums), to opening it up to all community members (e.g. SE). What is common between most of such communities is that the process of marking best answers is almost entirely manual. The side effect is that many threads are left without any such markings. In our datasets, about 50% of the threads lacked pointers to best answer. Although, much research has investigated the automatic assessment of answer quality and the identification of best answers [1], little work has been devoted to the comparison of such models across different communities.

In this paper we apply a model for identifying best answers on three different enquiry communities: the *SCN forums* (SCN), *Server Fault*[4] (SF) and the *Cooking* community[5] (CO). We test our model using various combinations of *user*, *content*, and *thread* features to discover how such groups of features influence best answer identification. We also study the impact of community-specific features to evaluate how platform design impacts best answers identification. Accordingly, the main contributions of our paper are:

1. Perform a comparative study on performance of a typical model for best answer identification on three online enquiry communities.
2. Introduce a new set of features based on the characteristics and structure of Q&A threads.
3. Study the influence of user, content, and thread features on best answer identification and show how combining these features increases accuracy of best answer identification.
4. Investigate the impact of platform-specific features on performance of best answer identification, and demonstrate the value of public ratings for best answer prediction.

In addition to the above contributions, we also developed an ontology for representing Q&A features as well as a methodology for converting and extending our model using augmentation functions. Furthermore, we introduce several ratio features, e.g. *ratio of scores of an answers in comparison to others*. We show that such ratio features have a good impact on our model.

In the following section we analyse existing research in *best answer identification*. In the third section, the features used by our model are introduced. Following the *user*, *content* and *thread* features introduction, we present an ontology based methodology for mapping and generating our three different dataset features. The fourth section describes our best answer model an presents our results. The results and future work are discussed in section five. Finally, we conclude our paper in section six.

[3] SAP Community Network, `http://scn.sap.com`
[4] Server Fault, `http://serverfault.com`
[5] Cooking community, `http://cooking.stackexchange.com`

2 Related Work

Many different approaches have been investigated for assessing content quality on various social media platforms [1]. Most of those approaches are based on estimating content quality from two groups of features; *content* features, and *user* attributes. Content-based quality estimation postulates that the content and the metadata associated with a particular answer can be used for deriving the value of an answer. While user features considers that behavioural information about answerers is relevant for identifying the merit of a post.

Content based assessment of quality has been applied to both textual [2,3,4] and non textual [5,2,3,4] content. Textual features normally include readability measures such as the *Gunning-Fog index*, *n-grams* or *words overlap* [3,4]. Content metadata like *ratings*, *length* and *creation date* [5,2,3,4] have also been investigated in this context. In this paper we also use common content features, such as *content length* and*Gunning-Fog index*, alongside user features and other novel features related to the online community platform.

Some approaches for assessing answer quality rely on assessing the expertise or importance of the users themselves who provided the answers. Such assessment is usually performed by applying link based algorithms such as *ExpertiseRank* [6] which incorporates user expertise with *PageRank*, and *HITS* for measuring popularity or connectivity of users [7,3,8].

Another line of research focused on identifying existing answers to new questions on Q&A systems. Ontologies and Natural Language Processing (NLP) methods have been proposed for extracting relevant entities from questions and matching them to existing answers [9,10,11,12]. Other methods involved more standard Information Retrieval (IR) techniques like Probabilistic Latent Semantic Analysis[13], query rewriting [14] and translation models [5]. Most of these works however focus on measuring the relevance of answers to questions, rather than on the quality of those answers. Other approaches analyse the role of posts, to distinguish between conversational and informational questions [15], or between questions, acknowledges, and answers [16]. Although out of the scope of this paper, such approaches could be used to filter out non-answer posts from discussion threads that could improve best answer prediction.

Our work differs from all the above in that in addition to using common content and user features, we also use thread features that take into account certain characterises of the individual threads; such as scores ratios, order of answers, etc. In addition to those features, we also present a contextual topical reputation model for estimating how knowledgeable the answerer is likely to be. Also, much of previous work concentrated on studying single communities, whereas in this paper we investigate and compare the results across three communities, thus establishing a better idea of how generic the findings are.

3 Predicting Quality of Answers

Measuring content quality and identifying best answers require the training and validation of prediction models and discovering the influence of the various

features on these predictions. For training our answer classifier, we use three main types of features; *content*, *user*, and *thread* features. All these features are strictly generated from the information available at the time of the feature extraction (i.e. future information are not taken into account while generating attributes). The different attributes are described in the following sections.

3.1 User Features

User features describe the characteristic and reputation of authors of questions and answers. Below is the list of 11 user features employed in this study.

