ISASE 2019

Further Tendency of Obtaining Potential Appropriate Respondents to Questions at Q&A Sites through Additional Categories

Yuya YOKOYAMA*, Teruhisa HOCHIN ** and Hiroki NOMIYA **

** Kyoto Prefectural University, 1-5 Hangi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan*
*** Kyoto Institute of Technology, Gosyokaidou-cho, Matsugasaki, Sakyo-ku, Kyoto, 606-8585, Japan*

Abstract: In order to eliminate mismatches between the intentions of questioners and respondents of Question and Answer (Q&A) sites, nine factors of impressions for statements have experimentally been obtained. Factor scores are then estimated by using the feature values of statements. So far the possibility of searching respondents capable of giving appropriate answers to a newly posted question has been established for Auction, PC and Love. It has been shown that the distance and the number of appearance may help us select users who can give appropriate answers to a question. In the similar fashion, this paper tries to find the possibility of detecting respondents who can appropriately answer a newly posted question for other categories such as Internet, Politics, etc. As a result of analysis, several newly regarded categories shows the similar tendency as the previous analysis, while some categories related to Yahoo! JAPAN show less outstanding tendency.
Keywords: *Q&A site, Respondents, Category*

1. INTRODUCTION

Recently, the number of people using Question and Answer (Q&A) sites on the Internet has been increasing. Q&A sites are online communities where users can manually post questions and answers. Thus, these sites are thought to be databases containing enormous amounts of knowledge to solve various problems. When a user posts a question, others may respond. The questioner selects the most appropriate response as the "Best Answer" (BA) and awards the respondent with a certain number of points that serve as a fee. The BA is the response statement the questioner subjectively finds most satisfying.

As the number of users of Q&A sites increases, and more questions are posted, it becomes harder for respondents to select questions that match their specialty and interests. Consequently, a question posed by a user may not be seen or answered by qualified respondents. Moreover, if an appropriate respondent is not encountered, mismatching may occur.

There have been a number of prior works investigating Q&A sites, such as introducing users to answer statements [1, 2] and inspecting the quality or tendency of answer statements [3-5], etc. These prior works have mainly used textual features or link analysis. However, some users may prefer a polite style, while others may write statements in a ruder style. These tendencies have not been considered. Meanwhile, our work focuses on using impressions as well as textual features. Moreover, although there are some prior works that introduce users to answer statements as described [1, 2], it is difficult to say that a way to introduce appropriate respondents to a questioner has yet been established. Thus, with using the impression of statements, our work aims to introduce appropriate respondents to a questioner.

Our goal is to present questions to users who are qualified to properly answer them, thus avoiding the problems described above. The impressions of sixty statements posted on Yahoo! Chiebukuro (Y!C) [6] have been evaluated [7]. By applying factor analysis to the experimental results, nine factors were obtained.

Factor scores obtained through factor analysis represent the impressions of the statements, and this is necessary for estimating the factor scores of other statements. The statements were estimated using multiple regression on the feature values assigned to the statements [8]. The overall estimation accuracy for all nine factors was proved to be good.

The possibility of detecting respondents who can appropriately answer a newly posted question statement was examined [9]. Here we assumed that a respondent whose impression is similar to that of a questioner is thought to be adequate to give an adequate answer. We attempted a process of getting proper respondents to a question by using questions actually already posted at Y!C. As a result of analysis, it was shown that there were some nonresponding users who have a shorter Euclidean distance between the factor score of a questioner and that of a respondent. Those users were thought to be capable of appropriately answering that question. It was also shown that with the consideration of categories, more appropriate users could be detected [9].

So far the categories of Q &A used for the previous analysis are only Auction, PC and Love. This paper aims to envision the tendency of obtaining appropriate respondents for other categories. Here, question statements of other categories are appended for further analysis. Among the dataset used for the previous analysis, categories whose answer statements account for over 800 are newly regarded. As a result of analysis, several newly regarded categories shows the similar tendency as previous analysis. Meanwhile, some categories related to Yahoo! JAPAN show less outstanding tendency.

The remainder of the paper is organized as follows. Our previous works are summarized in Section 2. A method to obtain appropriate potential answerers with additional categories is proposed and evaluated in Section 3. Finally, Section 4 concludes the paper.

