COMPLEMENTARY QA ANALYSIS FOR QUESTION-ANSWERING WEBSITES

Yu-Hsuan Chen, Pei-Jung Lu, Duen-Ren Liu*
Institute of Information Management
National Chiao Tung University, Hsinchu 300, Taiwan
dliu@mail.nctu.edu.tw

ABSTRACT

With the ubiquity of the Internet and the rapid development of Web 2.0 technology, Question Answering (QA) websites have become extremely popular knowledge sharing platforms. As the number of posted questions and answers continues to increase rapidly, the massive amount of question-answer knowledge is causing information overload. The problem is compounded by the growing number of redundant QAs. QA websites, such as Yahoo! Answer, are open platforms where users can ask or answer questions freely. Users may also wish to learn more about the information provided in an answer, so they can use related keywords in the answer to search for extended complementary information. In this paper, we propose a novel approach to identify complementary QAs of a target QA. We define two types of complementation - partial complementation and extended complementation. We utilize a decision-tree classification approach to construct a classification model and predict complementary relationships between QAs based on three measures: question similarity, answer novelty, and answer correlation. The results of experiments conducted on a dataset collected from Yahoo! Answers Taiwan show that the proposed approach can identify complementary QAs effectively.

Keywords: Knowledge complementation, Information novelty, Mutual information, Question-Answering Websites.

1 INTRODUCTION

With the ubiquity of the Internet and the rapid development of Web 2.0 technology, increasing numbers of individuals and organizations are searching for needed information on the Internet. The growth of Web 2.0 has enabled QA websites to become important knowledge sharing platforms, which accumulate question-answer knowledge through the mechanism of question posting and answering. The Yahoo! Answer Taiwan website (also called Yahoo! Knowledge Plus) is a community-driven knowledge website, where users can share their experience and exchange knowledge by asking and answering questions. Users can browse the questions that other users have asked, search for answers to particular questions, or post questions and wait for answers.

* Corresponding author.

QA websites are becoming increasingly popular knowledge sharing platforms because users can post natural language questions, as well as share miscellaneous information or obtain answers to their questions directly from the website. User participation and sharing have enriched the information resources of such websites. As the number of questions and answers is increasing rapidly, the massive amount of question-answering knowledge is causing information overload. To solve the problem, QA systems provide different functions to help people find required information. For example, users can post questions directly, or they can use the "keyword search" function to find and browse questions and answers of interest. Sometimes the answer to a question may only provide partial information, so users may wish to browse relevant QAs with partial complementation to get complete information. However, some relevant QAs may be redundant, because QA websites, such as Yahoo! Answer, are open platforms that allow users to ask or answer questions freely. On the other hand, if users wish to learn more about the information provided in an answer, they can use keywords in the answer to search for extended complementary information.

A great deal of research on Question-Answering websites has focused on finding experts to answer target questions [7, 9], or finding high quality answers [2, 8, 14]. By contrast, there has been relatively little research on finding complementary information. One line of research in this area exploits users' frequent QA browsing behavior to find related QAs [4]. In addition, a topic-structure-based complementary information retrieval approach has been proposed to help users retrieve complementary Web pages that augment the content of video or television programs [10]. Existing works do not address the issue of finding complementary QAs with partial or extended complementation. Most traditional QA systems use keywords to find relevant QAs without considering the issues of redundancy or complementation. However, users' information-seeking activities are becoming more sophisticated. Thus, besides returning relevant answers to questions through keyword search mechanisms, it is important that QA systems provide users with complementary QAs.

In this paper, we propose a novel way to identify complementary QAs of a target QA. We define two types of complementation - partial complementation and extended complementation. A partially complementary QA contains a partial answer to a related question. It provides a different perspective on the target QA's original answer. Thus a partially complementary QA supplements the information retrieved by the target QA. An extended complementary QA, which provides further information on unclear parts of the target QA's answer, contains extended information to enhance the target QA's answer. We use a decision-tree classification approach to build a classification model, and then predict the complementary relationships between QAs based on three criteria: question similarity, answer novelty, and answer correlation.

