Proposal for a Similar Question Search System on a Q&A Site

Katsutoshi Kanamori, Akinori Kanda, and Hayato Ohwada

Division of Industrial Administration
Graduate School of Science and Technology,
Tokyo University of Science,
2641 Yamazaki, Noda, Chiba, 278-8510, Japan
katsu@rs.tus.ac.jp, j7412607@ed.noda.tus.ac.jp and ohwada@rs.tus.ac.jp

ABSTRACT

There is a service to help Internet users obtain answers to specific questions when they visit a Q&A site. A Q&A site is very useful for the Internet user, but posted questions are often not answered immediately. This delay in answering occurs because in most cases another site user is answering the question manually. In this study, we propose a system that can present a question that is similar to a question posted by a user. An advantage of this system is that a user can refer to an answer to a similar question. This research measures the similarity of a candidate question based on word and dependency parsing. In an experiment, we examined the effectiveness of the proposed system for questions actually posted on the Q&A site. The result indicates that the system can show the questioner the answer to a similar question. However, the system still has a number of aspects that should be improved.

Keywords: Q&A site, Information retrieval, NLP, Tf-idf, Dependency parsing.

1. INTRODUCTION

Recently, Q&A sites have emerged as an Internet service. Using this service, a user(asker) posts a question on a Q&A site, and other users post answers. Major Q&A sites include Baidu Zhidao,[1] OKWave,[2] Quora,[3] and Yahoo! Chiebukuro[4] (Japanese Yahoo Answers[5]). Such Q&A site provide users information or results that they are not able to obtain by using simple search engines.

Although a search engine displays its results immediately with given keywords, a Q&A site cannot display the answer immediately. This delay in answering occurs because in most cases another site user is answering the question manually.

In contrast, the Q&A site has past data that consists of posted questions and answers. Thus without posting a question, users can see answers to a question with same meaning in the database. However, in general, the database is huge, and it is not easy to search for objective questions, because all questions and answers are written in natural language and they are not coordinated.

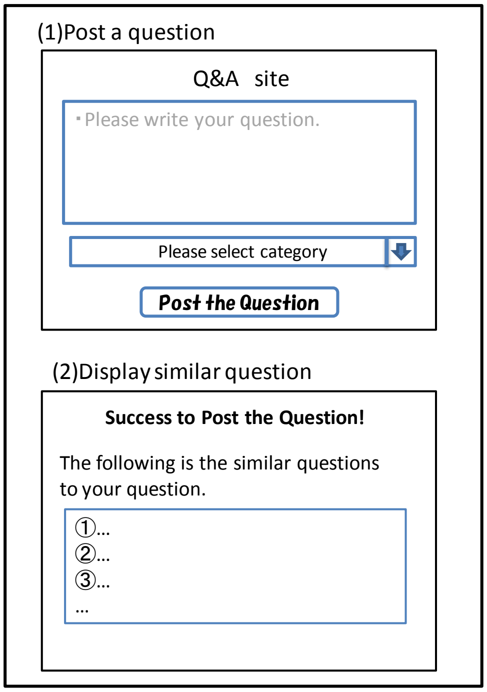
(1)Post a question
Q&A site
・Please write your question.
Please select category
Post the Question
(2)Display similar question
Success to Post the Question!
The following is the similar questions to your question.
①…
②…
③…
…

Fig. 1 System Image

In this study, we propose a system that can present past questions that are similar to a question posted by a user. Figure 1 illustrates the system. Screen (1) is used to post a question. A typical Q&A site has a space to enter a question, and a section for choosing the question category to which the question belongs. For example, a question about cooking belongs to "cooking", and a question about soccer belongs to "sports". Screen (2) displays all similar questions to the user. It should be noted that the system executes two tasks simultaneously: posting a question and searching for similar questions, users themselves do not need to search for similar questions, and they can see such questions immediately.

In order to create such a system, the following assumptions were considered in this research.

