Predicting Quality of Answer in Collaborative Q/A Community

Kohei Arai [1]
Graduate School of Science and Engineering
Saga University
Saga City, Japan

ANIK Nur Handayani[1,2]
[2] Electrical and Information Technology
State University of Malang
Malang, Indonesia

***Abstract*— Community Question Answering (CQA) services have emerged allowing information seekers pose their information need which is questions and receive answers from their fellow users, also participate in evaluating the questions or answers in a variety of topics. Within this community information seekers could interact and get information from a wide range of users, forming a heterogeneous social networks and interaction between users. A question may receive multiple answers from multiple users and the asker or the fellow users could choose the best answer. Freedom and convenience in participation, led to the diversity of the information. In this paper we present a general model to predict quality of information in a CQA by using non textual features. We showing and testing our quality measurement to a collection of question and answer pairs. In the future our models and predictions could be useful for predictor quality information as a recommender system to complete a collaborative learning.**

Keywords—component; Collaborative; diversity of information; questions; answers; predict; non-textual feature

I. INTRODUCTION

Community Question Answering (CQA) has recently become available for information seekers. Beside web search engines, information seekers today have an option to inform their questions on CQA sites and answered by other users. Comparing with information through search engines such as Google [1] [6], which the results are not always correspond to user requirements, in Community Question Answering (CQA) information seekers provides the information needed by other users such as Yahoo! Answer, Naver or Answer Bag.

These communities have become quite popular in the last several years for a number of reasons. First, because of the targeted response from users with knowledge or experience, it is making users more useful and easy to understand the information. Second, the information also provides consolidated communication environment in which the information related to the questions could be seen. This environment facilitates multiple answers (likely from a different perspective) and discussion (in the form of comments) which could benefit the questioner (and others as well).

By clarification and suggestion (using email or other means), it is possible for the questioner to interact with the answerer. This paradigm is, although, quite different from the instantaneous search for stored information, this is likely to provide the questioner with useful answer. Finally, the forum provides an incentive for the users to show their skills and in the process get acknowledged by the community. Such as collaborative learning, users could exploit and share their resources and skills by asking information, evaluating, monitoring one another's information and idea.

Many CQA service providing non-textual information related to their document collections. Usually textual features are used to measure relevance of the document to the query and non textual features can be utilized to estimate the quality of the document. The information from non-textual feature has potential for improving search quality [2] such as points, best answers, contributor etc. In the other hand, the quality of information given by traditional content could be favorable and trusted. For the social media of CQA, the quality of information is diverse, from the high-quality, low-quality or spam. The quality of an answer or of any information in document content for that matter could be subjective.

Jeon et al [2] [3] using non-textual features to predict quality of answers. They collected Q&A pair of data and 13 features from the Naver Q&A service which is written in Korean.

To handle various types of non-textual features and build a stochastic process that could predict the quality of documents, they use kernel density estimation [12] and maximum entropy approach. [13] Introduce the problem of predicting information seekers satisfaction in collaborative question answering communities. [16] also trying to predict selected information by using 13 quality criteria to evaluated the answers (5-point Likert scale) and 9 feature. Occasionally, answerer's temporal characteristic could significantly contribute to the quality of an answer beside activity feature [17]. This paper presents a method for systematically processing non-textual feature to predict the quality of information collected from specific Indonesian web service (id. Y!A) using classifier.

II. PROPOSED METHOD AND SYSTEM

A. System Architecture

The proposed method in this paper consists of four parts. There are data collection, feature extraction, coefficient correlation with an answers, and classification. Figure 1 showing the architecture of the proposed system.

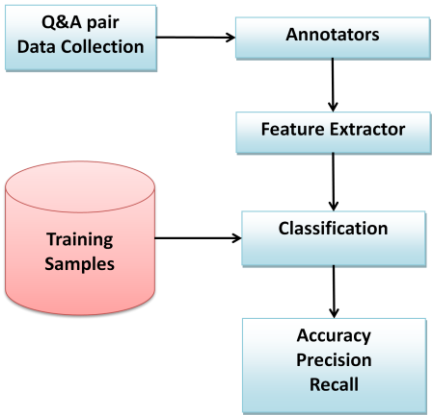

Fig. 1. Architecture of The Proposed System

B. Data Collection

Our data is based on a snapshot of Yahoo! Answer for Indonesian people (http://id.answers.yahoo.com/), a popular CQA site. Our first step is collecting categories that have the highest activity (question resolved) from the 26 category. From table 1, we could see a category that has high activity. There are music and entertainment category, society and culture category, computers and internet category, family and relationship category, and the last consumer and electronic category.