2 RELATED WORK

2.1 Question Answering Systems

Question answering (QA) systems, are platforms where users can share and exchange knowledge publicly, freely and conveniently. Users can ask any kind of question and hopefully receive answers from other users; or they can use the system's search function to look for information. Yahoo! Answers is one of the most well-known QA systems. It is also

called as Yahoo! Knowledge in Taiwan, where it is the most popular QA system. As mentioned earlier, we use data from Yahoo! Knowledge in our experiments. How to find high quality answers in question answering websites is a popular research issue. Basic studies use statistics to find the content factors that may influence the selection of the best answers [8]. More comprehensive studies have exploited non-textual and textual features to identify high quality answers [2, 3, 6]. Suryanto et al. [14] proposed a quality-aware framework that considers the relevance and quality of answers derived from answer features and answerers' expertise.

2.2 Information Novelty

The emergence of Web 2.0 has enabled Internet users to share information more easily, resulting in the rapid accumulation of huge amounts of information. Although much of the information shared by users is new or different, it is inevitable that some content will be repeated in different documents. Consequently, a search for information on a particular topic may yield several documents that contain redundant information. Information novelty can be measured by the degree of overlap between two documents, i.e., the number of terms that appear in both documents [16]. It can also be inferred as the inverse of similarity; that is, the greater the similarity between two documents, the lower will be the novelty of the information they provide. For example, Collins-Thompson et al. [5] used the cosine similarity to measure information novelty. Language models have also been adopted to measure information novelty [17].

2.3 Complementary Information Retrieval

Identifying complementation is a subjective process that depends on whether the user perceives the information as useful. Ma and Tanaka [10] measure the complementary degree between two Web pages by using the concept of topic-structure, which is represented by a directed acyclic topic graph. Their model uses a topic corpus to identify the subject terms and content terms of a topic, and then generates a topic structure for each web page based on the relationships between the two types of terms. Two topic-graphs can be combined to form a join graph if they have at least one node that is the same. The model measures the difference between the original topic-graph and the join graph as the complementary degree, and provides a means of quantifying the complementation. It infers that the complementary degree will be high if two web pages have a significant amount of "novel" information and a small amount of similar content. The approach requires a reliable topic corpus that can identify the subject terms and content terms of a topic. However, deriving a reliable topic corpus for QA websites is a difficult because of the huge number of QAs on various kinds of topics and subject terms. In addition, the quality of some QAs may be poor due to the open platform nature of QA websites. Thus, it is not feasible to extract appropriate topics and subject terms from QA websites.

2.4 Information Measures

In this work, we propose a method that identifies the complementary relationships between QAs based on three criteria, namely, question similarity, answer novelty, and answer correlation. In general, two QAs are complementary if their answers correlate to some extent and are not redundant. We adopt mutual information and all-confidence measures to determine the correlation between the answers of two QAs. Mutual information is a quantity that measures the mutual dependence of the two variables in probability theory and information theory [11]. Formally, the mutual information of two discrete random variables, *X* and *Y*, can be defined as follows:

$$MI(X,Y) = \sum_{y \in Y} \sum_{x \in X} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \quad \text{Eq. 1}$$

where $P(x,y)$ is the joint probability density function of *X* and *Y*; and $P(x)$ and $P(y)$ denote the marginal probability density functions of *X* and *Y* respectively. The mutual information is applied in document clustering. Wei and Yang [15] proposed an context similarity estimation method that employs World Wide Web (WWW) as the information source to estimate the similarity between two sets of term. They issue three queries to a search engine (particularly, Google in their study) and obtain the number of hits (matching documents) returned for each query. They estimated the relevance weight between a pair of terms q_i and q_j by the pointwise mutual information (PMI) measure.

The all-confidence metric [13], an alternative interest measure for association rules, can also be used to measure the association degree of items. Let *Z* be a set of items. The all_confidence measure (*Z*) is defined as follows:

$$\text{all\_confidence}(Z) = \frac{P(Z)}{\max_{x \in Z}(P(x))} \quad \text{Eq. 2}$$

The denominator in Eq. 2 is the probability (support) of the item with the highest probability in *Z*. The all-confidence metric is used to determine if all the rules generated from *Z* have at least a confidence of all-confidence(*Z*). The higher the value of all-confidence, the closer will be the association of items in *Z*.

3 PROPOSED COMPLEMENTARY QA ANALYSIS

In this section, we discuss the proposed approach for finding complementary QAs of a target QA, and explain the concept of partial complementation and extended complementation. We also consider the decision-tree classification method used to identify complementary relationships between QAs based on three criteria: question novelty, answer novelty, and answer correlation.