- *Reputation*: Represents how active and knowledgeable a user is. It can be approximated from the number of good answers written by the user.
- *Age*: The user age. It measures how old is a user in years.
- *Post Rate*: Average number of questions or answers the user posts per day.
- *Number of Answers*: The number of answers posted by a user.
- *Answers Ratio*: The proportion of answers posted by a user.
- *Number of Best Answers*: The number of best answers posted by a user.
- *Best Answers Ratio*: The proportion of best answers posted by a user.
- *Number of Questions*: The number of questions posted by a user.
- *Questions Ratio*: The proportion of questions posted by a user.
- *Normalised Activity Entropy*: A normalised entropy measure (H_a) represents how predictable is the activity of a user. In enquiry platforms, a user u_i can either post questions (Q) or answers (A). Lower entropy indicates focus on one activity. The normalised activity entropy is calculated from the probabilities of a user posting answers or questions:

$$H_A(u_i) = -\frac{1}{2}\left(P(Q|u_i)\log P(Q|u_i) + P(A|u_i)\log P(A|u_i)\right) \tag{1}$$

- *Normalised Topic Entropy*: Calculates the concentration (H_T) of a user's posts across different topics. Low entropy indicates focus on particular topics. In our case, topics are given by the tags associated with a question or the category of the post. Each user's tags T_{u_i} are derived from the topics attached to the questions asked or answered by the user. This can be used to calculate the probability $P(t_j|u_i)$ of having a topic t_j given a user u_i:

$$H_T(u_i) = -\frac{1}{|T_{u_i}|}\sum_{j=1}^{|T_{u_i}|} P(t_j|u_i)\log P(t_j|u_i) \tag{2}$$

- *Topical Reputation*: A measure of the user's reputation with a particular post. It is derived from the topics T_{q_k} associated with the question q_k for which the post belongs. By adding the score values of each user's answers $S(a)$, where $a \in A_{u_i,t_j}$, about a particular topic t_j, we obtain the general user topical reputation $E_{u_i}(t_j)$ for a particular topic. Given a post user u_i,

the user topical reputation function E_{u_i} and a question q with a set of topics T_q, the reputation embedded within a post related to question q is given by:

$$E_P(q, u_i) = \sum_{j=1}^{|T_q|} E_{u_i}(t_j) \tag{3}$$

$$E_{u_i}(t_j) = \sum_{a \in A_{u_i,t_j}} S(a) \tag{4}$$

3.2 Content Features

Content features represent the attributes of questions and answers, and can be used for estimating the quality of a particular question or answer as well as their importance. We use the following content features in our analysis:

- *Score:* Represents the rating of an answer, and it normally collected from users in the form of *votes* or *thumbs up/thumbs down* flags.
- *Answer Age:* Difference between the question creation date and the date of the answer.
- *Number of Question Views:* The number of views or hits on a question.
- *Number of Comments:* The number of comments associated with a post.
- *Number of Words:* The number of words contained in a post.
- *Readability with Gunning Fog Index*: Used to measure post readability using the Gunning index of a post p_i which is calculated using the average sentence length asl_{p_i} and the percentage of complex words pcw_{p_i}:

$$G_{p_i}(asl_{p_i}, pcw_{p_i}) = 0.4\ (asl_{p_i} + pcw_{p_i}) \tag{5}$$

- *Readability with Flesch-Kincaid Grade*: Calculated from the average number of words per sentence $awps_{p_i}$ and average number of syllables per word $aspw_{p_i}$ of a post p_i:

$$FK_{p_i}(awps_{p_i}, aspw_{p_i}) = 0.39\ awps_{p_i} + 11.8\ aspw_{p_i} - 15.59 \tag{6}$$

3.3 Thread Features

Our final set of features represents relations between answers in a particular thread. Each question tends to have more than one answer and most Q&A platforms allow only one answer to be selected as the best answer. As a consequence, each answer competes for being the best answer. In such context, relational features such as the proportion of votes to a particular answer can be used for estimating the relative importance of a particular post.

- *Score Ratio*: The proportion of scores given to a post from all the scores received in a question thread.
- *Number of Answers*: Number of answers received by a particular question.

– *Answer position*: The absolute order location of a given answer within a question thread (e.g. first, second).
– *Relative Answer Position*: The relative position of an answer within a post thread. Given a question q, its answers a_q, and the position of an answer $pos_{a_{q_i}}$, the relative answer position of an answer a_{q_i} is given by:

$$RP(a_{q_i}) = 1 - \frac{pos_{a_{q_i}}}{|a_q|} \tag{7}$$

– *Topical Reputation Ratio*: The proportion of topical reputation associated with a particular answer. Given the sum of topical reputation of all the answers, the ratio of topical reputation attributed to a particular answer.