TABLE I. ACTIVITY FOR 5 HIGHSET CATEGORY RESOLVED

Category	Resolved	Resolved question for each category ratio
Music and entertainment	436224	39656
Society and culture	377993	37799
Internet and computer	258870	36981
Family and relationship	123846	20641
Consumer and electronic	102513	9319

(Taken in July - August, 2012)

In order to focus on a realistic question and answer, we choose internet and computer category. The selection is based on the idea that several sub category on music entertainment and society culture providing highly subjective answer such as religion and spirituality.

We collected 258870 Q&A pairs from id.Y!A service (internet and computer), all question and answer are written in Indonesian. We randomly selected resolved question from 7 sub category and all we found 1500 Q&A pairs. The quality of a Q&A depends on the question part and answer part. For the question part we use most popular resolved question. Users could not get any useful information from bad questions. The reality bad questions always lead to bad quality answers. Therefore we decide to estimate only the quality of answers and consider it as the quality of the Q&A. In the Y!A CQA, multiple answers are possible for a single question and the questioners selects the best answer. We extract features only from the best answer. We use statement for evaluating answers [13]. The asker personally has closed the question and selected the best answer; also provide a rating of at least 3 stars for the best answer quality.

The information of CQA is typically complex and subjective. We use annotators for manual judgment of answer quality and relevance. General, good answers tend to be relevant, information, objective, sincere and readable. We may separately measure these individual factors and combine scores to calculate overall the quality of the answer. Therefore, we propose to use a holistic view to decide the quality of an answer. Our annotators read answers, consider all of the above factors and specify the quality of answers in three levels: Bad, Medium and Good (in the future classified as good, medium and bad).

C. Feature Extraction

First we will extract feature vectors from a Q&A pair (answer yahoo). We extract 18 non-textual features, divide as answer feature/AF (feature 1 to 8) and answerer user history/AUH (feature 9 to16). Because in community question answer, multiple answers for single answer are possible. We extract features only form the questioner selects (best answer). The features are;

(1) Star: Number of stars that given by questioners from one to five stars to the answer.

(2) Reference: When answer the question; sometime answerer's give the reference for the answer.

(3) Vote-up: Number of positive votes.

(4) Vote down: Number of negative votes.

(5) Contributor: Answerer's, who are specifically in several categories.

(6) Character length: Number of characters for the answer.

(7) World length: Number of words for the answer.

(8) Sentences length: Number of sentences for the answer.

(9) Member since: How long since last registration from the all activity.

(10) Answerer's activity level: Answerer's activity level.

(11) Answerer's total point: Total point from all the answer.

(12) Total number of answer: Total number of all answerer's that answers answered previously.

(13) Number of best answer: Total number of best answer.

(14) Best answerers acceptance ratio: The ratio between best answers to all the answers that the answers answered previously.

(15) Number of other answer: Total number of other answer (not best answer) that answerer's answered previously.

(16) Answerers other acceptance ratio: Ratio of other answers (not best answer) to all the answerer's answered previously.

(17) Best and other answer ratio: Ratio of best answers to the other answers previously.

(18) Answer question ratio: Ratio of all answer to the entire question previously.

D. Correlation Coefficient

The function of the correlation coefficient is to know how closely one variable is related to another variable [4], in this case the correlation between individual features and the annotators scores (good answers have higher scores: Bad = 0, Medium = 1, Good = 2). Table 2 showing 13 features' that have strongest correlation with the quality of answer. Surprisingly, number of char and number of word have the strongest correlation with the quality of the answer. On the other side, number of star is not the feature that has strongest correlation with the quality of the answer. This means the number of stars that given by questioners evaluation is subjectively, some of users opinion does not agree with the answer. Almost users appreciate getting answers regardless of the quality of the answers. This user behavior may be related to the culture of Indonesian users, same as Korean users [2].