- The system is given only a sentence (a keyword is not given). Information retrieval generally uses a search by keywords. When we use the Internet to search for information, we usually use a search engine (e.g., Google or Yahoo!) by entering keywords. However, in the proposed system, a sentence is entered in natural language. Therefore, a method that does not search by keywords must be used, or keywords must be selected from the sentence.

[1] http://zhidao.baidu.com/

[2] http://okwave.jp/

[3] https://www.quora.com/

[4] http://chiebukuro.yahoo.co.jp/

[5] http://answers.yahoo.com/

- It is possible to perform parsing.
  The proposed system can parse the input sentence. In this research, the dependency relationships among words are investigated using dependency analysis.

We investigated what method is the most effective as a system, considering the above assumptions. This paper is organized as follows. In Section 2, we review related work. Section 3 presents the details of the proposed system. Section 4 presents our detailed experiment results, and Section 5 discusses them. Finally, Section 6 presents the conclusion, along with some directions for future work.

2. RELATED WORK

Studies on Q&A sites have been conducted for a variety of purposes. For example, some researchers have studied the prediction of the best answer and the clustering of questions.

In research on the best answer, Kim et al. investigated the standard by which the best answer is chosen [1]. Nishihara et al. suggest a method for judging an answer that is more likely to become the best answer when a question and certain answers are given [2]. Ishikawa et al. conducted research on what predicts a good-quality answer [3]. Rather than focusing on the best answer, they judge a good answer based on such qualities as detail, reason, and graciousness.

In research on clustering questions on a Q&A site, the classification method depends on the purpose. For example, Watanabe et al. classified questions into five types, such as seeking a fact and a reason [4] in order to recommend questions to a user who posts an answer. Harper et al. also classified questions on a Q&A site [5]. Their work involved the standards for being valued as archival. Long et al. classified questions into three types [6] in an effort to find a potential answer and question similar to the question a user posted. Tamura et al. classified questions having two or more sentences [7] in order to develop a question-answering system. To classify a text, they extract the most important sentence (“core sentence”) in a text, and the intention of a question is classified using this sentence.

Research has also been conducted to complement an answer by showing the user information related to a posted question from outside a Q&A site. For example, Nie et al. are conducting research that adds a picture and image information to a question, in order to give an answer more clearly [8].

We investigated the importance of whether an answer the system presents is appropriate for the user who posted the question. Therefore, our purpose differs from the research introduced in this section.

3. PROPOSED SYSTEM MODEL

The proposed system consists of calculation of similarity among four methods, and integration of the calculation result.

Calculation of similarity

We use four methods to calculate similarity: N-gram, the number of appearances of a morpheme, tf-idf, and dependency parsing.

(1) N-gram

N-gram is a method of separating the unit that specifies the text and counting its frequency of appearance. If the value of n is 1, it is a uni-gram. If N is 2, it is a bi-gram. If N is 3, it is a tri-gram. For example, the text "I studied at home yesterday" is separated by word unit and bi-gram, resulting in:
[I studied], [studied at], [at home], [home yesterday], [yesterday.], [.]

We separate each question by a character unit. When the above example is separated by a character unit, we get:
[I s],[s t],[t u]...[a y],[y .], [.]

We use bi-grams and tri-grams for calculating similarity. When a sentence S is given, let $g_2(S, c_i, c_j)$ be the number of appearances of a character unit $[c_i, c_j]$, and let $G_2(S)$ be its vector.

$$G_2(S) = (g_2(S, c_1, c_1), g_2(S, c_1, c_2), \cdots)$$

The similarity between sentences S_A and S_B, $SIM_{bi-gram}(S_A, S_B)$, is then defined as follows:

$$SIM_{bi-gram}(S_A, S_B) = \frac{G_2(S_A) \cdot G_2(S_B)}{|G_2(S_A)||G_2(S_B)|} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} G_2(S_A)_{\mathrm{i}} \cdot G_2(S_B)_{\mathrm{i}}}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (G_2(S_A)_{\mathrm{i}})^2} \times \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (G_2(S_B)_{\mathrm{i}})^2} \quad (1)$$

