Recommending QA Documents for Communities of Question-Answering Websites

Duen-Ren Liu*, Chun-Kai Huang, and Yu-Hsuan Chen

Institute of Information Management
National Chiao Tung University, Hsinchu 300, Taiwan
dliu@mail.nctu.edu.tw

Abstract. Question & Answering (Q&A) websites have become an essential knowledge-sharing platform. This platform provides knowledge-community services where users with common interests or expertise can form a knowledge community to collect and share QA documents. However, due to the massive amount of QAs, information overload can become a major problem. Consequently, a recommendation mechanism is needed to recommend QAs for communities of Q&A websites. Existing studies did not investigate the recommendation mechanisms for knowledge collections in communities of Q&A Websites. In this work, we propose a novel recommendation method to recommend related QAs for communities of Q&A websites. The proposed method recommends QAs by considering the community members' reputations, the push scores and collection time of QAs, the complementary relationships between QAs and their relevance to the communities. Experimental results show that the proposed method outperforms other conventional methods, providing a more effective manner to recommend QA documents to knowledge communities.

Keywords: Knowledge Community, Group Recommendation, Knowledge Complementation, Question-Answering Websites, Link Analysis.

1 Introduction

Question & Answering (QA) websites become an important knowledge sharing platform, where question answering knowledge is formed through the mechanism of question posting and answering. The Yahoo! Answers Taiwan website (http://tw.knowledge.yahoo.com/) is a community-driven knowledge website which provides a knowledge community service, so that users with common interests or expertise can form a knowledge community to collect and share question answering knowledge regarding their interests. As the number of posting questions and answers increases rapidly through time, the massive amount of question answering knowledge creates a problem of information overload. Consequently, a recommendation mechanism is needed to recommend QAs to communities of Q&A websites and enhance the effectiveness of knowledge sharing.

* Corresponding author.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 139–147, 2013.

Currently, related research in Question Answering Websites focuses on finding appropriate experts for answering target questions [1]. Previous researches did not investigate the recommendation mechanisms for knowledge collection in question answering websites. Moreover, previous studies on recommender systems focus on recommending items of interest to individual users via collaborative filtering or content-based approaches [2, 3]. Traditional group-based recommendation methods mainly include two kinds of approaches [4]. The first one aggregates interest profiles for each member in a group to form the group's interest profile. The group's interest profile is then used to filter recommended items. The second kind of approach generates a group recommendation list via aggregating the recommendation list of each member derived from personalized recommendations. To the best of our knowledge, there is no study on the recommendation mechanisms for knowledge collections in communities of question answering websites. Traditional recommendation mechanisms have not considered certain factors, such as knowledge complementation and the reputation of the community member in terms of his/her collected QAs.

In this work, a novel group recommendation method is proposed to recommend QA documents to communities of QA websites. The proposed recommendation method generates community profiles from previously collected QAs by considering community members' reputations in collecting and answering QAs, the push scores and collection time of QAs. Moreover, users are usually interested in browsing relevant QAs of related questions to get more complete and complementary information. The proposed approach generates recommendations of QA documents by considering the complementary knowledge of the documents and the relevance degree between the QA document and the community profile. Finally, we use the data collected from Yahoo! Answers Taiwan to conduct our experimental evaluation. Experimental results show that the proposed method outperforms other conventional methods, providing a more effective manner of recommending QAs to communities.

The rest of this paper is organized as follows. Section 2 describes related works. Section 3 describes the proposed methods for recommendation. Section 4 presents experiments and evaluation results. Finally, the conclusion is presented in Section 5.

2 Related Work

Existing group-based recommendation researches were divided into two aspects: the first kind of method aggregates the interest profile of each member in a group to form the group's interest profile. The group's interest profile is then used to filter recommended items. The second kind of approach generates a group recommendation list via aggregating the recommendation list of each member derived from personalized recommendations [4]. However, the second method does not take into account the importance of each member and the interaction between members. The current group-based recommendation systems are widely utilized in different fields, especially in the life and entertainment field. For example, in MusicFx [5], each member can give a rating to the music based on their preference. Group-based recommendations are also used for movies or TV programs such as PolyLens [6] and TV4M [7]. These recommendation systems combine individual preference of movies or programs and then

generate a common recommendation list for the group. In addition, group recommendation is generally used to recommend tourism schedules or scenic spots [8].