3.4 Core vs Extended Feature Sets

As mentioned, we want to investigate the impact of platform-specific features on predictability of best answers. Hence the features above contain some that are not common across our datasets. For example, in SCN only the owner of a question can rate its answers, and select the best answer, whereas in SF and CO communities anyone with over 200 points of reputation can vote for any answer, and hence the selections of best answers can emerge collectively. The platform that supports SF and CO offer more features than SCN. In Table 3.4 we list the *core features set*, which is shared across all three datasets, and the *extended features set*, which is only valid for SF and CO datasets.

Table 1. Differences between the Core Features Set and the Extended Features Set

	Features Set	
Type	Core Features Set (19)	Extended Features Set† (23)
User	*Reputation, Post Rate, Normalised Activity Entropy, Number of Answers, Answers Ratio, Number of Best Answers, Best Answers Ratio, Number of Questions, Questions Ratio, Normalised Topic Entropy, Topical Reputation.* (10)	*Reputation,* ***Age****, Post Rate, Normalised Activity Entropy, Number of Answers, Answers Ratio, Number of Best Answers, Best Answers Ratio, Number of Questions, Questions Ratio, Normalised Topic Entropy, Topical Reputation.* (11)
Content	*Answer Age, Number of Question Views, Number of Words, Gunning Fog Index, Flesch-Kinkaid Grade Level.* (5)	***Score****, Answer Age, Number of Question Views,* ***Number of Comments****, Number of Words, Gunning Fog Index, Flesch-Kinkaid Grade Level.* (7)
Thread	*Number of Answers, Answer Position, Relative Answer Position, Topical Reputation Ratio.* (4)	***Score****, Number of Answers, Answer Position, Relative Answer Position, Topical Reputation Ratio.* (5)

†Only valid for the *Server Fault* and *Cooking* datasets.

4 Datasets

Our experiments are conducted on three different datasets. The first two are subs communities extracted from the April 2011 *Stack Exchange* (SE) public datasets:[6] *Server Fault* (SF) user group and the non technical *Cooking* website

[6] As part of the public *Stack Exchange* dataset, the *Server Fault* and *Cooking* datasets are available online at `http://www.clearbits.net/get/1698-apr-2011.torrent`

(CO) composed of cooking enthusiasts. The other dataset is obtained from the *SAP Community Network* (SCN) forums and consists of posts submitted to 33 different forums between December 2003 and July 2011.[7]

4.1 SAP Community Network

The *SAP Community Network* (SCN) is a set of forums designated for supporting SAP customers and developers. SCN integrates traditional Q&A functionalities systems such as best answer selection, user reputation and moderation. Each SCN thread is initiated with a question and each answer in that thread is a reply to that question. Thread authors can assign a limited number of points to the answers they like (unlimited two-points for *helpful answers*, two sets of six-points for *very helpful answers* and one ten-points for the *best answer*). Points given to answers add to the reputation of their authors. Users can be flagged as topic experts, get promoted to moderators, or be invited to particular SAP events if their online reputation is high.

Our dataset consists of 95,015 threads and 427,221 posts divided between 32,942 users collected from 33 different forums between December 2003 and July 2011. Within those threads, we only select threads that have best answers. Our final dataset consists of 29,960 (32%) questions and 111,719 (26%) answers.

4.2 Server Fault

Server Fault (SF) is a Q&A community of IT support professionals and is hosted on the SE platform. SE provides social features such as *voting*, *reputation* and *best answer selection* while making sure that each posted answer is self-contained. However, SF differences reside in its rewarding program where each user gains access to additional features like ability to vote and advertising removal depending on their reputation.

Compared to SCN, SF editing policy is completely community driven. Depending on the user reputation, each community member is allowed to refine other people's questions and answers. Hence, instead of adding additional posts for elaborating questions or answers, SF users can directly edit existing content.

To keep the community engaged, the SF platform offers rewards and badges for various type of contributions. For example, users can earn the *Autobiographer* badge if they fill their profiles completely. SF users' reputation is calculated from the votes that have been cast on a particular question or answer. For each post, community members vote up or down depending on the quality and usefulness that is then pushed to the post owner. As community members gain/lose reputation, they gain/lose particular levels and abilities. Our SF dataset is extracted from the April 2011 public dataset, and consists of 71,962 questions, 162,401 answers and 51,727 users. Within those questions we selected only the questions that have best answers. The final SF dataset consist of 36,717 (51%) questions and 95,367 (59%) answers.

[7] SAP is planning to migrate their community to a new platform in February 2012, with several new features that were not available at the time of our data collection.

