Finding Appropriate Respondents to Questions at Question and Answer Sites

Yuya YOKOYAMA † Teruhisa HOCHIN ‡ and Hiroki NOMIYA ‡

† ‡ Graduate School of Science and Technology, Kyoto Institute of Technology, Gosyokaido-cho, Matsugasaki, Sakyo-ku, Kyoto, 606-8585 Japan

E-mail: † yuya_dan_yokoyama.0719@hotmail.co.jp, ‡ {hochin, nomiya}@kit.ac.jp

Abstract This paper proposes a method of selecting respondents who can give an appropriate answer to a question in order to eliminate mismatches between the questioners and respondents at Question and Answer sites. The proposed method uses the number of appearance of respondents and the score based on the distance between the factor scores of a question and an answer already posted. Nine factors of impressions for statements have experimentally been obtained. Factor scores have been estimated through multiple regression by using feature values of the statements. The possibility of detecting respondents capable of appropriately answering to a newly posted question has been examined. The proposed method is based on this observation. It is experimentally evaluated by comparing it with the methods based on average scores and distances through precision and recall. It is shown that the proposed method outperforms the methods compared with it. It is also shown that the proposed method could successfully select the respondents that are more than almost averagely appropriate to a question.

Keyword Q&A Sites, Factor Score, Impression

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. Hence, these sites can be considered as databases containing enormous amounts of knowledge that can be used to solve various problems. When a user posts a question, other users may respond. The questioner selects the most appropriate response as the “Best Answer” and awards the respondent with some points as a form of fee. The Best Answer is the response statement that the questioner subjectively finds most satisfying.

As the number of users of Q&A sites increase 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, which may cause the following problems:

- A questioner may acquire incorrect knowledge from inappropriate answers.
- Respondents may not have the necessary knowledge to properly answer the question, and thus the problem remains unsolved.
- Users may be offended by answers that contain abusive words, slanders, or statements against public order and standards of decency.

Our objective is to present questions to qualified users who can appropriately answer them, thus avoiding the problems described above. The impressions of sixty statements posted on Yahoo! Chiebukuro [1] have been evaluated [2]. By applying factor analysis to the scores obtained in the experiment, nine factors were obtained.

Factor scores obtained through factor analysis represent the impressions of statements. However, mere factor scores for the statements used in the experiment can be obtained. It is required to estimate the factor scores of other statements. They are estimated through multiple regression by using feature values of the statements. Feature values include the syntactic information of the statements, such as word classes (such as nouns and verbs), and the number of appearances (or the percentage) of alphanumeric characters and kanji [3]. Moreover, word imageability, closing sentence expressions, word familiarity, and notation validity are also adopted as feature values. It has been shown that the overall estimation accuracy for all the nine factors is good. The validity of estimating the scores of each factor by obtaining the major feature values is confirmed.

The possibility of detecting respondents who can give an appropriate answer to a newly posted question has been examined [4]. It has been shown that there is possibility that users other than actually posted respondents could be an appropriate answerer. It has also been shown that there are some users who appear several times in higher rank in ascending order of Euclidean distance between the factor score of a questioner and that of a respondent. Those users

are thought to be capable of answering a question.

On the basis of the observation described above, the method of selecting respondents who can give an appropriate answer to a newly posted question is proposed. The proposed method determines and ranks the possibly qualified respondents according to the appearance and score on the basis of distance. It is experimentally evaluated by comparing the methods based on average scores and distances through precision and recall. It is shown that the proposed method outperforms the methods compared with it. It is also shown that the proposed method could successfully select the respondents that are more than almost averagely appropriate to a question.

The remainder of this paper is organized as follows. In Section 2, related works are described. Our previous works are summarized in Section 3. Experimental evaluation is shown in Section 4. Considerations towards the evaluation are provided in Section 5. Finally, Section 6 concludes the paper.