3.1 Overview of the Approach for Finding Complementary QAs

Below, we explain the rationale behind the approach. Users normally use the “keyword search” function in the search engines of QA systems to find and browse questions and

answers (QAs) of interest. As some questions in a system are related, users often wish to browse the QAs of related questions. The information provided in the answer part of a target QA may be partial and incomplete, so the user may wish to search for related QAs to get complete information. However, the information in some related QAs may be redundant to the target QA and of no interest to the user. QAs that provide related information that is not redundant are called partially complementary QAs of a target QA. Moreover, some information in the target QA's answer may not be clear, so the user may wish to conduct an extended search by using keywords in the original answer to search for related QAs that contain extended complementary information. Such QAs are called extended complementary QAs of a target QA. Examples of partial complementation and extended complementation are shown in FIGURE 1. More specifically, a partial complementary QA provides a different perspective on the answer part of the target QA; thus, it supplements the original answer by making up for insufficient information in the target QA. On the other hand, an extended complementary QA provides further information that clarifies some aspect of the original answer. It contains extra information that extends and improves the original answer; thus, it is an extended complement of the target QA.

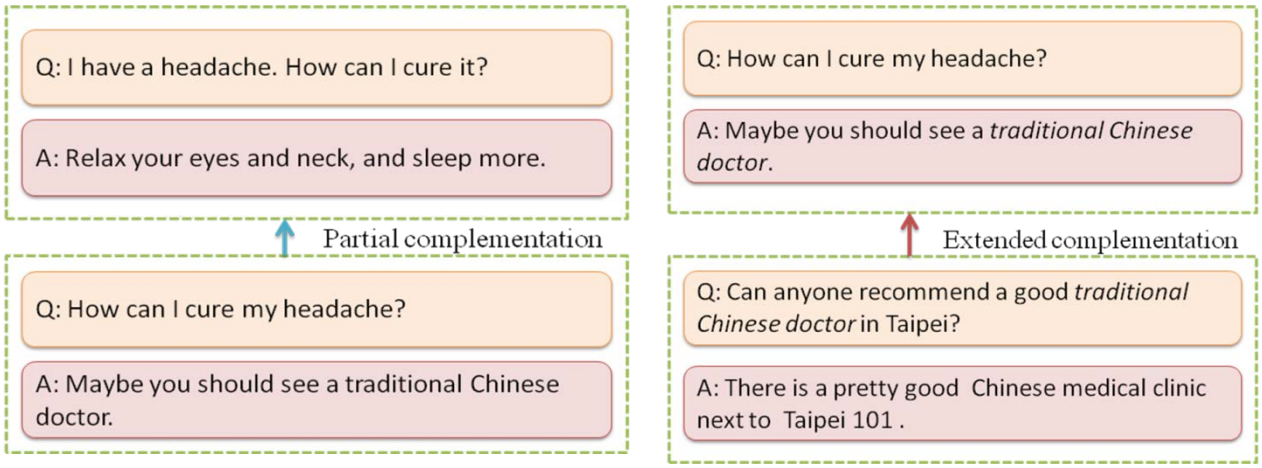

FIGURE 1. Examples of partial and extended complementation

To determine the type of relationship, we use three criteria: question similarity, answer novelty, and answer correlation. Given two QAs, suppose one is called the target QA and the other is called a candidate QA. If the question similarity score is high, it implies that the two questions are related; and if the answers are not redundant, they are regarded as novel and partial complementation is inferred. On the other hand, if the question similarity is low, the two questions are different; thus, we have to check if any term appears in both the answer of the target QA and the question of the candidate QA. If such a term exists, we consider that the candidate QA may contain some information that can explain the unknown subject (term) in the target QA's answer.

However, the answers of the two QAs may be redundant or unrelated, so we have to check the answer novelty and correlation between the target QA and the candidate QA. Answer novelty is measured by the inverse of the answer similarity; and the answer correlation is measured by the correlation of terms in the answers of two QAs. Extended complementation can be inferred if the answer novelty and answer correlation are high.

We utilize a decision-tree classification approach to build a classification model and predict the complementary relationships between QAs based on three criteria: question similarity, answer novelty, and answer correlation.