The identification of knowledge complementation is unclear due to the definition of complementary knowledge depending on the users themselves. Ma and Tanaka [9] use the concept of topic-structure to measure the complementary degree between two documents. Liu, Chen and Lu [10] define two types of knowledge complementation in a QA website: partial complementation and extended complementation, and propose a method to predict complementation relationships between QA documents by building a classification model based on three measures: question similarity, answer novelty and answer correlation.

3 Proposed QA Recommendation Approach

3.1 Overview of Recommendation for Community Knowledge Collection

The framework of our proposed recommendation method for a knowledge community contains three stages. In the first stage, the content of each QA document is preprocessed into a document profile vector. The term vector of a QA d is denoted as KP_d. The content of each QA is analyzed using the TF-IDF approach [11] to calculate the weight of term i in a profile of QA d, KP_d. In the second stage, the collected QAs are grouped into several topics according to their tags. A community topic profile is derived from a weighted aggregation of document profiles of a topic's collected documents by considering members' reputations in collecting QAs and answering questions, push scores of QAs as well as the time factors of collected QAs. QAs with higher push scores more clearly represent the community's interests. The most recent QAs collected can better reflect the current interest of the community. In the third stage, each target QA is compared with each collected QA of the community to determine a complementary score based on question similarities, answer novelty and answer correlation. Finally, the approach combines the community preference score and complementary score of each target QA to generate a recommendation list.

3.2 Preference Analysis of Knowledge Community

The topic relevance score of target QA q to a topic of community G can be derived by calculating the cosine similarity between q's profile and the community topic profile. A community G's preference score on the target QA q can then be derived as the maximal topic relevance score over all topics of G. The diversity of QAs exists in each topic of a community. Accordingly, we derive a community G's preference score on a target QA q by considering the top-k QAs in each topic collected by G that have highest weighted relevance scores to the target QA q, as shown in Eq. (1). Let $D_{G,q}^{z,topk}$ be the set of top-k QAs in topic z collected by community G that have highest weighted relevance scores to the target QA q. The weighted relevance score of a QA d to the target QA q is derived from their cosine similarity multiplied with the QA d's collection weights, including the collection member's reputation, push score of QA d, and the collection time of QA d. The community G's topic-based preference score on

target QA q in topic z, denoted as $TPR^{z}_{G,q}$, is an aggregation of the weighted relevance scores between the target QA and the QAs in $D^{z,topk}_{G,q}$.

$$TPR^{z}_{G,q} = \frac{\sum_{d \in D^{z,topk}_{G,q}} sim\left(KP_d, KP_q\right) \times MI^{z:d}_{u_c,G} \times WRec^{z}_{d,G} \times WT_{d,G}}{\left|D^{z,topk}_{G,q}\right|}, \quad (1)$$

$$GTPR_{G,q} = Max_z(TPR^{z}_{G,q})$$

where KP_d is the document profile of QA d; $MI^{z:d}_{u_c,G}$ is the importance of member u_c that collected QA d for topic z; $WRec^{z}_{d,G}$ is the push score of d within topic z; $WT_{d,G}$ is the weight of d's collection time. The community's preference score on the target QA q, denoted as $GTPR_{G,q}$, is the maximal topic-based preference score over all topics of community G.

Important community members usually play an important role in collecting QAs. A community member u's importance in topic z, $MI^{z}_{u,G}$ consists of two parts: the reputation of member u for collecting/pushing QA documents in community G, $MCR_{u,G}$. and the reputation of member u for answering questions on topic z on behalf of community G, $MAR^{z}_{u,G}$. The importance of community members is defined, as shown in Eq. (2), which adjusts the relative importance between the member's reputation for collecting QA ($MCR_{u,G}$) and for answering questions ($MAR^{z}_{u,G}$) by parameter α:

$$MI^{z}_{u,G} = \alpha \times MCR_{u,G} + (1-\alpha) MAR^{z}_{u,G} \quad (2)$$

$MCR_{u,G}$ is derived from the link analysis of the knowledge collection and push interactions between community members, while $MAR^{z}_{u,G}$ is derived based on the number of best answers obtained by member u. We adopt a link analysis algorithm, PageRank [12] to calculate members' reputations according to the collect/push relationships among community members. $MAR^{z}_{u,G}$ is a normalized number of best answers obtained by member u on topic z for knowledge community G.