4.3 Cooking Websites

The *Cooking* community (CO) is composed of enthusiasts seeking cooking advice and recipes. It is also hosted on the SE platform and thus shares the same attributes and functionalities as SF above. CO is a smaller dataset with 3,065 questions, 9,820 answers and 4,941 users. Similarly to the other datasets, we only select the questions that have best answers. The final dataset is composed of 2,154 (70%) questions and 7,039 (72%) answers.

4.4 Features Inferencing and Representation

Our three datasets come in different formats and structures. To facilitate their representation, integration, and analysis, we converted all three datasets to a common RDF format and structure (Figure1). Data is dumped into an SQL database (1) then converted to RDF based on SIOC[8] ontology using the D2RQ[9] (2). RDF is then loaded into a triple store where knowledge augmentation functions are executed (3). Such functions simply extend the knowledge graph of each dataset by adding additional statements and properties (i.e. topical reputation, answer length, votes ratio, etc.). This workflow serves as the input of the learning algorithms used for predicting content quality (4). We extended SIOC to represent Q&A vocabulary[10]. The flexibility of RDF enabled us to add features without requiring schema redesign. Summary of mappings of our datasets to SIOC classes is illustrated in Table 5.

5 Best Answer Identification

Ability to accurately identify best answers automatically is not only a compliment to the fitness and preciseness of the prediction model, but also to the fit of the community and platform features that are enabling such task to be performed accurately. If a platform fails to support the gathering of information that correlates with content quality, then automating content quality prediction becomes much harder. More importantly, such difficulty will also be faced by the users who need to quickly find the best solving answers to their problems.

The experiment described next aims at measuring the importance of our core and extended feature sets for best answer prediction, as well as highlighting how each feature impacts prediction accuracy in a given platform.

5.1 Experimental Setting

In our experiments we train a categorical learning model for identifying the best answers in our three datasets. For each thread, the *best answer* annotation is used for training and validating the model. Because SCN *best answer* annotation is

[8] SIOC Ontology, `http://sioc-project.org`
[9] D2RQ Platform, `http://www4.wiwiss.fu-berlin.de/bizer/d2rq`
[10] Q&A Vocabulary, `http://purl.org/net/qa/ns#`

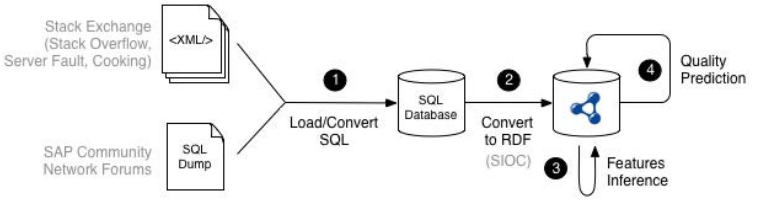

Fig. 1. Dataset Conversion and Inferencing Workflow

Table 2. SIOC Class Mappings of the *Stack Exchange* and *SCN Forums* Datasets

Input Dataset		
SCN	SF and CO	RDF Output
`User`	`User`	`sioc:OnlineAccount/foaf:Person`
`Thread` (first thread Post)	`Question`	`sioct:Question`
`Post` (not in first position)	`Answer`	`sioct:Answer`
`Post` (with 10 points)	`Best Answer`	`sioct:BestAnswer`
-	`Comment`	`sioct:Comment`
`Forum`	`Tag`	`sioct:Tag` (topic)

based on the author ratings, we use the *best answer* rating (i.e. 10) as the model class and discard the other ratings (i.e. 2 and 6) for training the SCN model.

A standard 10-folds cross validation scheme is applied for evaluating the generated model. Each model uses the features described earlier in the paper. Decision tree algorithms have been found to be the most successful in such contexts [3,17]. We use the *Multi-Class Alternating Decision Tree* learning algorithm due to its consistent and superior results to other decision tree algorithms we tested (*J48*, *Random Forests*, *Alternating Tree* and *Random Trees*).

To evaluate the performance of the learning algorithm, we use precision (P), recall (R) and the harmonic mean F-measure (F_1) as well as the area under the Receiver Operator Curve (ROC) measure. The precision measure represents the proportion of retrieved best answers that were real best answers. Recall measures the proportion of best answers that were successfully retrieved. We also plot the ROC curve and use the Area Under the Curve (AUC) metrics for estimating the classifier accuracy.

We run two experiments, the first compare the performance of our model for identifying best answers across all three datasets, using the core and extended feature sets. The second experiment focuses on evaluating the influence of each features on best answers identification.