FIGURE 2 shows the procedure for identifying the complementary QAs of a target QA. The procedure involves two steps: data preprocessing and identifying complementary QAs. In the first step, QAs are preprocessed to extract their knowledge subjects. We use the vector space model [11] for representing the content of QAs as vectors. In the second step, we compare each candidate QA with the target QA to calculate its similarity, novelty, and correlation to the target QA. Then, we can determine the type of complementation to the target QA.

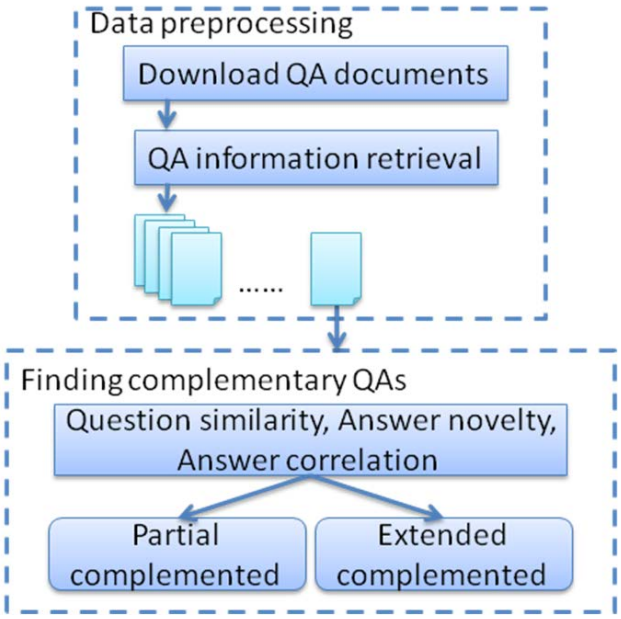

FIGURE 2. The procedure for identifying complementary QAs

3.2 Data Preprocessing

We use the TF-IDF approach [12] to analyze the title, question description, and answer of each QA and extract important terms that represent the knowledge subjects of the QA. The data pre-processing steps are Chinese Knowledge and Information Processing (CKIP)[1], removal of stop words, and calculation of the TF-IDF of each term. CKIP contains a corpus of about one hundred thousand terms, which are used as a base to automatically segment an article into meaningful terms and the corresponding parts-of-speech tags (POST). We use the POST to filter out unimportant words, and only consider words tagged as nouns, verbs or foreign words. As some words, such as pronouns, are not suitable to represent the original article, we have to compile a stop word list to remove those words in this step. TF-IDF, which is used to derive the weights of terms in a QA, can be calculated for a given category or a whole data set. In this step, we use a term vector to represent the knowledge subjects (terms) in the question title, question content, and answer fields of a QA. There are two ways

[1] CKIP is a system developed by the Chinese Knowledge and Information Processing (CKIP) Group at Academia Sinica, Taiwan.

to derive the term vector. The first calculates a term's weight based on the term's frequency in the QA without considering where the term occurs (i.e., the field). The second method ranks a term's importance (weight) according to the field where the term is located (i.e., question title, question description or answer field).

3.3 Identifying Complementary Relationships

In this section, we discuss the proposed model for identifying the complementary relationships between QAs. As mentioned earlier, we use three criteria, question similarity, answer novelty, and answer correlation, to determine the type of complementation between two QAs. A QA is comprised of a question, including the title and description, and an answer. Given a target QA, qa_t and a candidate QA, qa_c, the question similarity represents the degree of similarity between qa_t's question and qa_c's question. The answer novelty denotes the degree of novelty between qa_t's answer and qa_c's answer; and the answer correlation denotes the degree of correlation between qa_t's answer and qa_c's answer.

FIGURE 3 shows the procedure used to identify the two types of complementary QAs. Generally, two QAs that are partially complementary should have high question similarity and high answer novelty. High question similarity indicates that the QAs' questions are related, and high answer novelty confirms that the answers are not redundant. Questions that have extended complementation should generally have low question similarity, high answer novelty and high answer correlation, implying that (1) the questions are different; and (2) the answers correlate to some extent, but they are not redundant. We also assume that the extended complementary QA qa_c contains some extended information from qa_t. Thus, the question of qa_c should contain at least one term that appears in the answer of qa_t.