2. Related Works

There have been a number of prior works that investigate Q&A sites. Blooma *et al.* used both textual and non-textual features to predict the Best Answers [5]. They used five textual and five non-textual features. It was found that textual features influence the quality of answers more than non-textual features. Accuracy, completeness, language, reasonableness, and length are considered as textual features. The analogical reasoning approach [6] finds the Best Answer by using links between questions and answers contained in previous knowledge. In their approach, three textual features, seven statistical features, and five user interactions were used. Nishihara *et al.* have also proposed a method of detecting the Best Answer to a question [7]. They obtain the Best Answer to a question by noticing the affinity between closing sentence expressions of questioners and respondents, and the clustering combinations of questions and Best Answers. Adamic *et al.* analyze the characteristics of knowledge of Yahoo! Answers and cluster the categories into three groups by thread length [8]. The prediction of Best Answer is attempted by using network analysis, entropy and cross-validation. The results show that reply length, the number of competing answers, and the track record of the user were the most predictive of whether the answer would be chosen.

Several researches are attempted in terms of introducing users to answers. Riahi *et al.* investigate the way to provide appropriate experts with a newly posted question [9]. Profiles are constructed on the basis of their answering history and then used through several measures. For some of the dataset, their proposal model shows better performance than other methods in recommending new questions to experts. Jurczyk *et al.* detect an authority uses for specific question categories through link analysis [10]. Link structures of communities are analyzed in order to improve the quality of answers.

These prior works, however, have focused mainly on textual features or link analysis. Some users may prefer polite style, while others may write statements in rude style. Some users often use abstract words, whereas others express specific words. These tendencies have not been considered. Meanwhile, our work focuses on using impression as well as textual feature values. To the best of our knowledge, the way of introducing appropriate respondents to questioners has not been established yet.

3. Previous Works

3.1. Factors of Statements

Impression evaluation experiment was conducted for 41 subjects, with using 50 impression words that represent the impression or evaluation from the style or content of statements [2]. Experiment materials are twelve sets of question and answer statements (three each from four major categories: Auction, PC, Love, and political & social problems), out of the statements actually posted at Yahoo! Chiebukuro in 2005 [1]. Factor analysis was applied to the experimental result and nine factors were obtained. A factor means the nature of statements explained by several impression words. Factors are named *accuracy, displeasure, creativity, ease, persistence, ambiguity, moving, effort,* and *hotness*, respectively.

3.2. Estimation of Factor Scores

3.2.1. Feature Values of Statements

The factor scores obtained, however, are only on 60 statements used in the experiment. To be able to estimate the factor scores of any statements, multiple regression analysis was applied to the feature values of statements [3]. Overall 77 feature values adopted are briefly depicted:

- Syntactic information: the number and length of statements, number or percentage of word classes e.g. nouns and verbs, and concrete feature values e.g. exclamation and question marks, etc.
- Word imageability: a subjective characteristic that

implies how we can remind of various imaginations aroused by words.

- Closing sentence expressions: fundamental Japanese words e.g. “zo,” “da,” “yo,” “ne,” “ka,” “na,” “shi,” “desu,” “masu,” “tai,” and “nai.”
- Word Familiarity: an index representing the familiarity a subject feels with a word.
- Notation validity: an index representing the validity of a word.

3.2.2. Estimation Result

Multiple regression analysis was performed on the sixty questions and answers employed in the impression evaluation experiment, using the 281 quadratic terms (the product of two explanatory variables) based on 77 explanatory variables, and the respondent variables with factor scores for the nine factors explained in Section 3.1.

Multiple correlation coefficients, which show goodness of estimation, were above 0.9 for all nine factors [3]. Therefore, it can be said that the estimation accuracy of all the factors is very good.

3.3. Impression and Suitability of Q&A

3.3.1. Purpose

Given a question statement, the differences between the impressions of the question and answers already posted in Yahoo! Chiebukuro can be obtained by calculating the Euclidean distance between the factor scores of the question and the answers. If a premise is possible that the impression of answers similar to that of question could lead to searching for an appropriate respondents who can answer the question, these differences might be used [4].

The differences of the impressions of questions and answers and the suitability of answers to questions are examined in order to inspect the possibility of seeking for qualified users expected to reply an appropriate answer. The categories of the questions statements inspected [11-13] are Auction, PC and Love, one each.

Euclidean distances between each question statement and 66,238 answer statements are calculated. The distance *D* is calculated by using the formula (1):

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

where Fac_{Q_k} and Fac_{A_k} are the *k*th factor score of a question and that of the answer, respectively. The distances are sorted by ascending order.

3.3.2. Consideration

As a result of inspection, there are some users who appear across categories. This means that they posted several answers whose impressions are similar to that of the question. The possibility that such users give appropriate answers to the questioner is considered to be high. The number of appearance can help us select appropriate answerers.