The formula for Pearson's Correlation Coefficient:

$$r_{xy} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n_x}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{n_y}\right]}} \quad (1)$$

TABLE II. COEFFICIENT CORRELATION

Features	Correlation
Star	0.3391
Contributor	0.3323
Member since	0.2147
Activity level	0.4705
Total point	0.4285
Total answer	0.4464
Best answer	0.4435
Ratio best answer	0.3323
Other answer	0.3846
Number of char	0.6391
Number of word	0.6607
Number of sentence	0.5740
Answer question ratio	0.2303

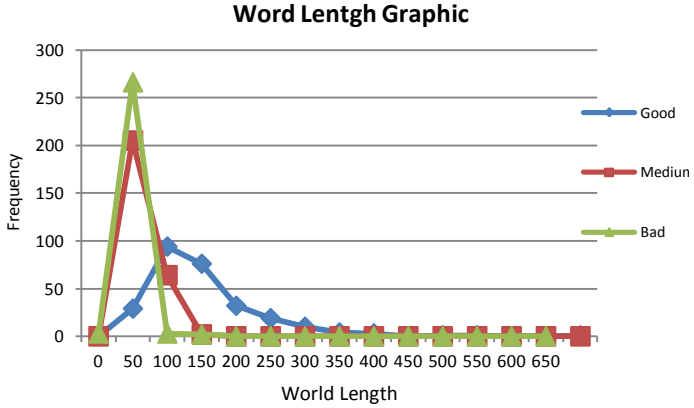

Fig. 2. Distributions of Word Length

Figure 2 show the distributions of good, medium and bad quality answer for word length. Good answers are usually longer than bad and medium answer.

E. Classification Algorithms

We explored Decision Tress, Boosting and Naïve Bayes, using Weka framework [15]. Using a decision tree classifier, we expect o get high precision on the target class. Support vector machines are considered the classifier of choice for many tasks, and to handle the noisy features use AdaBoost. Using Naïve Bayes cause has performed very simple and fast, effective method to investigate the success of our experiment.

III. IMPLEMENTATION AND RESULTS

We l implement the proposed methods to the Q&A pair of data. There are four kind data for the classification, data from the entire feature, data with high correlation (> 0.1 and > -0.1), data from answer feature, and data from answer user history. We build the predictor using 815 training data and 302 testing data (from the annotators we get 1117 related Q&A pair data). Table 3 reports prediction accuracy for different implementations, comparing the choice in classifier algorithm and features for training set, testing set also in 5 cross validation

TABLE III. ACCURACY OF TRAINING FOR EACH FEATURE

Classifier	All	Corr	AF	AUH
Naïve Bayes	73.13	69.73	79.14	50.80
Adaboost	81.10	80.27	81.10	53
C4.5	91.90	91.42	88.83	66.50

Table 3 reports prediction accuracy for the different implementation of answer quality, in particular comparing the choice in classifier algorithm, feature sets (using all feature, Correlation feature, answer feature, answerer history feature) and test option. Surprisingly C4.5 results in the best performance of all the classification variants, with accuracy on the satisfied class of 91.9 for all features. From the same table we could see that by using answer feature (AF) and answerer user history (AUH) the accuracy it is not so good, especially for answerer user history. For the answer feature is closed to within 3.07 with all feature and 2.59 with Correlation feature.

The geometric mean of precision and recall measures (F1) reported in Table 4. We could see from all feature set and Correlation feature set by using test option, C4.5 have higher F1 for 91.9, training set, 89.1 testing set and 81 using 5 cross validation. Another interesting result from Table 4 and 5 we could see that the differences between all features and Correlation feature, is not too significant for accuracy it is about 0,52. This indicates that feature which does not have high correlation is not too pretty significant impact for classification results.

TABLE IV. PRECISSION AND RECALL OF ALL FEATURE, CORR

Classifier		all			Correlation		
		cv = 5	Train	Test	cv = 5	Train	Test
Naïve Bayes	F1	72.3	72.7	65.8	68.9	62.6	68.6
	Accuracy	72.26	73.13	65.56	69.98	69.73	62.91
Ada boost	F1	80	80.6	79.7	79.2	79.8	76.6
	Accuracy	80.12	81.10	79.80	79.66	80.27	76.82
C4.5	F1	**81**	**91.9**	**89.1**	79.9	91.4	89.1
	Accuracy	**80.98**	**91.90**	**89.07**	79.90	91.42	87.09

FEATURE

TABLE V. ACCURACY OF ALL FEATURE AND CORR FEATURE

Classifier		all			Correlation		
		cv = 5	Train	Test	cv = 5	Train	Test
Naïve Bayes	Precission	72.4	72.8	66.7	69.8	63.9	69.7
	Recall	72.8	73.1	65.6	70	62.9	69.7
Ada boost	Precission	80.2	81	80	79.9	80.7	78
	Recall	80.1	81.1	79.8	79.7	80.3	76.8
C4.5	Precission	**81.1**	**91.9**	**89.2**	80	91.4	89.3
	Recall	**81**	**91.9**	**89.1**	79.9	91.4	89.1