QAs pushed by members with greater importance hold more importance and should generally have higher push scores for the community. In addition, a QA with a higher number of recommendations will be given a higher push score. The push score of a QA d in topic z of community G, $WRec^{z}_{d,G}$ is shown in Eq. (3). $MI^{z}_{u_r,G}$ is the importance of recommender u_r in topic z of community G; $UR^{z}_{d,G}$ is the set of members who recommend the collected QA d in topic z of community G.

$$WRec^{z}_{d,G} = 1 + \left[\frac{\sum_{u_r \in UR_{d,G}} MI^{z}_{u_r,G}}{\left|UR^{z}_{d,G}\right|} \times \left(1 - \frac{1}{\left|UR^{z}_{d,G}\right| + 1}\right) \right] \quad (3)$$

The more recent QA documents collected by a community can better reflect the current interest of the community. The time weight of a QA d collected by community G, $WT_{d,G}$ is adopted from the formula given in [13] to compute time factor.

3.3 Complementary Analysis and Recommendations of Complementary QAs

The complementary relationships among QAs include partial complementation and extended complementation. The information provided in the answer part of a collected QA may be partial and incomplete, so the community may wish to search for related QAs to get complete information. However, the information in some related QAs may be redundant to the collected QA and of no interest to the community. QAs that provide related information that is not redundant are called *partially complementary QA*s of a collected QA. Moreover, some information in the collected QA's answer may not be clear, so the community may wish to search for related QAs that contain extended complementary information. Such QAs are called *extended complementary QAs* of a collected QA. Given two QAs, suppose one is called the Collected QA and the other is called a Target QA. We use the cosine similarity measure to determine the degree of similarity between the question of a collected QA and the question of a target QA. If the question similarity is high, the questions of the two QAs are related, so we analyze their answers to derive each answer's novelty. Let A_d and A_k denote the answers of the Collected QA d and the Target QA q, respectively. We measure the novelty of the two answers, A_d and A_q by Eq. (4), which refers to [10]. 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, this means that the answers contain a lot of common information, so their novelty is low:

$$Nov\left(A_q, A_d\right) = 1 - sim\left(A_q, A_d\right) \tag{4}$$

If the question similarity score is high, this implies that the two questions are related; and if the answers are not redundant, i.e. the answer novelty score is high, partial complementation is inferred. If the question similarity is low, the two questions are different; thus, we have to check to see if any term appears in both the answer of the collected QA and the question of the target QA. If such a term exists, we consider that the target QA may contain some information that can explain the unknown subject in the collected 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 collected QA and the target QA. The answer correlation is measured by the correlation of terms in the answers of the two QAs. Extended complementation generally can be inferred if the answer novelty and answer correlation are high. We use the all-confidence metric [14], which measures the mutual dependence of two variables, to derive the answer correlation, as shown in Eq. (5). The correlation between the two answers, A_d and A_q, denoted by $AC(A_d, A_q)$, is derived by summing the all-confidence (x,y) scores for $x \in S_d^A$ and $y \in S_q^A$. Note, that S_d^A / S_q^A is the term set of A_d / A_q:

$$AC\left(A_q, A_d\right) = \sum_{x \in A_q} \sum_{y \in A_d} \frac{P(x \wedge y)}{MAX\left(P(x), P(y)\right)} \tag{5}$$

where *x/y* is the term contained in the answer for document *q/d*; *P(x)* is the probability of documents containing term *x* and $P(x \wedge y)$ is the probability of documents containing both term *x* and term *y*. The dependence of two terms (probability) is measured by the number of documents which contain the two terms returned by the Google search engine. We use a decision tree classification approach to build a classification model and predict the complementary relationships among QAs based on three input variables: question similarity, answer novelty and answer correlation. Specifically, we use Weka's Classification and Regression Tree (CART) model [15] to build a classification model. In the prediction of a target QA, the decision process reaches a leaf node of the classification tree based on question similarities, answer novelties and answer correlations between two QAs. The complementary score of target QA, *q,* to the collected QA, *d*, $CPS_{q,d}$, is the partial or extended probability which can be calculated as the ratio of the number of training cases in the leaf node with a positive label to the total number of training cases in the leaf node.