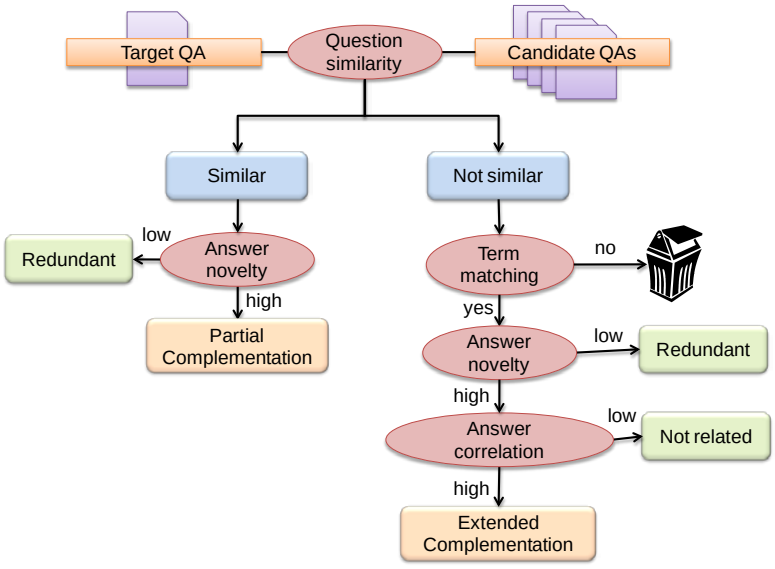

FIGURE 3. The rationale to identify the two types of complementary QAs

3.3.1 Rationale for identifying partial complementation

The left-hand side of FIGURE 3 shows the rationale for analyzing partial complementation. We use the cosine similarity measure to determine the degree of similarity between a target question and a candidate question. If the question similarity is high, the questions of the two QAs are related, so we analyze their answers to derive their answer novelty. Let qa_t^A and qa_c^A denote the answers of the target QA qa_t and the candidate QA qa_c respectively. We measure the novelty of the two answers, qa_t^A and qa_c^A by Eq. 3. We use the term vectors generated by TF-IDF to measure the cosine similarity between the answers of the two QAs. If the similarity is high, it means that the answers contain a lot of common information, so their novelty is low.

$$Novelty(qa_t^A, qa_c^A) = 1 - sim(qa_t^A, qa_c^A) \quad \text{Eq. 3}$$

3.3.2 Rationale for identifying extended complementation

The right-hand side of FIGURE 3 shows the rationale to identify cases of extended complementation. If the question similarity of two QAs is low, we cannot be certain that they are related. Assume that a candidate QA qa_c is analyzed for the extended complement of target QA qa_t.

First, we check if any terms in the answer of qa_t match terms in the question of qa_c. If a user would like to obtain further information about some parts or terms in the answer of qa_t, he can conduct an extended search by using certain terms in qa_t's answer. If any terms in the answer of qa_t and the question of qa_c match, qa_c might be the search result of the user's extended search; thus, it is an extended complementary QA candidate of qa_t. After checking for matching terms, we still have to ensure that the answer of qa_c contains novel information to avoid redundancy in the answers of qa_t and qa_c. The answer novelty of the two QAs can be measured by Eq. 3.

Although there is term matching between qa_t and qa_c, the QAs might be too different to be complementary QAs. We use two methods to measure the answer correlation between two QAs. One is based on mutual information (MI) and the other is based on all-confidence. MI is a quantity that measures the mutual dependence of two variables [11]. Using documents returned by Google's search engine, we measure the dependence of two terms by the number of documents that contain the two terms. Let $p(x \wedge y)$ denote the probability that two documents contain both term x and term y; and let $p(x)/p(y)$ denote the probability of documents containing term x / term y. In addition, let S_t^A / S_c^A denote the term set of qa_t^A / qa_c^A. The mutual information of the two answers, qa_t^A and qa_c^A, denoted by $MI(qa_t^A, qa_c^A)$, is measured by Eq. 4:

$$MI(qa_t^A, qa_c^A) = \sum_{x \in S_t^A} \sum_{y \in S_c^A; y \neq x} P(x \wedge y) \times \log_2(\frac{P(x \wedge y)}{P(x)P(y)})$$
$$= \sum_{x \in S_t^A} \sum_{y \in S_c^A; y \neq x} \frac{hits(x \wedge y)}{N} \times \log_2(\frac{N \times hits(x \wedge y)}{hits(x)hits(y)}) \quad \text{Eq. 4}$$
$$P(x \wedge y) = \frac{hits(x \wedge y)}{N}, P(x) = \frac{hits(x)}{N}, P(y) = \frac{hits(y)}{N}, x \neq y$$

where $hits(w)$ is the number of hits of word w returned by the search engine; $hits(x \wedge y)$ is the number of hits of word x and word y returned by the search engine; and N is the total number of documents in the repository. Because the exact value of N in the WWW environment is difficult to estimate, we employ an alternative approach that sets N as the largest hit value among all the terms we use to measure the mutual information.