2. PREVIOUS WORKS

2.1 Obtaining Factors of Statements

An experiment was conducted to evaluate impressions of answers. There were forty-one evaluators, and they evaluated the style or content of statements and assigned labels from a group of fifty words [7]. Twelve sets of questions and answers were evaluated, and these included three from each of four major categories: Auction, PC, Love, and Political/Social Problems; the categories were chosen from those actually posted at Y!C in 2005 [7]. Factor analysis was applied to the experimental results, and nine factors were obtained. The factors indicate the nature of a statements, as explained by the various impression words assigned to that statement; they were named accuracy, displeasure, creativity, ease, persistence, ambiguity, moving, effort, and hotness.

2.2 Estimation of Factor Scores

The factor scores were obtained for only the sixty statements used in the experiment. To be able to estimate the factor scores of other statements, multiple regression analysis was applied to their feature values [8]. Overall, seventy-seven feature values were adopted; these are summarized, syntactic information, word imageability, expression in the closing sentence, word familiarity and notation validity.

Multiple regression analysis was performed on the sixty questions and answers employed in the impression evaluation experiment. 281 quadratic terms (the product of two explanatory variables) were used according to seventy-seven explanatory variables, and the respondent variables with factor scores for the nine factors. Multiple correlation coefficients, which show the goodness of the estimation, were above 0.9 for all nine factors [8]. As a result of analysis, it was shown that the estimation accuracies of all of the factors are very good.

2.3 Impression and Suitability of Q&A

2.3.1 Purpose

The differences are determined between the impressions of a given question and the answers already posted for it in Y!C, by calculating the Euclidean distance between their factor scores [9]. These differences could be used to identify those users who could appropriately answer that question.

2.3.2 Dataset

The differences between the impressions of the questions and those of the answers, and the suitability of the answers were examined in order to determine if it would be possible to use this information to find users who would be expected to give an appropriate answer [9]. One question was chosen from each of the following categories: Auction, PC, and Love. For the respondents who posted the answers to those questions, the factor scores for most of the answers they posted in 2005 are obtained, by using the multiple regression equations.

2.3.3 Result

Distances were calculated for a total of 66,238 answer statements, using Formula (1):

$$D = \sqrt{\sum_{k=1}^{9} \left(Fac_{Q_k} - Fac_{A_k} \right)^2} \qquad (1)$$

where Fac_{Q_k}, Fac_{A_k} are the score for the kth factor for a question and for an answer, respectively. The distances were then sorted in ascending order. The answer statements whose distances are ranked in shortest fifteen are shown in Table 1. Each row indicates a separate answer statement, while each column entitled is explained as follows:

- "No." means the ascending order rank of distance for each answer statement.
- "Distance" means the Euclidean distance between the factor score of the answer statement and that of a question one, obtained through Formula (1).
- "User" indicates users who appear at least twice. Those users are denoted as (A), (B) and (C). They are shaded in Table 1 to emphasize that they appear at least twice.

The users denoted as (A), (B) and (C) in the column entitled "User" appear repeatedly for Auction, PC, and Love, respectively. Therefore, users who posted answer

Table 1 Shortest 15 Answer Statements for Each Category (Auction, PC, Love)

(a) Auction

No.	Distance	User
1	0.770	
2	0.796	(A)
3	0.805	(A)
4	0.831	
5	0.834	(A)
6	0.862	(A)
7	0.899	
8	0.910	
9	0.913	(A)
10	0.914	(A)
11	0.933	(A)
12	0.935	(A)
13	0.939	
14	0.943	
15	0.944	(A)

(b) PC

No.	Distance	User
1	0.999	(A)
2	1.026	
3	1.080	(B)
4	1.141	(B)
5	1.143	
6	1.161	(B)
7	1.176	
8	1.213	
9	1.229	(A)
10	1.239	(C)
11	1.241	
12	1.266	
13	1.282	
14	1.283	(B)
15	1.285	(B)

(c) Love

No.	Distance	User
1	0.850	(C)
2	0.864	
3	0.875	(C)
4	0.925	(C)
5	0.930	(C)
6	0.934	(C)
7	0.939	(C)
8	0.941	(C)
9	0.944	(C)
10	0.945	
11	0.949	(C)
12	0.953	(C)
13	0.961	
14	0.962	
15	0.966	(C)

statements of the same categories appear in Table 1. From these results, with the consideration of categories of answer statements, particularly some users appear most in each dataset [9]. Thus, users who have posted answers on a certain category can be thought to appropriately answer a question of the identical category.