We can use a tri-gram case in the same way. $G_3(S)$ is a vector of $g_3(S, c_i, c_j, c_k)$, and $SIM_{tri-gram}(S_A, S_B)$ is defined as follows:

$$SIM_{tri-gram}(S_A, S_B) = \frac{G_3(S_A) \cdot G_3(S_B)}{|G_3(S_A)||G_3(S_B)|} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} G_3(S_A)_{\mathrm{i}} \cdot G_3(S_B)_{\mathrm{i}}}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (G_3(S_A)_{\mathrm{i}})^2} \times \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (G_3(S_B)_{\mathrm{i}})^2} \quad (2)$$

(2) The number of appearances of a morpheme

A morpheme is the minimum linguistic unit with a meaning. We calculate similarity using this method when a text is separated by a morpheme. The morphemes appearing in each text are compared. Our proposed system covers Japanese, referring to [9] for our Japanese morphological analysis. When a sentence S is given, let $freq(S, m_i)$ be the number of appearances of morpheme m_i and let $F(S)$ be a vector of these.

$$F(S) = (freq(S, m_1), freq(S, m_2), \cdots)$$

The similarity between sentences S_A and S_B, $SIM_{morph}(S_A, S_B)$, is then defined as follows.

$$SIM_{morph}(S_A, S_B) = \frac{F(S_A) \cdot F(S_B)}{|F(S_A)||F(S_B)|} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} F(S_A)_{\mathrm{i}} \cdot F(S_B)_{\mathrm{i}}}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (F(S_A)_{\mathrm{i}})^2} \times \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (F(S_B)_{\mathrm{i}})^2} \quad (3)$$

(3) Tf-idf

Tf-idf is a means of weighting the words in the text. Tf-idf is calculated based on the indexes of Term Frequency (tf) and Inverse Document Frequency (idf).

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \quad (4)$$

$$idf_i = \log \frac{|D|}{|\{d: d \ni t_j\}|} \quad (5)$$

$$tfidf_{ij} = tf_{ij} \cdot idf_i \qquad (6)$$

Here, n is the number of occurrences in question j of word I, D is the total amount of the question, and d is each question text. Let $TI(S)$ be a vector of sentence S's $tfidf$; then $SIM_{tfidf}(S_A, S_B)$ is defined as follows.

$$SIM_{tfidf}(S_A, S_B) = \frac{TI(S_A) \cdot TI(S_B)}{|TI(S_A)||TI(S_B)|} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} TI(S_A)_{\mathrm{i}} \cdot TI(S_B)_{\mathrm{i}}}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (TI(S_A)_{\mathrm{i}})^2} \times \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (TI(S_B)_{\mathrm{i}})^2}} \qquad (7)$$

(4) Dependency parsing
Dependency analysis is a method of investigating the dependency relationships among clauses.

To calculate similarity by dependency analysis using the method below, we first apply dependency parsing to a question. Next, the clause obtained is separated into morphemes. Finally, we investigate the dependency relationship between nouns and verbs (basic form) as well as adjectives.
For example, if the question sentence "I looked at the beautiful picture in the art museum yesterday." is given, we can find the following dependency relationship.
[I=>look], [yesterday=>look], [beautiful=>picture], etc.

At the risk of repetition, since our system covers only Japanese, English dependencies may differ.

When dependency is used, it becomes clear that the meanings of the following two sentences differ.
(1) “The woman saw a beautiful bird in the town.”
(2) “I saw the beautiful woman in the town.”
These two sentences seem to be similar when focusing on word frequency. However, since (1) is [beautiful =>bird] and (2) is [beautiful =>woman], they can be distinguished, and it can be determined that their meanings differ.