A target QA may be complementary to very few QAs collected in a community. Therefore, in order to enhance the effect of complementary QA recommendations, we derive the complementary score of target QA *q* to a topic *z* of a community *G*, $CPS^{z}_{G,q}$ by aggregating the complementary scores of a target QA to the QAs collected in *z*, as shown in Eq. (6):

$$CPS^{z}_{G,q} = f_{d \in D^{z}_{G}}\left(CPS_{q,d}\right) \tag{6}$$

where D^{z}_{G} is the set of QAs in topic *z* collected by community *G*; $CPS_{q,d}$ is the complementary score of target QA *q* to collected QA *d*; and *f*() is an aggregation function, such as the average, max or sum of the complementary scores of target QA to the QAs collected in *z*, that can be used to determine the complementary score of target QA to a topic. In the experiment, the max function is applied for measuring the complementary score. Once the complementary score of target QA *q* to a topic *z* $CPS^{z}_{G,q}$ is derived, we consider $CPS^{z}_{G,q}$ to enhance the effect of recommending complementary QAs in deriving community *G*'s preference score on target QA *q*, $GPRC^{topic}_{G,q}$ by the topic-based complementary approach, as shown in Eq. (7):

$$TPRC^{z,topic}_{G,q} = \frac{\sum_{d \in D^{z,topk}_{G,q}} sim\left(KP_d, KP_q\right) \times MI^{z:d}_{u_c,G} \times WRec^{z}_{d,G} \times WT_{d,G} \times \left(1 + CPS^{z}_{G,q}\right)}{\left|D^{z,topk}_{G,q}\right|}, \tag{7}$$

$$GTPRC^{topic}_{G,q} = Max_{z}\left(TPRC^{z,topic}_{G,q}\right)$$

where $CPS^{z}_{G,q}$ is the complementary score of target QA *q* to a topic *z* collected by community *G*. Community *G*'s preference score on target QA *q* of topic *z* is obtained by multiplying the two factors, including the weighted relevance scores of target QA *q* to top-*k* QAs in topic *z* and the complementary score of target QA to topic *z*. Finally, community *G*'s preference score on the target QA document *q*, $GTPRC^{topic}_{G,q}$, is the maximal topic-based preference score over all topics of community *G*. We enhance the effect of the complementary QA recommendation by using the *Max* function to

derive complementary scores of target QA to topics. QAs with high preference scores are used to compile a recommendation list, from which the top-*N* QAs are chosen and recommended to the target user.

4 Experimental Evaluations

4.1 Experiment Design

We evaluate the performance of the proposed approach by using the QA documents collected in knowledge communities at Yahoo! Answers Taiwan. We choose 15 knowledge communities from three domains: computer, medicine and finance. The F1 performance metric [3, 16] is used to evaluate the performance of the proposed approach. F1-measure is the harmonic means of precision and recall. We divide the data set into training data and testing data. The data from each community is separated into two parts, 80% for training data and 20% for testing data. Our proposed methods are compared with the traditional content-based group recommendation method. The content-based group recommendation method consolidates individual profiles to generate group profiles, which in turn are used to filter out items of recommendation. The top-*N* QAs are recommended to the target user.

The traditional ***GP-CB*** method mainly considers the content similarity between the recommended document and the community profile in order to recommend related QAs to the community without considering community topics and QA collection weights. The ***GPT*** method recommends QAs to the community based on the relevance of target QA to the community-topic profiles without considering QA collection weights. A community *G*'s preference score on a target QA *q* is derived by considering the top-*k* relevant QAs in each topic. The ***GTPR*** method uses QA collection weights to derive weighted relevance scores and derive a community *G*'s preference score from top-*k* QAs in each topic. The topic-based complementary method (*GTPRC-T*) recommends QAs to the community not only considering the relevance of QAs and QA collection weights, but also the complementary scores of QAs.

4.2 Experimental Results

A community *G*'s preference is derived by considering the top-*k* QAs rather than all QAs of each topic. Our experiment result shows the recommendation quality is the best when *k* equals 10. Based on the result, we choose top-10 QAs of each topic to derive community preferences for *GPT* method and our proposed methods. The *GTRPC-T* performs better than the *GTPR*. The results imply that considering complementary QAs helps to improve the recommendation quality. Fig. 1 shows the performance comparison (F1 measures) among various recommendation methods. The *GPT* recommends QAs based on top-*k* (top-10) QAs in topics that are most relevant to target QA without considering QA collection weights. The *GTPR* uses QA collection weights to derive weighted relevance scores and derive a community *G*'s preference score from top-*k* QAs in each topic. The *GP-CB* method does not consider the topics and the three QA collection factors. The result shows that the recommendation quality of *GPT* is better than that of the traditional *GP-CB* method. The *GTPR* method

performs better than the *GPT* and the *GP-CB* method. Considering the topic profiles and the QA collection factors can achieve better recommendation performance than the traditional content-based group profiling method. Moreover, the results show that the *GTPRC-T* performs the best among all the methods. The recommendation quality is improved when we consider the complementary scores of the target QAs. In summary, our proposed approach is effective in recommending complementary QA documents to knowledge communities.