5.2 Results: Model Comparison

For our first experiment, we train the *Multi-Class Alternating Decision Tree* classifier on different features subsets and compare the results using the metrics that we described in the previous section (Table 5.2).

Baseline Models: We used the *number of words* feature to train a baseline model since it was argued to be a good predictor [5,3]. Additionally, for the SF

Table 3. Average *Precision*, *Recall*, F_1 and *AUC* for the *SCN Forums*, *Server Fault* and *Cooking* datasets for different feature sets and extended features sets (marked with +) using the *Multi-Class Alternating Decision Tree* classifier

	SCN Forums				Server Fault				Cooking			
Feature	*P*	*R*	F_1	*AUC*	*P*	*R*	F_1	*AUC*	*P*	*R*	F_1	*AUC*
Words	0.536	0.732	0.619	0.616	0.592	0.621	0.537	0.567	0.671	0.705	0.644	0.651
Answer Score	-	-	-	-	0.643	0.656	0.625	0.673	0.751	0.760	0.753	0.797
Answer Score Ratio	-	-	-	-	0.808	0.809	0.806	0.848	0.866	0.868	0.866	0.916
Users	0.716	0.746	0.687	0.752	0.637	0.651	0.626	0.664	0.687	0.714	0.681	0.686
Content	0.712	0.740	0.659	0.678	0.647	0.659	0.628	0.679	0.708	0.727	0.707	0.754
Thread	0.820	0.827	0.817	0.865	0.753	0.756	0.749	0.809	0.765	0.772	0.751	0.785
All	0.833	0.839	0.831	0.880	0.770	0.769	0.760	0.827	0.777	0.784	0.767	0.816
Users+	-	-	-	-	0.637	0.651	0.626	0.664	0.687	0.714	0.681	0.686
Content+	-	-	-	-	0.700	0.707	0.699	0.761	0.788	0.793	0.789	0.842
Thread+	-	-	-	-	0.844	0.845	0.844	0.910	0.867	0.869	0.867	0.919
All+	-	-	-	-	0.848	0.847	0.844	0.912	0.870	0.872	0.870	0.919

and CO datasets, we also train another basic model based on *answer scores* and *answer scores ratios* since such features are normally especially designed as a rating of content quality and usefulness.

Surprisingly, our results from all three datasets do not confirm previous research on the importance of content length for quality prediction. For each of our datasets, *precision* and *recall* were very low with a F_1 median of 0.619 (SCN: 0.619/SF: 0.537/Cooking: 0.644). This might be due to the difference of our data to those from literature which were taken from general Q&A communities such as Yahoo! Answers [3] and the Naver community [5]).

The SF and CO models trained on the *answer scores* highlight positive correlations between *best answers* and *scores*. However, this positive influence is reduced when the data grow in SF over CO. CO shows high F_1 results with 0.753 with Answer Score, whereas SF result is 0.625. Training the SE models on *Answer Score Ratios* shows even higher results with a F_1 of 0.806 for SF and 0.866 for *Cooking*. Overall, *answer score ratio* appear to be a good predicator for answer quality which shows that SF and CO collaborative voting models are effective. In particular, it shows that taking into account the relative voting proportions between answers (i.e. *scores ratio*) is a better approach than only considering absolute *scores*.

Core Features Models: Here we focus on the comparison of feature types (i.e. *users*, *content* and *threads*) and the impact of using the extended feature set on the identification process. We trained a model for each dataset and features set. Results in Table 5.2 show that using the thread features we introduced in this paper increases accuracy in all three datasets over user and content features. Results also show that F_1 when combining all core user, content, and thread features was 11%, 9.3%, and 5.4% higher for SCN, SF, and CO respectively, than the best F_1 achieved when using these features individually.

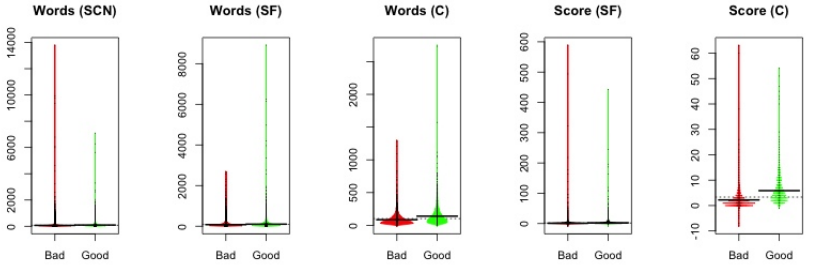

Fig. 2. Bean Plots representing the distribution of different features and best answers for the *SCN Forums* (SCN), the *Server Fault* (SF) and *Cooking* (C) datasets

Overall, when using all the core features (common to all datasets), SCN performed better than SF (+7.1%) and CO (+6.4%). Predictions for CO were slightly more accurate than for SF, probably due to its smaller size. However, results in Table 5.2 show that F_1 with all core features is lower than the *Answer Score Ratio* by 4.6% for SF and 9.9% for CO. This reflects the value of this particular feature for best answer identification on such platforms.