Besides mutual information, we use the all-confidence metric [13] to measure the answer correlation, as shown in Eq. 5. The higher the value of all-confidence ($\{x,y\}$), the closer will be the association of x and y. The correlation between the two answers, qa_t^A and qa_c^A, denoted by $AC(qa_t^A, qa_c^A)$, is derived by summing the all-confidence ($\{x,y\}$) scores for $x \in S_t^A$ and $y \in S_c^A$. Note that S_t^A / S_c^A is the term set of qa_t^A / qa_c^A.

$$AC(qa_t^A, qa_c^A) = \sum_{x \in S_t^A} \sum_{y \in S_c^A; y \neq x} \frac{P(x \wedge y)}{\max(P(x), P(y))}$$
$$= \sum_{x \in S_t^A} \sum_{y \in S_c^A; y \neq x} \frac{hits(x \wedge y)}{\max(hits(x), hits(y))} \quad \text{Eq. 5}$$
$$P(x \wedge y) = \frac{hits(x \wedge y)}{N}, P(x) = \frac{hits(x)}{N}, P(y) = \frac{hits(y)}{N}, x \neq y$$

3.4 Decision Tree Classification

In general, the type of complementation can be determined according to the rationale shown in Figure 3. However, there may be complex situations that make the task difficult. Moreover, it is difficult to set thresholds for the three criteria. Accordingly, we use a decision tree classification approach to build a classification model and predict the complementary relationships between QAs.

Decision tree learning is widely used in the data mining field because it is easy to understand and interpret, and it can handle numerical and categorical data. We use the decision tree classification approach to build a model that can predict the complementary relationships between two QAs based on three input variables: question similarity, answer novelty, and answer correlation.

Specifically, we use Weka's Classification and Regression Tree (CART) model to build a classification model. We use CART because our input variables are numerical and the predicted complementary type is categorical. We train partial and extended complementary classification models separately, as their input variables are different. The complementary relationship of a target QA and a candidate QA can be determined by the classification

models. FIGURE 4 shows the process used to identify the complementary relationship based on the classification models for partial complementation and extended complementation respectively.

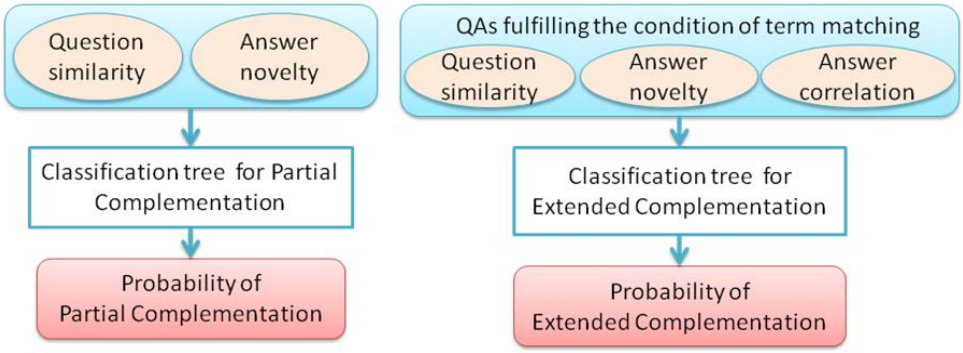

FIGURE 4. The process for identifying complementary relationships

In a decision tree, the training cases on a leaf node may not have the same class label. Besides predicting the class labels (partial or extended complement) of the complementary relationship between the two QAs, we calculate the probability that the predicted relationship will have a positive class label (partial complementation or extended complementation). The probability is measured as the ratio of the number of training cases with positive class labels on the leaf node to the total number of training cases on the node.