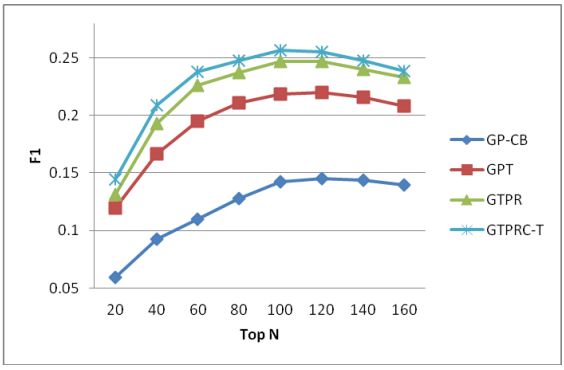

Fig. 1. F1 measures of various recommendation methods

5 Conclusion

In this research, a novel recommendation approach is proposed on recommending relevant and complementary QA documents to knowledge communities of Q&A websites. Recommending complementary QAs is important to increase the effectiveness of knowledge collections. The novel ideas of our proposed approach are as follows: 1) It generates community topic profiles by considering QA collection factors such as community members' reputations in collecting and answering QAs, push scores of QAs and the collection time of QAs from the historically collected QA documents on specific topics. 2) It predicts the complementary scores of QAs based on question similarity, answer novelty and answer correlation. 3) It proposes a QA-based complementary approach and topic-based complementary approach to recommend complementary QA documents. Experimental results show that consider partial or extended complementary QAs help improve the recommendation quality. Moreover, our proposed approach, that considers community topic profiles with QA collection factors and complementary scores of QAs, performs better than traditional recommendation methods. Our proposed approach is effective in recommending complementary QA documents to knowledge communities.

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

References

1. Liu, D.-R., Chen, Y.-H., Kao, W.-C., Wang, H.-W.: Integrating Expert Profile, Reputation and Link Analysis for Expert Finding in Question-Answering Websites. Information Processing and Management 49, 312–329 (2013)
2. Balabanovic, M., Shoham, Y.: Fab: content-based, collaborative recommendation. Communication of the ACM 40(3), 66–72 (1997)
3. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Analysis of recommendation algorithms for e-commerce. In: Proceedings of ACM Conference on Electronic Commerce, pp. 158–167 (2000)
4. Kim, J.K., Kim, H.K., Oh, H.Y., Ryu, Y.U.: A group recommendation system for online communities. International Journal of Information Management 30, 212–219 (2010)
5. McCarthy, J.F., Anagnost, T.D.: MusicFX: an arbiter of group preferences for computer supported collaborative workouts. In: Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work (CSCW), pp. 363–372. ACM, Seattle (1998)
6. O'Connor, M., Cosley, D., Konstan, J.A., Riedl, J.: PolyLens: a recommender system for groups of users. In: Proceedings of the Seventh Conference on European Conference on Computer Supported Cooperative Work, pp. 199–218. Kluwer Academic Publishers, Bonn (2001)
7. Yu, Z., Zhou, X., Hao, Y., Gu, J.: TV Program Recommendation for Multiple Viewers Based on user Profile Merging. User Modeling and User-Adapted Interaction 16, 63–82 (2006)
8. Jameson, A.: More than the sum of its members: challenges for group recommender systems. In: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 48–54. ACM, Gallipoli (2004)
9. Ma, Q., Tanaka, K.: Topic-structure-based complementary information retrieval and its application. ACM Transactions on Asian Language Information Processing 4, 475–503 (2005)
10. Liu, D.-R., Chen, Y.-H., Lu, P.-J.: Complementary QA-Network Analysis for Q&A Retrieval in Question-Answering Websites (2012) (submitted manuscript)
11. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management 24, 513–523 (1988)
12. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)
13. Zhang, J., Ackerman, M., Adamic, L., Nam, K.: QuME: a mechanism to support expertise finding in online help-seeking communities. In: Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology, pp. 111–114. ACM (2007)
14. Omiecinski, E.R.: Alternative interest measures for mining associations in databases. IEEE Transactions on Knowledge and Data Engineering 15, 57–69 (2003)
15. Weka: Data Mining Software at URL: `http://www.cs.waikato.ac.nz/ml/weka/`
16. Rijsbergen, C.J.V.: Information retrieval. Butterworth-Heinemann, London (1979)