Distance is the difference between the factor scores of the question statement and that of an answer one. It is considered that the answerers, who posted the answer statements close to the question one, may pose the answer to the question. The users who posted these answers are considered to be appropriate answerers. It is considered that distance may also be helpful for selecting appropriate answerers.

It is also shown that there is possibility that users other than actually posted respondents could be an appropriate answerer.

4. Finding Appropriate Respondents to Question

4.1. Method

Based on the observation described in Section 3.3, the method of selecting respondents who can appropriately answer a newly posted question is proposed. The proposed method determines the possibly qualified respondents according to the following two criteria:

1) The number of times a respondent appears in the top *N* rank of ascending order of distance between a set of factor scores of a question and that of an answer.

2) The sum of $Score_k$ if several respondents appear the same times. $Score_k$ is calculated through the formula (2):

$$Score_k = \frac{Dis\_A_1}{Dis\_A_k} \quad (1 \le k \le N) \qquad (2)$$

Here, Dis_A_k indicates the distance of the *kth* answer statement in ascending order of distance. Therefore, when the answers of a user are $A_{k_1}, A_{k_2}, \cdots,$ and A_{k_n}, the user has the score *S* calculated through the formula (3):

$$S = \sum_{i=1}^{n} Score_{k_i} \qquad (3)$$

4.2. The Number of Answers Used

For the criterion 1), the number of answer statements

must be determined. It is determined by inspecting the number of unique users among the top *N* answer statements. The threshold *N* is set as 100, 80, 70, 60, 50, 40, 30, 25, 20, 15, and 10.

The number of unique users is shown in Table 1. As the threshold *N*, the number of top answer statements, is not equal to the number of unique users, it is considered that the answers written by the same user appear more than once in the top *N* answer statements. The number of the answer statements written by top fifteen unique users is shown in Table 2. It is seen that answers written by a few users appear many times. For example, in the case of *N=100* of Auction, the number of the answers written by top two users is 51, which is more than half of all the answers. These users are considered as appropriate ones to the question. From these results, top 100 answer statements are decided to be used because it is considered that appropriate users are included in the top 100 answer statements.

4.3. Evaluation

Impression evaluation experiment is carried out in order to evaluate the proposed method.

4.3.1. Dataset

Experiment materials are three sets of question and answer statements (one each from Auction, PC, and Love), out of those actually posted at Yahoo! Chiebukuro in 2004 and 2005. Each set consists of one question and one hundred answers. The questions are the same ones [11-13] used and described in Section 3.3. The one hundred answers used are described in Section 4.2.

4.3.2. Obtaining Objective Fitness Rates

In order to evaluate the proposed method, fitness rates of the answer statements to a question in the dataset need to be obtained. Fitness rates are the degrees of answer statements objectively assessed by humans. They are experimentally obtained in order to determine a set of relevant respondents. The experiment is conducted for twelve subjects (9 males and 3 females, age of 22-29). Subjects are asked to read a question statement first and then read answer statements to evaluate. The criteria of evaluating answer statements are determined by the following five levels:

- 5: The respondent who wrote the answer statement is highly expected to appropriately answer the question.

Table 1: The Number of Unique Users.

Threshold *N*	Auction	PC	Love
10	6	9	7
15	9	14	9
20	13	16	13
25	18	18	14
30	19	21	16
40	23	29	19
50	28	33	23
60	29	34	26
70	31	41	31
80	37	45	35
100	43	53	42

Table 2: The Number of Answers Written by Fifteen Unique Users for Each Statement.

(a)Auction

rank \ N	100	80	70	60	50	40	30	25	20	15	10
1	28	22	20	17	14	11	8	5	5	5	4
2	23	18	17	13	9	7	5	4	4	3	2
3	5	4	3	3	2	2	1	1	1	1	1
4	3	2	2	2	1	1	1	1	1	1	1
5	2	2	2	1	1	1	1	1	1	1	1
6	2	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1	–
8	1	1	1	1	1	1	1	1	1	1	–
9	1	1	1	1	1	1	1	1	1	1	–
10	1	1	1	1	1	1	1	1	1	–	–
11	1	1	1	1	1	1	1	1	1	–	–
12	1	1	1	1	1	1	1	1	1	–	–
13	1	1	1	1	1	1	1	1	1	–	–
14	1	1	1	1	1	1	1	1	–	–	–
15	1	1	1	1	1	1	1	1	–	–	–

