Literature Review: Exploring user behavior patterns based on Zhihu dataset - a Chinese online Q&A community

Andi Liao

1. Introduction

Online Q&A communities, which enable users to post questions and obtain answers, are becoming more and more popular these days. Typical examples include Quora, Yahoo! Answer and Stack Overflow. In my research project, I choose Zhihu, which is the largest social based questions and answers online community in China with 100 million registered individual users, as the target community.

This research will mainly concentrate on user behaviors in online Q&A communities. The next two sections will review researches about user interaction, especially user expertise inside online Q&A communities, and the final session will discuss how these work lead to my research project.

2. User interactions

Researchers study user interaction behaviors from different perspectives, and here it is discussed from the angle of user participation and user role, which serve as the theoretical foundation of my research project.

2.1 User participation

There are many factors contributing to participation behavior in online Q&A communities, including personal features, network features and mental motivation. Focusing on mental motivation, researchers collect user activities data from Zhihu to build a hierarchical regression model to show that identity-based trust, positive feedback, social exposure, norms of reciprocity and self-presentation information all facilitate continuous knowledge contribution behavior inside this community(Guan, Wang, Jin, & Song, 2018). In other words, social capital theory, social exchange theory, social cognitive theory and communication theory of identity all play a part in explaining knowledge-sharing behavior.

Researchers also believe that user participation is the most important factors contributing to the success of online Q&A communities. A group of researchers choose Yahoo! Answers, a social Q&A site, and Google Answers, a paid expert Q&A site, to compare, and they figure out that the number of contributors and consumers are well balanced in Yahoo! Answers but not in Google Answers(Shah, Oh, & Oh, 2008). By encouraging user participation, Yahoo! Answers develops as a responsive community and attracts users to repeatedly raise questions or provide answers, while Google Answers has a high percentage of one-time consumers because of restricting user participation.

2.2 User role

Basically, user roles in online Q&A communities can be divided into: contributors and consumers(Shah et al., 2008); administrators, content contributors(questioners/answer people/discussion people and technical editors) and marginal roles(fans/lurkers)(Wang & Zhang, 2016).

A recent research identify Zhihu users as starters, answerers, technical editors and followers in health-related topics(Wang & Zhang, 2016). Researchers discover that men are more likely to be answerers and technical editors, while women prefer to be starters and followers. Most users are in IT industry, followed by public health and social work. In terms of content contribution, technical editors and starters post more questions, and answerers answer post more answers. As for social connection, followers follow more topics and users, and answerers have more followers. Speaking of popularity and recognition, technical editors have more pageviews, votes and likes. And these activity measurements are not isolated - they are all significantly and positively related to each other.

Given the review above, it can be seen that user interactions are extremely diverse in online Q&A communities. In particular, researchers are fascinated by expert users, the representatives of active participators and enthusiastic contributors. The following section will summarize papers relevant to user expertise.

3. User expertise

Currently, researches about online Q&A communities expertise can be divided into two categories: one is user expertise estimation utilizing interaction graph analysis and interest modelling; the other one is quality prediction, including questions quality and answers quality based on historical data(Xiao, Zhao, Wang, & Xiao, 2014).

As my research project emphasize on user behaviors, only the first perspective, the user expertise using user behavior data will be taken into consideration.

3.1 Detecting user expertise

The simplest way to measure user expertise to use existing metrics extracted from online Q&A communities. (Shah et al., 2008) utilize the points and levels system embedded in Yahoo! Answers to cluster users, where the level of a particular user is largely determined by the contribution of answers. Answers from users in higher levels are routinely selected as best answers, and it is a clear reflection of expertise knowledge in a specific area.

There are more complicated methods, including link analysis, clustering and unsupervised learning techniques applied in detecting expert users.

It is a common-belief that authoritative users contribute to the production of high-quality contents. Therefore, researchers(Bouguessa, Dumoulin, & Wang, 2008) model Indegree, a measurement of user authority considering the sum of weights of

edges that point to the current node, as a combination of two gamma components, and then use this technique to automatically discriminate expert users and non-expert users.

Researchers(Furtado, Oliveira, & Andrade, 2014) also use Wald algorithm and k-means cluster method to classify contributors in Super User community into four categories: no marked skill contributors, unskilled answerers, experts and activists. The key measurement here is the combination of motivation metrics and ability metrics based on the quantity and quality of user contributions.