3. ANALYSIS WITH ADDITIONAL CATEGORIES

3.1 Additional Categories

So far the categories of Q &A used for the previous analysis explained in Section 2.3 are only Auction, PC and Love. Therefore, additional analysis is required with more categories other than Auction, PC and Love. In this paper, we aim to envision the tendency of obtaining appropriate respondents for further categories.

Question statements of other categories are added for further analysis. Here, among the dataset used for the previous analysis explained in Section 2.3, the categories whose answer statements account for over 800 are shown in Table 2. The categories other than Auction, PC and Love are newly regarded; Yahoo! Chiebukuro (Y!C), Internet, Politics & Social Issue (Pol), Yahoo! Service (Y!S), Nippon Professional Baseball (NPB), and Tax. These are shaded in Table 2.

Table 2 Categories of Answer Statements (only 800+)

Category	Number of Answer Statements
Auction	23528
PC	4071
Love	2824
Yahoo! Chiebukuro (Y!C)	1980
Internet	1803
Politics & Social Issue (Politics)	946
Yahoo! Service (Y!S)	840
Nippon Professional Baseball (NPB)	824
Tax	817

Table 3 Shortest 15 Answer Statements for Each Category (Y!C, Internet, Politics, Y!S, NPB, Tax)

(a) Y!C

No.	Distance	User
1	1.93	
2	1.97	
3	2.15	(D)
4	2.22	
5	2.28	(E)
6	2.32	
7	2.38	(E)
8	2.55	(F)
9	2.55	(F)
10	2.59	
11	2.60	
12	2.62	
13	2.65	
14	2.66	
15	2.71	(D)

(b) Internet

No.	Distance	User
1	0.187	
2	0.247	(G)
3	0.251	
4	0.267	
5	0.278	
6	0.321	(H)
7	0.341	
8	0.356	
9	0.358	
10	0.362	(H)
11	0.401	
12	0.411	(G)
13	0.413	(D)
14	0.421	(H)
15	0.431	(B)

(c) Politics

No.	Distance	User
1	0.381	(B)
2	0.404	(D)
3	0.425	(C)
4	0.431	(D)
5	0.452	(D)
6	0.482	(A)
7	0.482	(C)
8	0.549	(D)
9	0.576	
10	0.588	(D)
11	0.593	(D)
12	0.596	
13	0.603	(D)
14	0.606	(D)
15	0.613	(D)

(d) Y!S

No.	Distance	User
1	0.868	
2	0.908	
3	0.994	
4	1.08	
5	1.12	
6	1.26	
7	1.28	
8	1.33	
9	1.34	
10	1.36	
11	1.38	
12	1.38	
13	1.39	
14	1.39	
15	1.40	

(e) NPB

No.	Distance	User
1	0.722	(D)
2	1.12	(D)
3	1.22	
4	1.23	
5	1.24	(D)
6	1.25	(D)
7	1.26	(E)
8	1.29	(C)
9	1.29	(E)
10	1.30	(D)
11	1.32	(E)
12	1.33	(D)
13	1.33	(D)
14	1.34	(A)
15	1.35	

(f) Tax

No.	Distance	User
1	1.25	(D)
2	1.49	(D)
3	1.51	(D)
4	1.56	(D)
5	1.60	(D)
6	1.67	(D)
7	1.69	(D)
8	1.69	(D)
9	1.69	(D)
10	1.72	(D)
11	1.78	(D)
12	1.80	(D)
13	1.80	(D)
14	1.80	
15	1.83	(D)

3.2 Dataset

Six question statements and 66,238 answer statements are used as experimental materials. The question statements consist of one question for each six categories depicted in Section 3.1. The answer statements are the same ones used for the previous analysis as well. The distances are then sorted in ascending order.

3.3 Result

Similar as the previous analysis explained in Section 2.3, Euclidean distances between question and answer statements are calculated by using Formula (1). Here, the answer statements whose distances are ranked in shorter fifteen are shown in Table 3. Each column is explained in Section 2.3. From Table 3, the users denoted as (D) to (H) in the column entitled “User” appear repeatedly. Here, the users denoted as (A) to (C) indicates the same respondents that appeared in Table 1.

From the results shown in Table 3, for the categories of Internet, Politics, NPB and Tax, particularly several users appear most in each dataset. Therefore, users who have posted answers on a certain category can be thought

to appropriately answer a question of the identical category. This tendency is similarly observed for Auction, PC and Love shown in Table 1. The user (D) appears in Politics, NPB, and Tax. Above all, 14 out of 15 answer statements are posted by the user (D).