(b)PC

rank \ N	100	80	70	60	50	40	30	25	20	15	10
1	27	21	19	19	11	8	7	5	3	2	2
2	8	8	8	7	7	4	3	3	2	1	1
3	7	3	3	2	2	2	2	2	2	1	1
4	4	3	2	2	1	1	1	1	1	1	1
5	2	2	2	1	1	1	1	1	1	1	1
6	2	2	1	1	1	1	1	1	1	1	1
7	2	2	1	1	1	1	1	1	1	1	1
8	2	2	1	1	1	1	1	1	1	1	1
9	2	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1	–
11	1	1	1	1	1	1	1	1	1	1	–
12	1	1	1	1	1	1	1	1	1	1	–
13	1	1	1	1	1	1	1	1	1	1	–
14	1	1	1	1	1	1	1	1	1	1	–
15	1	1	1	1	1	1	1	1	1	–	–

(c)Love

rank \ N	100	80	70	60	50	40	30	25	20	15	10
1	32	24	23	20	16	13	10	8	6	5	4
2	14	10	9	9	8	6	4	4	2	2	1
3	10	9	5	4	3	3	2	2	2	2	1
4	3	3	3	3	2	2	2	1	1	1	1
5	3	3	3	2	2	2	1	1	1	1	1
6	2	2	2	2	2	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1	–
9	1	1	1	1	1	1	1	1	1	1	–
10	1	1	1	1	1	1	1	1	1	–	–
11	1	1	1	1	1	1	1	1	1	–	–
12	1	1	1	1	1	1	1	1	1	–	–
13	1	1	1	1	1	1	1	1	1	–	–
14	1	1	1	1	1	1	1	1	–	–	–
15	1	1	1	1	1	1	1	–	–	–	–

- 4: The respondent is expected to give an appropriate answer statement to the question.
- 3: It is undecided if the respondent is expected to give an appropriate answer statement to the question.
- 2: The respondent is less expected to give an appropriate answer statement to the question.
- 1: The respondent is not expected to give an appropriate answer statement to the question at all.

The evaluation order is Auction, PC, and Love. Several statements are actually given by the identical respondents, which is not minutely informed to the subjects.

As the averages of the fitness rates of a user are considered as objective fitness rates of the user, the averages of the fitness rates are called objective fitness rates of the user.

To examine if gender difference affected impression evaluation, *t*-test was applied with the significance level at 1% between answers of nine males and those of three females. As a result, *p*-value is 4.68×10^{-6}, which is much smaller than 0.01. It is shown that there is a significant difference between the scores of males and those of females. Thus, the answer statements of males and those of females are separately treated.

4.3.3. Evaluation Procedure

The proposed method is compared with the methods using averaged scores and the distances. The method using averaged scores ranks users based on the average scores of the answers of each user. That using distances ranks users based on the answer having the smallest distance from a question in the answers of each user.

The soundness of these methods is evaluated through precision and recall. Precision is calculated by the ratio of the number of retrieved relevant respondents to that of all retrieved ones. A relevant respondent is defined as "the respondent whose answer statements include the best objective fitness rate over the threshold *n*." Recall is calculated by the ratio of the number of retrieved relevant respondents to that of all relevant ones. Precision and recall are calculated by setting the threshold *n* as 4.0, 3.5, 3.0, 2.5, and 2.0.

4.3.4. Evaluation Result

The answer statements of males and those of females are separately treated as described in Section 4.3.2. When *n*=3.5, mere one relevant answer is obtained on Love. When *n*=4.0, only two relevant answers are obtained on Auction and PC, and no answer is to be obtained on Love.

Precision and recall based on the proposed method evaluated by only males (females, respectively), those based on the average of scores, and those on the basis of distance are shown in Figure 1 (Figure 10), Figure 2 (Figure 11), and Figure 3 (Figure 12) for Auction, Figure 4 (Figure 13), Figure 5 (Figure 14), and Figure 6 (Figure 15) for PC, and Figure 7 (Figure 16), Figure 8 (Figure 17), and Figure 9 (Figure 18) for Love.

Most of the analysis eventually resulted in the same precision, while this tendency sometimes did not hold especially when the threshold was set 4.0.