When a sentence S is given, let $dp(S, m_i, m_j)$ be the number of appearances of a dependency relationship [m_i => m_j], and let $DP(S)$ be its vector.

$$DP(S) = (dp(S, m_1, m_1), dp(S, m_1, m_2), \cdots)$$

Then, the similarity between sentences S_A and S_B, $SIM_{dp}(S_A, S_B)$, is then defined as follows.

$$SIM_{dp}(S_A, S_B) = \frac{DP(S_A) \cdot DP(S_B)}{|DP(S_A)||DP(S_B)|} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} DP(S_A)_{\mathrm{i}} \cdot DP(S_B)_{\mathrm{i}}}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (DP(S_A)_{\mathrm{i}})^2} \times \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (DP(S_B)_{\mathrm{i}})^2}} \qquad (8)$$

All methods are calculated using cosine similarity, which is a measure of the similarity between two vectors of an inner product space that measures the cosine of the angle between them. In this study, the two vectors are two question sentences. The minimum value of cosine similarity is 0 (they are not alike at all), and the maximum value is 1 (they are completely the same.). Given two vectors for attributes A and B, the cosine similarity cos(θ) is represented using the dot product and magnitude as

$$\cos(\theta) = \frac{\mathrm{A} \cdot \mathrm{B}}{|\mathrm{A}||\mathrm{B}|} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{A_i} \cdot \mathrm{B_i}}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{A_i})^2} \times \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{B_i})^2}} \qquad (9)$$

Integration method for the result

We integrate the results of the similarity calculation in the preceding section. We present a question that was calculated as similar using each method and has a higher value than the set-up threshold value. This is set up for each method. Moreover, we determine its value from the results of a preliminary experiment.

4. EXPERIMENT

Dataset

In our experiments, we collected data from questions on the Yahoo! Chiebukuro site, which contains 3,116,009 questions posted between April 2004 and April 2009.We chose the questions classified as "Internet." We then randomly selected 15,000 data points and used them as a data set.

Preliminary experiment

The preliminary experiment was conducted to determine the threshold, as mentioned in discussion of the integration method for the result. First, we randomly chose ten questions from the data set. Next, we calculated their similarity by applying each method to each question. Finally, we evaluated them manually. A valuation method is used to determine the appropriateness of a question. The question presented to the system (i.e., the question posted by the user) is $\mathrm{Q_p}$; the question that was judged by the system as being similar to $\mathrm{Q_p}$ is $\mathrm{Q_s}$; and the answer to $\mathrm{Q_s}$ is $\mathrm{A_s}$. We then evaluate whether $\mathrm{A_s}$ is an appropriate answer to $\mathrm{Q_p}$. (See the example in Table 1.)
We determine the threshold based on the following standards. This is the average value of similarity. We list the threshold for each method in Table 2.

Table 1. Example of a posted question, a similar question, and its answer

Posted Question	“Please tell me how to ask a question on the Q&A site.”
Similar Question	“What should I do to ask a question on a Q&A site? Which category should be chosen in order to get a good answer? ”
Answer	“First, you enter a question. Next, you choose a category. Finally, you click the button marked [post your question] to post your question. I think that the tips that you can use to get a good answer are choosing a category suitable for your question and inputting your question in detail.”

Table2. Thresholdof each method

Method	Threshold
bi-gram	0.404
tri-gram	0.308
morpheme	0.346
tf-idf	0.440
dependency parsing	0.136

Experiment result

We randomly chose ten questions from a data set and calculated their similarity using each method. We define a question whose value exceeds the threshold determined in the preceding section as a similar question. An evaluator evaluates these questions. The valuation basis is whether A_s is an appropriate answer to Q_p. We calculated precision using the following formula.

$$\mathrm{Precision} = \frac{\#\mathrm{total\ correct\ answers}}{\#\mathrm{total\ \ answers\ judged\ by\ the\ system}} \quad (10)$$

The result is precision = 0.24.

Table3. Correct example

Posted Question	"What is the concrete difference between a 'blog' (a currently popular word), and a 'BBS'? "
Similar Question	"What is a blog? Is it different from a BBS? "
Answer	"A blog is a web diary. Ａ BBS is a web page written by the general public."