Figure 2 shows the distributions of best answers (good) and non-best answers (answer) for posts length for all our datasets and answer scores for SF and CO. Best answers seem to likely be shorter in SCN, and longer in SF and CO. This variation could be driven by the difference in data sizes and topics as well as external factors such as community policies (e.g. community editing in SE).

Extended Features Models: Now we recompute the models using extended *users*, *content* and *threads* feature sets. Remember that the extended features (Table 3.4) are only supported by SF and CO. No change in accuracy can be witnessed when extending the user features. However, F_1 increases by average of 8.3% for SF and 14.9% for CO when extending content and thread features. The only difference between the core user features and extended ones is the *user age* attribute. Hence the age of the answerer does not seem to have an effect on best answers identification. As for extended content and thread features, they contain extra features such as *number of comments* and *scores*, as well as the *scores ratios* which we compute per thread.

Table 5.2 shows that the F_1 for SF and CO when using all extended features combined (*All+* in 5.2) has increased by 8.4% and 10.3% for SF and CO respectively over using core features (*All* row in Table 5.2). This is mainly due to the addition of the *scores/ratings* based features. Furthermore, F_1 from the combined extended features was even higher than the Answer Score Ratio model, by 3.8% for SF and a mere 0.4% for CO.

In general, we can see that thread features are consistently more beneficial than others for identifying best answers. When available, scoring (or rating) features improve prediction results significantly, which demonstrates the value of community feedback and reputation for identifying valuable answers.

5.3 Results: Feature Comparison

Following on from the previous experiments, our second round of analysis focus on evaluating the importance of each feature for best answer identification. For each dataset, we rank all our predictors using Information Gain Ration (IGR) with respect to the best answers annotations. The top 15 are shown in Table 5.3.

Table 4. Top features ranked by Information Gain Ratio for the *SCN*, *Server Fault* and *Cooking* datasets. Type of feature is indicated by $U/C/T$ for *User*/*Content*/*Thread*

	SCN		Server Fault		Cooking	
R.	IG	Feature	IG	Feature	IG	Feature
1	0.217	*Topic. Rep. Ratio (T)*	0.332	***Score Ratio (T)***	0.430	***Score Ratio (T)***
2	0.196	*No. Answers (T)*	0.275	*No. Answers (T)*	0.190	***Score (C)***
3	0.108	*Bests Ratio (U)*	0.126	*Answer Position (T)*	0.164	*No. Answers (T)*
4	0.105	*Questions Ratio (U)*	0.117	*Topic. Rep. Ratio (T)*	0.120	*Answer Position (T)*
5	0.105	*Answers Ratio (U)*	0.097	*Relative Position (T)*	0.083	*Topic. Rep. Ratio (T)*
6	0.104	*Relative Position (T)*	0.070	***Score (C)***	0.074	*Bests Ratio (U)*
7	0.097	*Reputation (U)*	0.056	*Q. Views (C)*	0.070	*No. Bests (U)*
8	0.093	*Topic. Rep. (U)*	0.046	*Bests Ratio (U)*	0.069	*Reputation (U)*
9	0.090	*No. Bests (U)*	0.037	***No. Comments (C)***	0.065	*Answer Age (C)*
10	0.089	*Activity Entropy (U)*	0.022	*Topic Entropy (U)*	0.055	*Topic Entropy (U)*
11	0.064	*Answer Position (T)*	0.021	*Answer Age (C)*	0.054	***No. Comments (C)***
12	0.048	*No. Answers (U)*	0.019	*Post Rate (U)*	0.054	*No. Words (C)*
13	0.035	*Topic Entropy (U)*	0.018	*Reputation (U)*	0.053	*No. Answers (U)*
14	0.033	*Q. Views (C)*	0.017	*No. Bests (U)*	0.045	*Relative Position (T)*
15	0.027	*No. Words (C)*	0.016	*No. Answers (U)*	0.039	*Topic. Rep. (U)*

Core Features: First we focus the analysis on the *core features set*. Table 5.3 shows that SCN's most important feature for best answer identification appear to be the *topical reputation ratio*, which also came high up the list with 3rd rank in SF and 5th in CO. The *number of answers* also comes high in each dataset: 2nd for SCN and SF, and 3rd for CO. Note that our training datasets only contained threads with best answers. Hence the shorter the thread is (i.e. less answers) the easier it is to identify the best answer. Similarly, *best answers ratio* and *number of best answers* also proved to be good features for best answer prediction. Figure 3 shows the correlations with best answers (good) and non-best answers (bad) for the top five features in each datasets.