On the other hand, less outstanding tendency is observed for Y!C and Y!S, which are related to Yahoo! JAPAN service. There are not more than two answers posted by an identical user for Y!C, while all the fifteen answer statements are posted by unique fifteen users for Y!S. This tendency could be attributed to the fact that these Q&A statements were posted in 2005, which is the second year since Y!C service started in 2004. Therefore, it could be suggested that there were not many users who had enough knowledge about the systems of Y!C or Y!S at that time.

4. CONCLUSION

In this paper the tendency of obtaining appropriate respondents was investigated for additional categories other than Auction, PC and Love. Among the dataset used for the previous analysis, categories whose answer statements account for over 800 were newly regarded; Y!C, Internet, Politics, Y!S, NPB and Tax. Six question statements (one each for those categories) and 66,238 answer statements were used as dataset. As a result of analysis, for Internet, Politics, NPB and Tax showed the similar tendency as previous analysis; Auction, PC and Love. Meanwhile, Y!C and Y!S, which are related to Yahoo! JAPAN service, showed less outstanding tendency as the other four categories.

For future work, additional Q&A statements and categories will be required to investigate the further tendency of obtaining appropriate respondents. In addition, the contents of answer statements must be considered for the improvement of precision and the clarification of tendency among categories. It is also required to determine a way to estimate the objective scores of the answers. After that, those scores will be used to estimate the BAs. In order to find appropriate answerers, the characteristics of users (both questioners and answerers) must be used for investigation. As most of the feature values in this study are dependent on Japanese, generalization to other languages is included in our future work as well.

ACKNOWLEDGMENTS

This research was partially supported by the Japan Society for the Promotion of Science, Grant Number 26008587, and 2015–2016. This research used the data of "Yahoo! Chiebukuro" that was given to the National Institute of Informatics by Yahoo Japan Corporation.

REFFERENCES

[1] Pawal Jurczyk and Eugene Agichtein: Discovering Authorities in Question Answer Communities by Using Link Analysis, Proc. of 16th ACM Conference on Information and Knowledge Management (CIKM), pp. 919-922, 2007.

[2] Fatemeh Riahi , Zainab Zolaktaf , Mahdi Shafiei and Evangelos Milios: Finding Expert users in Community Question Answering, Proc. of the 21st International Conference Companion on World Wide Web (WWW12), pp.791-798, 2012.

[3] Eugene Agichtein, Castillo Carlos, Debora Donato, Aristides Gionis and Mishne Gilad: Finding High-Quality Content in Social Media, Proc. of the Int'l Conf. on Web Search and Web Data Mining (WSDM08), pp.183-194, 2008.

[4] Yuanjie Liu , Shasha Li, Yunbo Cao, Chin-yew Lin, Dingyi Han, and Yong Yu: Understanding and Summarizing Answers in Community-Based Question Answering Services, Proc. of the 22nd International Conference on Computational Linguistics, pp.497-504, 2008.

[5] Daphne Ruth Raban: Self-Presentation and the Value of Information in Q&A Websites, Journal of the American Society for Information Science and Technology Volume 60, Issue 12, pp.2465-2473, 2009.

[6] Yahoo! Chiebukuro (URL, in Japanese), http://chiebukuro.yahoo.co.jp/, 2018.

[7] Yuya Yokoyama, Teruhisa Hochin, Hiroki Nomiya, and Tetsuji Satoh: Obtaining Factors Describing Impression of Questions and Answers and Estimation of their Scores from Feature Values of Statements, Studies in Computational Intelligence, Volume 413, pp.1-13, Springer, 2012.

[8] Yuya Yokoyama, Teruhisa Hochin, and Hiroki Nomiya: Using Feature Values of Statements to Improve the Estimation Accuracy of Factor Scores of Impressions of Question and Answer Statements, International Journal of Affective Engineering, Vol. 13, No. 1, pp.19-26, 2014.

[9] Yuya Yokoyama, Teruhisa Hochin, and Hiroki Nomiya: Improvement of Obtaining Potential Appropriate Respondents to Questions at Q&A Sites by Considering Categories of Answer Statements, International Journal of Affective Engineering, Vol.16, No.2, pp.63-73, 2017.