To detect expert users on Quora, researchers collect profiles of users who follow a certain set of topics, and label them as either expert or non-expert(Patil & Lee, 2016). Researchers find out that expert users and non-experts behave differently: expert users have more followers, make more edits, generally post longer answers, post more questions and prefer lexical words compared to non-expert users. Then researchers build J48, SVM and Random Forest classifier using best features selected from the activity features, quality of answer features and linguistic features from previous step. Random Forest classifier outperforms other models by achieving 95.94% accuracy in the general-topic dataset, and even better in specific-topic dataset. Additionally, researchers categorize experts into three subcategories, fluctuating experts, stable experts and idle experts according to their temporal activities features.

3.2 Predicting user expertise

Apart from identifying expert users, researchers also try to predict expert users using different approaches, such as link analysis and machine learning methods.

By comparing the HITS algorithm, another link analysis method, together with Degree algorithm to predict experts in Yahoo! Answers, researchers(Jurczyk & Agichtein, 2007) discover that the HITS model is more robust, and successfully predict experts in a specific domain as well as the percentage of best answers generated by expert users in the next 5 months.

Researchers(Xiao et al., 2014) also predict the performance of new-coming Zhihu users using PageRank algorithms and related social media features. Researchers choose new Zhihu users who logged in using the social media account - Sina Weibo, and evaluate prestige and relevance from previous Weibo's activities. Researchers test combinations of PageRank algorithms(unbiased/biased) and Weibo's features(prestige/relevance) to predict best answers and top experts, and they discover that using biased PageRank algorithms and prestige feature can greatly improve the prediction performance, especially when the history window is small.

To find potential experts in TurboTax Live Community, researchers(Pal, Farzan, Konstan, & Kraut, 2011) utilize SVM and Decision Tree over motivation and ability features of users. SVM can reach the precision 0.89 when combing two features, covering the number of answers/votes/best answers, frequency of login, average time gap between answers and usage of pronoun.

This part reviews mainstream methods of detecting and predicting user expertise in online Q&A communities, which builds the methodology foundation of my research.

4. Conclusion

User behaviors have their own patterns in online Q&A communities. Mental motivation elicits user participation in knowledge sharing, which leads to the success of online open Q&A communities(Guan et al., 2018, @shah2008exploring), and users can be divided into different roles manually according to behavior data(Wang & Zhang, 2016).Together, these work lay the theoretical basis of my research project - users have internal motivation to share knowledge to as contributors.

As for methods, k-means(Furtado et al., 2014) and Random Forest classifier(Patil & Lee, 2016) will be ideal given their outstanding performance in identifying experts.

Previous research use profile information, account activity, linguistic features in questions and answers, content in social media accounts and other type of data to build their models. The highlight as well as the challenge of my project is that only profile information and activity summary data will be available.

In my project, I will first explore the distribution of Zhihu users in the sample given different dimensions, and then use k-means method to cluster users after extracting features from profile data. Random forest classifiers will also be utilized to detect expert users within the dataset, and if time permitted, I will also try to find early experts using prediction.

5. Reference

Bouguessa, M., Dumoulin, B., & Wang, S. (2008). Identifying authoritative actors in question-answering forums: The case of yahoo! Answers. In *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining* (pp. 866–874). ACM.

Furtado, A., Oliveira, N., & Andrade, N. (2014). A case study of contributor behavior in q&A site and tags: The importance of prominent profiles in community productivity. *Journal of the Brazilian Computer Society* , *20* (1), 5.

Guan, T., Wang, L., Jin, J., & Song, X. (2018). Knowledge contribution behavior in online q&A communities: An empirical investigation. *Computers in Human Behavior* , *81* , 137–147. Elsevier.

Jurczyk, P., & Agichtein, E. (2007). Discovering authorities in question answer communities by using link analysis. In *Proceedings of the sixteenth acm conference on conference on information and knowledge management* (pp. 919–922). ACM.

Pal, A., Farzan, R., Konstan, J. A., & Kraut, R. E. (2011). Early detection of potential experts in question answering communities. In *International conference on user modeling, adaptation, and personalization* (pp. 231–242). Springer.

Patil, S., & Lee, K. (2016). Detecting experts on quora: By their activity, quality of answers, linguistic characteristics and temporal behaviors. *Social network analysis and mining* , *6* (1), 5. Springer.

Shah, C., Oh, J. S., & Oh, S. (2008). Exploring characteristics and effects of user participation in online social q&A sites. *First Monday* , *13* (9).

Wang, Z., & Zhang, P. (2016). Examining user roles in social q&A: The case of health topics in zhihu. Com. *Proceedings of the Association for Information Science and Technology* , *53* (1), 1–6. Wiley Online Library.

Xiao, Y., Zhao, W. X., Wang, K., & Xiao, Z. (2014). Knowledge sharing via social login: Exploiting microblogging service for warming up social question answering websites. In *Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 656–666).