4 EXPERIMENT EVALUATIONS

We evaluated the performance of the proposed classification model in predicting the level of complementation between QAs. We compared the classification performance of using both question similarity and answer novelty with that of using question similarity alone to predict partial complementation; and we assessed the classification performance of using different answer correlation measurements to predict extended complementation.

4.1 Data Collection

In the experiments, we used data collected from Yahoo! Answer Taiwan. We chose 181 medical keywords and selected the top 20 QAs returned by the search with each keyword. Because too much data would have overburdened the human judges, we selected 250 QA relationships (pairs of QAs) to evaluate the performance of the decision tree classification model. We evaluated partial complementation and extended complementation separately. TABLE 1 shows the dataset used in the evaluation. The total number of QAs used to assess extended complementation is the number of QA pairs that satisfy the term matching conditions.

TABLE 1. Dataset used for the evaluation of the complementation classification model

	Partial	Extended
Complement	63	80
Not complement	187	71
Total	250	151(term matching)

4.2 Evaluation Metrics

To evaluate the performance of the classifiers, we use two standard classification performance metrics, the *precision rate* and *recall rate*, which are widely used in the field of information retrieval [1, 11]. For a complementation type *i*:

$$\text{Precision}(i) = \frac{\#\text{ of correctly identified QAs of type } i}{\text{total \# of QAs identified as type } i}$$

$$\text{Recall}(i) = \frac{\#\text{ of correctly identified QAs of type } i}{\text{total \# of QAs in type } i}$$

Finally, to obtain a single performance measure, we used the *F_1-measure* to balance the *precision* and *recall* scores:

$$\text{F}_1\text{-measure}(i) = \frac{2 \times \text{precision}(i) \times \text{recall}(i)}{\text{precision}(i) + \text{recall}(i)}$$

4.3 Experiment Results and Implications

We generate two decision tree classification models (partial and extended) with a publicly available data mining tool called Weka, and assess the performance of the model by the precision and recall rates. In addition, we use 10-fold cross-validation to evaluate the classification models as well as the precision and recall rates.

For partial complementation, we compare the performance of using question similarity alone against using both question similarity and answer novelty as input variables. TABLE 2 shows the performance of the partial complementation classification model under different input variables. We focus on the performance of predicting complementary QAs. The partial complementation results derived by using both question similarity and answer novelty yields a better performance than using question similarity alone.

TABLE 2. The performance of the partial complementation classification model

Partially complementary decision tree		Question Similarity & Answer Novelty	Question similarity (No answer novelty)
Complementary QAs	Precision	0.820	0.618
	Recall	0.794	0.667
	F_1-measure	0.806	0.641
Non-complementary QAs	Precision	0.931	0.884
	Recall	0.941	0.86
	F_1-measure	0.936	0.872
Average	Precision	0.903	0.817
	Recall	0.904	0.811
	F_1-measure	0.903	0.814

To identify extended complementary relationships between QAs, we use three criteria: question similarity, answer novelty, and answer correlation. Mutual information (MI) and all-confidence measures can be exploited to determine the answer correlation between two QAs, as mentioned in Section 3.3.2. In this experiment, we compare the performance of different measurements of the answer correlation in identifying extended complementary QAs. We derive the answer correlation of two QAs by (1) considering all term pairs of the two QAs' answers; and (2) using the top-5 term pairs with the highest MI or all-confidence values.

TABLE 3 shows the performance of the extended complementation classification model under different measurements of the answer correlation. We focus on the performance of predicting complementary QAs. The best performance is achieved when the answer correlation is measured by the all-confidence metric using the top-5 term pairs with the highest all-confidence values.

TABLE 3. The performance of the extended complementation classification model under different measurements of the answer correlation of QAs

Extended complementary decision tree		Mutual information	MI using top 5 term pairs	All-confidence	*All-conf* using top 5 term pairs
Complementary QAs	Precision	0.679	0.684	0.663	0.75
	Recall	0.663	0.675	0.713	0.675
	F_1-measure	0.671	0.679	0.687	0.711
Non-complementary QAs	Precision	0.630	0.639	0.646	0.671
	Recall	0.648	0.648	0.592	0.746
	F_1-measure	0.639	0.643	0.618	0.707
Average	Precision	0.656	0.663	0.655	0.713
	Recall	0.656	0.662	0.656	0.709
	F_1-measure	0.656	0.662	0.654	0.709