5. Consideration

As a whole, precision generally shows better as the threshold *n* is set smaller. Regardless of gender, precision shows the best when *n*=2.0 because the number of relevant answers increases with the decrease of threshold *n*.

The proposed method can select top several users when *n*=3.5 and *n*=3.0, whereas cannot when *n*=4.0. When *n*=2.5, it is considered that the proposed method can sufficiently be used. Therefore, it is considered that the proposed method could successfully select the respondents that are more than almost averagely appropriate to a question.

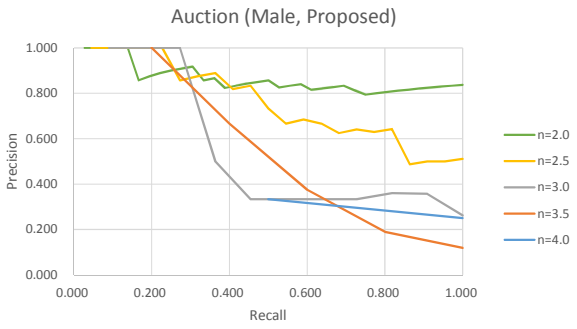

Figure 1: Result (Auction, Male, Proposed)

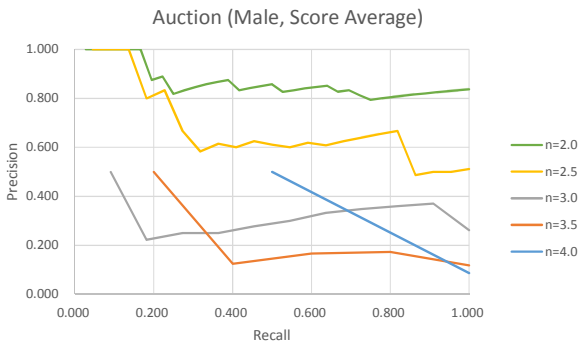

Figure 2: Result (Auction, Male, Score Average)

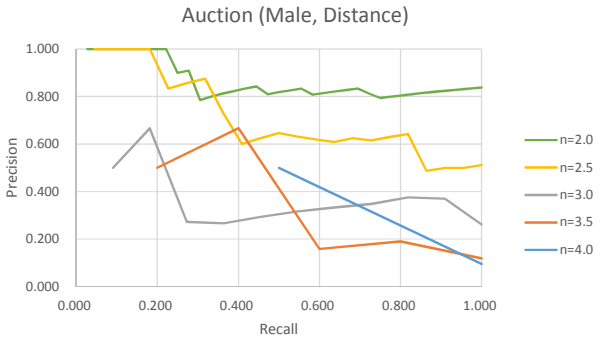

Figure 3: Result (Auction, Male, Distance)

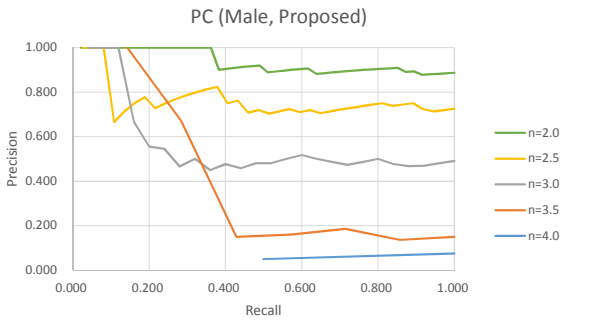

Figure 4: Result (PC, Male, Proposed)

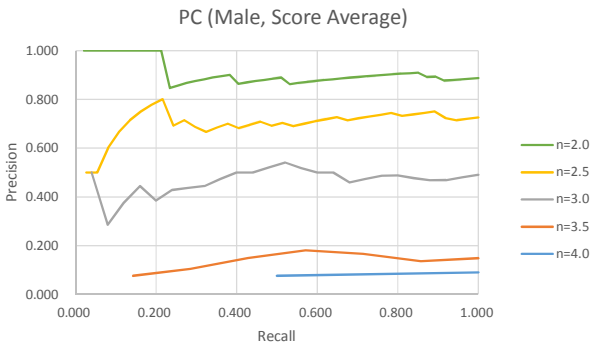

Figure 5: Result (PC, Male, Score Average)