Table4. Incorrect example

Posted Question	"What is the concrete difference between a 'blog' (a currently popular word), and a 'BBS'? "
Dissimilar Question	"What is the difference between a comment and a trackback on a blog?"
Answer	"A comment expresses an opinion and feedback about an article within the blog. Ａ trackback is a notice issued when others refer to the article."

5. DISCUSSION

Correct example

Table 3 provides a correct example. The answer to a similar question is an appropriate answer to the posted question.

Incorrect example

Table 4 provides an incorrect example. The system has judged a question that is not similar to be similar. We think this error is caused by the method used to set the threshold.

Idea for an improvement

To improve the threshold, we considered how to calculate the average of similar question pairs**.** Using such a method, we manually prepare pairs of questions that are similar, and the average of these similarities is defined as a threshold.

In addition, we considered each technique separately this time. Therefore, we think that precision may be improved using an ensemble method that mixes the techniques. We calculate similarity using the following formula.

$$\mathrm{similarity} = \frac{\sum_{i=1}^{n} V_i}{n} \quad (11)$$

Here, V_i is the value that normalizes the similarity in method i. In the future, we will conduct experiments to confirm the effectiveness of this method.

Furthermore, the system sometimes judges a question to be different from a similar question with regard to intention. For example, the question "Who is Washington?" is different from "Where is Washington?" However, because “Washington” appears in both questions, they might be judged to be similar. We may be able to prevent such incorrect conclusions if we automatically (or manually) add a tag to the question. For example, we may tag the first question as "a question about a place" and the second question as "a question about a person." We would also like to confirm the effectiveness of this approach.

6. CONCLUSIONS

In this study, we proposed a system that can present questions that are similar to a question posted by a user. The proposed system measures the similarity of the question sentence by calculating the cosine similarity based on bi-grams, tri-grams, morphemes, tf-idf, and dependency parsing. The experiment result demonstrates that the system can show a questioner the answer to a similar question. Our system enables us to search for the answers we want in a short time.

However, the system still has a number of aspects that should be improved. In the future, we would like to reconsider the method for setting the threshold and to develop a new method for distinguishing questions.

7. REFERENCES

[1] Soojung Kim, Jung Sun Oh, Sanghee Oh,"Best-Answer Selection Criteria in a Social Q&A site from the User-Oriented Relevance Perspective", **Proceedings of the American Society for Information Science and Technology,** Volume 44, Issue 1, pp.1–15, 2007.

[2] Yoko Nishihara, Naohiro Matsumura, 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**, 2008.

[3] Daisuke Ishikawa, Tetsuya Sakai, Yohei Seki, Kazuko Kuriyama, Noriko Kando, "Automatic Prediction of High-Quality Answers in Community QA ", **The Japanese Society for Artificial Intelligence**, Vol. 21, No. 3, 2011, pp.362-382.

[4] Naoto Watanabe, Satoshi Shimada, Yohei Seki, Noriko Kando, and Tetsuji Satoh, "A Study for Questions Classication based on Questioner Demands in QA Communities", **DEIM Forum**, 2011.

[5] F. Maxwell Harper, Daniel Moy, Joseph A. Konstan, "Facts or Friends? Distinguishing Informational and Conversational Questions in Social Q&A Sites", **CHI Q&A Systems**, 2009, pp.759-768.

[6] Long Chen, Dell Zhang, Mark Levene, "Understanding User Intent in Community Question Answering", **WWW 2012 – CQA'12 Workshop**, 2012, pp.823-828.

[7] Akihiro Tamura, Hiroya Takamura, Manabu Okumura, "Classification of Multiple-sentence Questions", **Information Processing Society of Japan**, Vol. 47, No. 6, 2006, pp. 1954-1962.

[8] Liqiang Nie, Meng Wang, Zheng-Jun Zha, Guangda Li and Tat-Seng Chua, "Multimedia Answering: Enriching Text QA with Media Information", **SIGIR '11**, 2011, pp.695-704.

[9] Y. Matsumoto, A. Kitauchi, T. Yamashita, andY. Hirano. Japanese morphological analysis system chasen version 2.0 manual. Technical report, **NAIST**, 1999.