5 CONCLUSIONS

In this paper, we propose an approach for finding complementary QAs. We define two types of complementation, namely, extended complementation and partial complementation, to describe the complementary relationships between QAs. To analyze the complementation probability of extended complementary QAs and partial complementary QAs, we utilize a decision tree classification method that considers the question similarity, answer novelty, and answer correlation between a target QA and candidate QAs. The contribution of this work is

twofold: (a) we provide a method for finding complementary QAs for a target QA; (b) we analyze two types of complementary QAs: extended complementary QAs and partially complementary QAs. Our experiment results demonstrate that, in the analysis of partially complementary QAs, using both question similarity and answer novelty outperforms using question similarity alone; and in the analysis of extended complementary QAs, using the top-5 term pairs with highest all-confidence values to measure the answer correlation yields better classification results than other measures.

ACKNOWLEDGEMENTS

This research was supported by the National Science Council of the Taiwan under the grant NSC 100-2410-H-009-016 and NSC 99-2410-H-009-034-MY3.

REFERENCES

[1] Baeza-Yates, R. and Ribeiro-Neto, B., *Modern information retrieval*. 1999, New York: Addison Wesley Longman.

[2] Bian, J., Liu, Y., Agichtein, E., and Zha, H.: Finding the right facts in the crowd: Factoid question answering over social media, in *Proceeding of the 17th international conference on World Wide Web*. Beijing, China: ACM (2008)

[3] Blooma, M.J., Chua, A.Y.K., and Goh, D.H.-L.: A predictive framework for retrieving the best answer, in *Proceedings of the 2008 ACM symposium on Applied computing*. Fortaleza, Ceara, Brazil: ACM (2008)

[4] Chiang, M.-F., Wang, T.-W., and Peng, W.-C.: Parallelizing random walk with restart for large-scale query recommendation, in *Proceedings of the Workshop on Massive Data Analytics on the Cloud*. Raleigh, North Carolina: ACM (2010)

[5] Collins-Thompson, K., Ogilvie, P., Zhang, Y., and Callan, J.: Information filtering, novelty detection, and named-page finding, in *Proceedings of the 2002 Text Retrieval Conference* (2003)

[6] Jeon, J., Croft, W.B., Lee, J.H., and Park, S.: A framework to predict the quality of answers with non-textual features, in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. Seattle, Washington, USA: ACM (2006)

[7] Kao, W.-C., Liu, D.-R., and Wang, S.-W.: Expert finding in question-answering websites: A novel hybrid approach, in *Proceedings of the 2010 ACM Symposium on Applied Computing*. Sierre, Switzerland: ACM (2010)

[8] Kim, S., Oh, J.S., and Oh, S.: 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. **44**(1), 1-15 (2007)

[9] Liu, X., Croft, W.B., and Koll, M.: Finding experts in community-based question-answering services, in *Proceedings of the 14th ACM international conference on Information and knowledge management*. Bremen, Germany: ACM (2005)

[10] Ma, Q. and Tanaka, K.: Topic-structure-based complementary information retrieval and its application. ACM Transactions on Asian Language Information Processing (TALIP). **4**(4), 475-503 (2005)

[11] Manning, C.D., Raghavan, P., and Schütze, H., *An introduction to information retrieval*. 2008, New York: Cambridge University Press.

[12] Mitra, M., Singhal, A., and Buckley, C.: Improving automatic query expansion, in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. Melbourne, Australia: ACM (1998)
[13] Omiecinski, E.R.: Alternative interest measures for mining associations in databases. IEEE Transactions on Knowledge and Data Engineering. **15**(1), 57-69 (2003)
[14] Suryanto, M.A., Lim, E.P., Sun, A., and Chiang, R.H.L.: Quality-aware collaborative question answering: Methods and evaluation, in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. Barcelona, Spain: ACM (2009)
[15] Wei, C.-P. and Yang, C.-S.: Collaborative filtering-based context-aware document-clustering (cf-cac) technique, in *PACIS 2008 Proceedings* (2008)
[16] Zhang, M., *et al.*: Expansion-based technologies in finding relevant and new information: Thu trec 2002: Novelty track experiments, in *Proceedings of the 11th Text Retrieval Conference* (2002)
[17] Zhang, Y., Callan, J., and Minka, T.: Novelty and redundancy detection in adaptive filtering, in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. Tampere, Finland: ACM (2002)