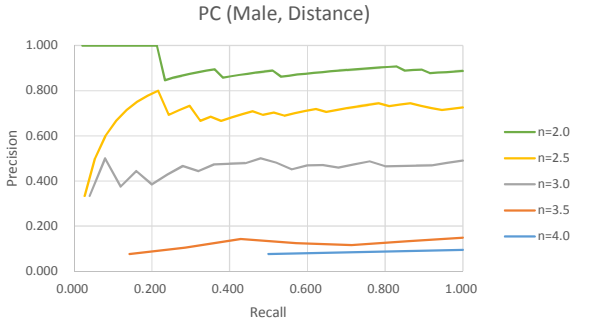

Figure 6: Result (PC, Male, Distance)

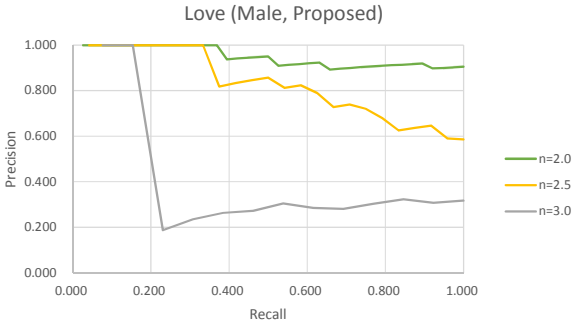

Figure 7: Result (Love, Male, Proposed)

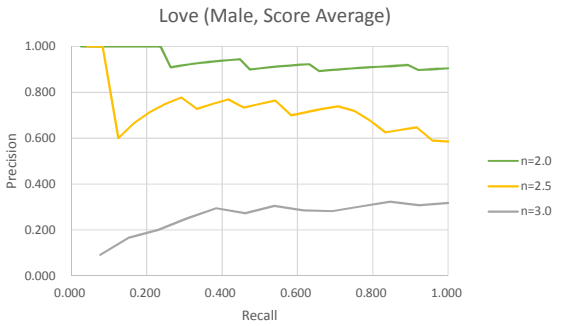

Figure 8: Result (Love, Male, Score Average)

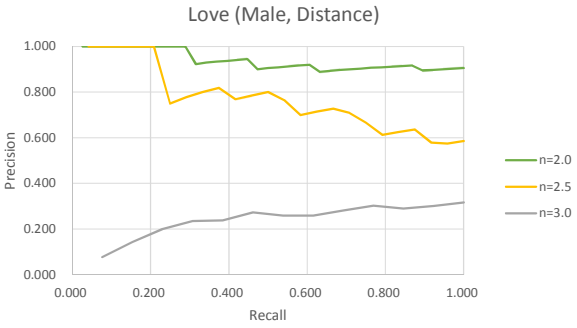

Figure 9: Result (Love, Male, Distance)

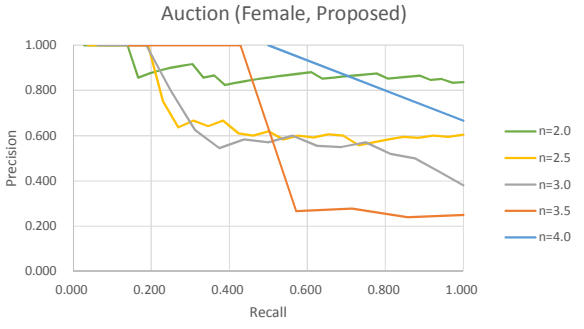

Figure 10: Result (Auction, Female, Proposed)

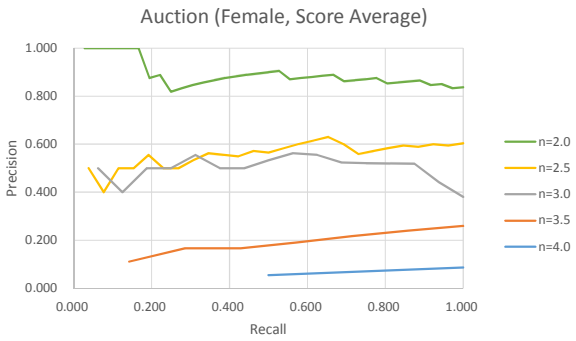

Figure 11: Result (Auction, Female, Score Average)