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         749               91.9018 %
Incorrectly Classified Instances        66                8.0982 %
Kappa statistic                          0.8785
Mean absolute error                      0.092
Root mean squared error                  0.2145
Relative absolute error                 20.7059 %
Root relative squared error             45.5037 %
Total Number of Instances              815
```

Fig. 3. Result of Classification on Training Data

```
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         269               89.0728 %
Incorrectly Classified Instances        33               10.9272 %
Kappa statistic                          0.8361
Mean absolute error                      0.111
Root mean squared error                  0.2515
Relative absolute error                 24.9781 %
Root relative squared error             53.3488 %
Total Number of Instances              302
```

Fig. 4. Result of Classification on Testing Data

IV. IMPLEMENTATION AND RESULTS

In this paper we presented our knowledge to quantify and predict quality of answer in question answering communities, especially for Indonesian CQA. Beyond developing models to select best answer and evaluate the quality of answers, there are several important lessons to learn here for measuring content quality in CQA. We find huge variety of question and answer on CQA services, and by given question may several answers are providing from the community.

With appropriate features, we could build models that could have significantly higher probability of identifying the best answer class than classifying a non-best answer.

From the entire system by using Q&A pairs from id.answer yahoo, 18 feature and 3 type classification. We conclude as following:

(19) From the four existing feature, the highest accuracy exist on all feature set (comparing with correlation coefficient set, AF set and AUH set)

(20) The best performance of all classification variants by using C4.5, with average accuracy 91.90 , precision 91.9 and recall 91.9

In the future our models and predictions could be useful for predictor quality information as a recommender system to complete a collaborative learning.

ACKNOWLEDGMENT

The authors would like to thank the students in the State University of Malang who participated to our experiments.

REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1{7):107{117, 1998.
[2] J. Jeon, W. B. Croft, and J. H. Lee. A framework to Predict the Quality of Answer with Non-Textual features. In Proceedings of SIGIR, Seattle, Washington,USA. 2065.
[3] J. Jeon, W. B. Croft, and J. H. Lee, Soyoen Park. Finding similar questions in large question and answer archives. In Proceedings of the ACM Fourteenth Conference on Information and Knowledge Management, pages 76{83, 2005.
[4] Higgins Jim, The correlation Coefficient, 2005
[5] H. Kim and J. Seo. High-performance faq retrieval using an automatic clustering method of query logs. Information Processing and Management, 42(3):650{661, 2006.
[6] [6] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604{632, 1999.
[7] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 275{281, 1998.
[8] D.M. Strong, Y.W. Lee and R.Y. Wang. Data Quality in Context. Communication of the ACM, 40(5):103-110,1997.
[9] C.-H. Wu, J.-F. Yeh, and M.-J. Chen. Domain-specific faq retrieval using independent aspects. ACM Transactions on Asian Language Information Processing, 4(1):1{17, 2005.
[10] Y. Zhou and W. B. Croft. Document quality models for web ad hoc retrieval. In Proceedings of the ACM Fourteenth Conference on Information and Knowledge Management, pages 331{332, 2005.
[11] X. Zhu and S. Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 288{295, 2000.
[12] Zucchini Walter. Applied Smoothing Techniques Kernel Density Estimation, 2003.
[13] Liu. Y, Bian. J, Agichtein. E. Predicting Information Seeker Satisfaction in Community Question Answering. SIGIR;08 July 20-24,2008, Singapore.
[14] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 275{281, 1998.
[15] I. Witten and E. Frank. Data Mining : Practical machine learning tools and techniques. Morgan Kaufman, 2nd edition, 2005.
[16] Chirag Shah, Jefferey Pomerantz. Evaluating and Predicting Answer Quality in Community QA, SIGIR'10, July 19-23, Geneva, 2010 Switzerland.

[17] Yuanzhe Cai, Sharma Chakravarthy. Predicting Answer Quality in Q/A Social Networks: Using Temporal Features. Technical Report, Department of Computer Science and Engineering University of Texas at Arlington, 2011.

AUTHORS PROFILE

Kohei Aarai He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 and also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission-A of ICSU/COSPAR since 2008. He wrote 30 books and published 332 journal papers.