Distribution of SCN *topical reputation* in Figure 3 is narrower than the distribution of SF and CO. This highlights the difference between the SCN and SE reputation models. Contrary to SE, SCN only allow positive reputation. For core features, SF, CO, and SCN have a generally similar mode of operation. However, SCN is less affected by *answer position* due to the difference of platform editing policies. SE favours small thread whereas SCN does not. Such difference leads to a better correlation of *number of answers* with best answers in SE.

According to Table 5.3, user features appear to be dominant, with some thread features amongst the most influential. Number of thread answers and historical activities of users are particularly useful (e.g. number and ratio of user's best answers). User reputation in SCN plays a more important role than in SF and

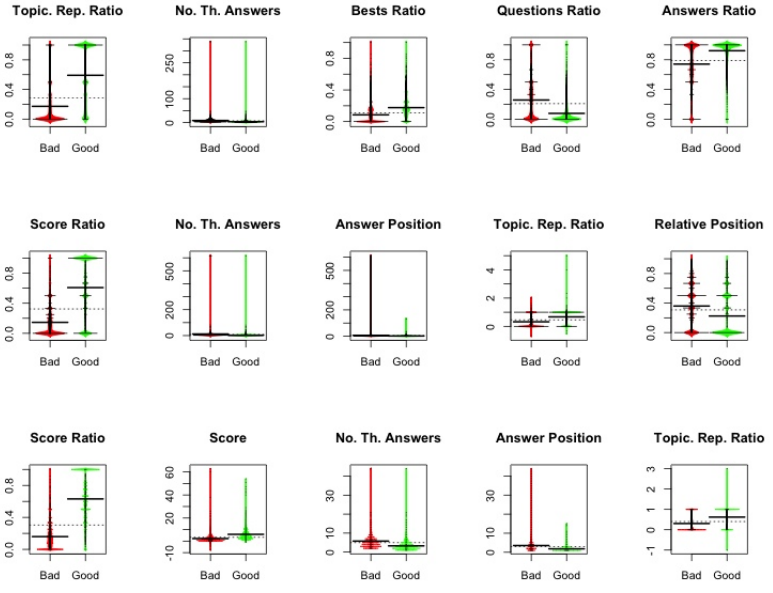

Fig. 3. Bean Plots representing the distribution of different the top five features for the *SCN Forums* (first row), the *Server Fault* (second row) and *Cooking* (third row) datasets

CO, which is probably a reflection of the community policies that puts emphasis on members' reputation.

In SAP's SCN, user activity focus seems to play a notable role (*topical reputation*, *answer and question ratios*, *activity entropy*, etc.). These features are further down the list for SF and CO.

Extended Features: The evaluation of extended features establishes the importance of *scores*. For SF and CO datasets, it is clear once again that the *score* features are the most important for identifying best answers.

SF has a *score ratio* IG of 0.332 and CO have IG score of 0.430 representing respectively around +5.7% and +24% more gain than the second ranked feature.

As in the general model evaluation, thread features compare the score of a single answer with the score of other thread answers. The higher the ratio, the better the answer. Note that the selection of best answers in SF and CO is left to the user who posted the question, who may or may not consider the scores given by the community or general site visitors.

6 Discussion and Future Work

Understanding which features impact best answer identification is important for improving community platform designs. However, different types of online communities tend to have different characteristics, goals, and behaviours. It is

therefore difficult to generalise any findings without a broad base of experimentation. Such observation is reinforced with the difference between our findings and previous research [5,3] concerning the value of content length. In this work, we widened our analysis to three communities to give our findings more scope. Our communities bear much similarity in terms of type, goals, and properties, and hence we can argue that our findings are transferable across such communities.

Reliable automatic identification of best answers can solve the common problem of scarcity of such valuable information in most online Q&A communities. However, automated methods must also find out when no best answers exist in a given thread. To this end, we are developing measures of answer quality, and will test them on threads with and without best answers. This will help ensuring that no best answers are enforced when none are above a certain quality.

Identifying best answers becomes more important the longer the threads are. It might be worth focusing such analysis on threads with more than one answer. It is worth mentioning than in SCN, SF, and CO datasets, the median number of answers per thread was 5,3, and 4 respectively, with averages of 13, 8.5, 5.