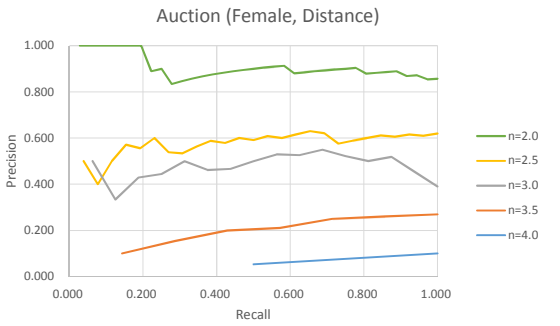

Figure 12: Result (Auction, Female, Distance)

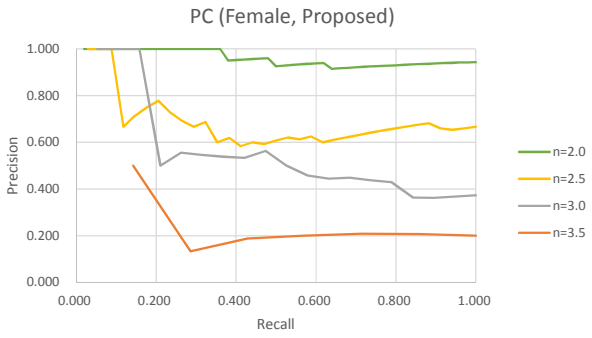

Figure 13: Result (PC, Female, Proposed)

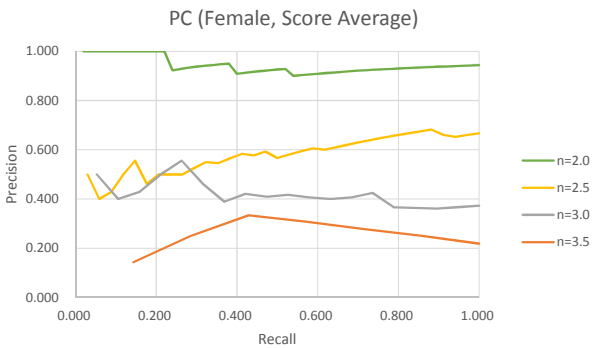

Figure 14: Result (PC, Female, Score Average)

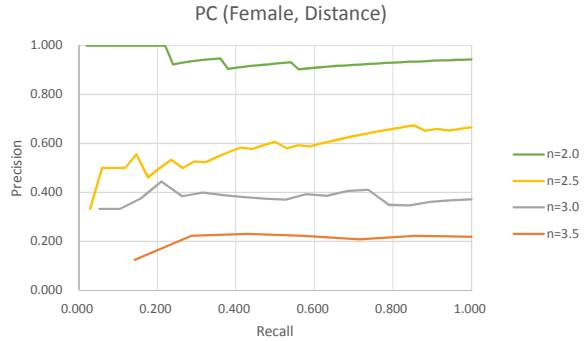

Figure 15: Result (PC, Female, Distance)

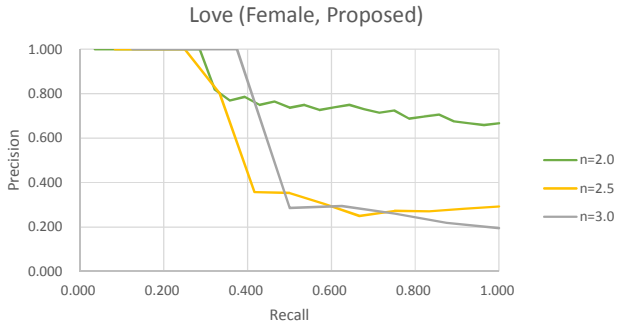

Figure 16: Result (Love, Female, Proposed)

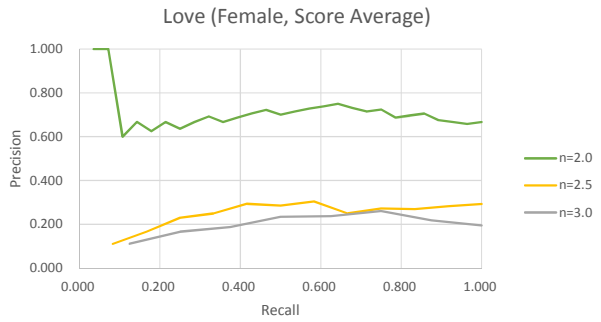

Figure 17: Result (Love, Female, Score Average)

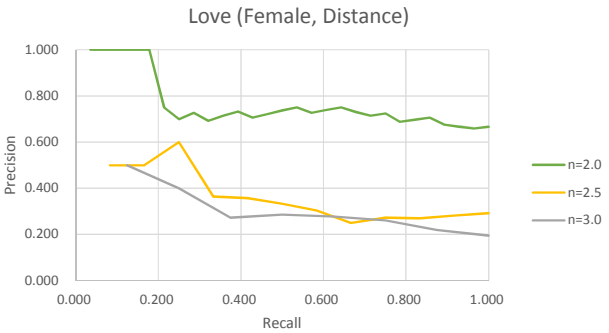

Figure 18: Result (Love, Female, Distance)

Compared with the methods using averaged scores and distances, the proposed method has higher precision. It is shown that the proposed method outperforms the other methods.

In the case of n=4.0 of female subjects, the proposed method shows the good performance. This may be caused by the few number of subjects, only three. The evaluation through more female subjects is required.

Auction shows a good performance when n=3.5 in the results on males for Auction. This is because the Auction category accounts for 38 answers posed by 11 unique users out of 100 answer statements used for the experiment. Meanwhile, PC (Love, respectively) category occupies only 2 (3) answers posted by 2 (2) unique users. For PC and Love, the answer statements whose categories are the same as a question one happened to be not included in the dataset used for the experiment and were not considered. It would be required to take categories into consideration at the stage of inspecting impression and suitability of questions and answers explained in Section 3.3.

6. Conclusion

This paper proposed the method of introducing appropriate respondents to questioners. The proposed method uses the number of appearance of respondents and the score based on the distance between the factor scores of a question and an answer already posted. The proposed method was compared with the methods using averaged scores and the distances through precision and recall for various relevant respondents. It has been shown that the proposed method outperformed the other methods. It was also shown that the proposed method could successfully select the respondents that are more than almost averagely appropriate to a question.

Estimating objective scores of answer statements is in future work. After the method of estimating objective scores as established, estimating Best Answers by using the estimated objective scores is also included in future work. Using characteristics of users, i.e., questioners and answerers, for finding appropriate answerers is also included in future work.

Acknowledgements

This research is partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 25280110, 2013-2016 and supported by the Japan Society for the Promotion of Science, Grant Number 26008587, 2014-2015. This research uses the data of “Yahoo! Chiebukuro” given to National Institute of Informatics by Yahoo Japan Corporation.

References

[1] Yahoo! Chiebukuro, http://chiebukuro.yahoo.co.jp/, 2014, in Japanese.

[2] 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, Software and Network Engineering,” pp.1-13, Springer, 2012.

[3] 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 (2014) No. 1 Special Issue on ISAE 2013, pp.19-26, 2014.

[4] Yuya Yokoyama, Teruhisa Hochin and Hiroki Nomiya, “Towards Detecting Appropriate Respondents to Questions Posted at Q&A Sites,” Submitted to International Symposium on Affective Science and Engineering 2015 (ISASE2015), 2015.

[5] Mohan John Blooma, Alton Y. K. Chua and Dion Hoe-Lian Goha, “Predictive Framework for Retrieving the Best Answer,” in the Proc. of 2008 ACM Symposium on Applied Computing (SAC08), pp.1107-1111, 2008.

[6] Xin-Jing Wang, Xudong Tu, Dan Feng and Lei Zhang, “Ranking Community Answers by Modeling Question-Answer Relationships via Analogical Reasoning,” Proceedings of the 32nd International ACM SIGIR Conference, pp.179-186, 2009.

[7] Yoko Nishihara, Naohiro Matsumura and Masahiko Yachida, “Understanding of Writing Style Patterns between Q&A in Knowledge Sharing Community,” The 22nd Annual Conference of the Japanese Society for Artificial Intelligence, 1H2-7, 2008.

[8] Lada A. Adamic, Jun Zhang, Eytan Bakshy and Mark S. Ackerman, “Knowledge Sharing and Yahoo Answers: Everyone Knows Something,” in the Proc. of 17th International World Wide Web Conference (WWW2008), 2008.

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

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

[11] Yahoo! Chiebukuro: Already solved Q&A, http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q13587188, 2014, Japanese.

[12] Yahoo! Chiebukuro: Already solved Q&A, http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q124489262, 2014, Japanese.

[13] Yahoo! Chiebukuro: Already solved Q&A, http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q141050114, 2014, Japanese.