For the SF and CO communities, we showed that the ratings given by community members to existing answers were good predictors of best answers. Although only the authors of questions can currently pick the best answers, their choices seem to be positively correlated with those of the public. Our results showed that the accuracy of using public ratings for best answer selection can be improved further when other features are considered. SCN currently lacks this feature altogether. Interestingly, SAP' SCN is migrating itself to the Jive Engage platform[11] in 2012. Jive offers many social features, including collaborative rating of answers.

7 Conclusions

Many popular online enquiry communities receive thousands of questions and answers on daily basis. Our work identified that around 50% of posted questions do not have best answers annotations, thus forcing site visitors to check all existing answers for identifying correct answers. We studied three online Q&A communities to learn about the influence of the various features they have on our automated best answer identification model which is based on a wide selection of *user*, *content* and *thread* features. Some of those features were common across all three communities, and some were community-specific. We achieved 83% accuracy with SCN community, 84% with SF and 87% with CO.

We found out that contrary to previous work [5,3], *answer length* seems uncorrelated with best answers. We also discovered that best answers in communities that support community-based answer ratings (i.e. SF and CO) can be identified much more accurately, with over 0.8 F_1 using this feature alone *(answer score ratio)*. Our thread-based features proved to be very influential for best answer identification in all three communities.

[11] Jive Software, `http://jivesoftware.com`

Acknowledgments. This work is funded by the EC-FP7 project Robust (grant number 257859). The authors would like to thank SE for publicly sharing their data. Also thanks to Adrian Mocan from SAP for providing SCN data to ROBUST project.

References

1. Chai, K., Potdar, V., Dillon, T.: Content quality assessment related frameworks for social media. In: Proc. Int. Conf. on Computational Science and Its Applications (ICCSA), Heidelberg (2009)
2. Liu, Y., Agichtein, E.: You've got answers: towards personalized models for predicting success in community question answering. In: Proc. 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Columbus, Ohio (2008)
3. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: First ACM Int. Conf. on Web Search and Data Mining, Palo Alto, CA (2008)
4. Bian, J., Liu, Y., Zhou, D., Agichtein, E., Zha, H.: Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In: Int. World Wide Web Conf., Madrid (2009)
5. Jeon, J., Croft, W.B., Lee, J.H., Park, S.: A framework to predict the quality of answers with non-textual features. In: SIGIR, Washington. ACM Press (2006)
6. Zhang, J., Ackerman, M., Adamic, L.: Expertise networks in online communities: structure and algorithms. In: Proc. 16th Int. World Wide Web Conf., Banff (2007)
7. Jurczyk, P., Agichtein, E.: Discovering authorities in question answer communities by using link analysis. In: ACM 16th Conf. Information and Knowledge Management, CIKM 2007 (2007)
8. Suryanto, M., Lim, E., Sun, A., Chiang, R.: Quality-aware collaborative question answering: methods and evaluation. In: Proc. 2nd ACM Int. Conf. on Web Search and Data Mining, Barcelona (2009)
9. McGuinness, D.: Question answering on the semantic web. IEEE Intelligent Systems 19(1) (2004)
10. Narayanan, S., Harabagiu, S.: Question answering based on semantic structures. In: Proc. 20th Int. Conf. on Computational Linguistics, Geneva (2004)
11. Lopez, V., Pasin, M., Motta, E.: AquaLog: An Ontology-Portable Question Answering System for the Semantic Web. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 546–562. Springer, Heidelberg (2005)
12. Wang, Y., Wang, W., Huang, C.: Enhanced semantic question answering system for e-learning environment. In: 21st Int. Conf. on Advanced Information Networking and Applications Workshops, AINAW, vol. 2 (2007)
13. Qu, M., Qiu, G., He, X., Zhang, C., Wu, H., Bu, J., Chen, C.: Probabilistic question recommendation for question answering communities. In: Proc. 18th Int. World Wide Web Conf., Madrid (2009)
14. Kwok, C., Etzioni, O., Weld, D.: Scaling question answering to the web. ACM Transactions on Information Systems (TOIS) 19(3) (2001)

15. Harper, F., Moy, D., Konstan, J.: Facts or friends?: distinguishing informational and conversational questions in social Q&A sites. In: Proc. 27th Int. Conf. on Human Factors in Computing Systems, CHI, Boston, MA (2009)
16. Kang, J., Kim, J.: Analyzing answers in threaded discussions using a Role-Based information network. In: Proc. IEEE Int. Conf. Social Computing, Boston, MA (2011)
17. Rowe, M., Angeletou, S., Alani, H.: Predicting Discussions on the Social Semantic Web. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part II. LNCS, vol. 6644, pp. 405–420. Springer, Heidelberg (2011)