Revisiting the Benefits of Duplicate Questions: Evidence from Knowledge Evolution on Stack Overflow

Zhang, Yiwei Nanjing University, China | zhangyiweinju@163.com
Jiang, Na BNU-HKBU United International College, China | najiang@uic.edu.cn
Liu, Xiaohui University of Shanghai for Science and Technology, China | xiaohuiliu23@usst.edu.cn
Zhang, Qi Nanjing University, China | qi.zhang@smail.nju.edu.cn
Deng, Sanhong Nanjing University, China | sanhong@nju.edu.cn

ABSTRACT

Stack Overflow (SO) represents one of the most vibrant Question Answering Communities (QACs), providing a crucial platform for developers to pose and respond to questions. SO preserved duplicate questions due to their potential for furnishing additional insights or suggestions. In this paper, we delve into the study of duplicates within SO, with the objective of unraveling their positive value, particularly through the lens of knowledge networks and the evolution. We propose a categorization of knowledge evolution within QACs into two key dimensions: depth and breadth. Our exploration reveals that duplicate questions play a constructive role in fostering both the depth and breadth of knowledge evolution. This finding illuminates the underestimated value of duplicate questions, underlining their significance for the ongoing expansion of knowledge within QACs.

KEYWORDS

Duplication; Question Answering Communities; Knowledge Evolution

INTRODUCTION

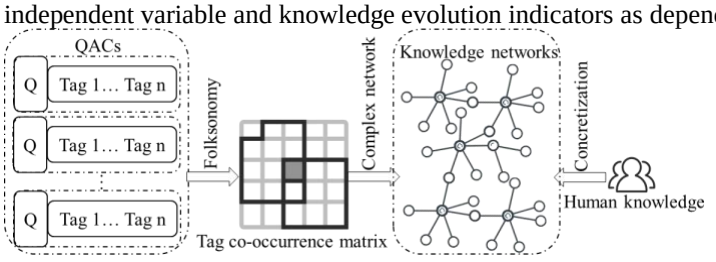
Question Answering Communities (QACs) have emerged as significant knowledge generators, aiding users in fulfilling their information requirements (Ahmad et al., 2018). Specialized QACs, such as Stack Overflow, serve as repositories for human knowledge. Questions on these platforms reflect the expertise and interests of professionals in various domains, fostering the generation and evolution of specialized knowledge (Liu et al., 2019). Typically, users seek pre-existing answers before posing new questions, a process that inevitably leads to the posting of duplicate questions. Despite being labeled as such, these duplicate questions are preserved for their potential informational value.

Prior research primarily centered on enhancing detection algorithms for duplicate questions (Kale et al. 2022), typically viewing duplicates negatively (Kumar et al. 2023) and overlooking their potential benefits. Despite this trend, studies by Mathias Ellmann (Ellmann 2019) and others have begun to uncover the potential value of duplicates on Stack Overflow. However, these works have not accurately pinpointed where this value lies. Despite these findings, research exploring the role of duplicate questions as key knowledge components in QACs, particularly in terms of their contribution to knowledge evolution, remains scant.

In this paper, we investigate the positive impact of duplicate questions in QACs, we conceptualize knowledge evolution into two dimensions: depth and breadth. Further, we quantified measure the positive influence of duplicate questions on the depth and breadth of knowledge evolution. Using the Stack Overflow Torrent dataset, we address the following research questions: 1) How can the evolution of knowledge networks be understood within CQA platforms? 2) What impact do duplicate questions have on the depth and breadth of knowledge network evolution?

METHODOLOGY

This study scrutinizes the iOS domain knowledge on Stack Overflow from 2008 to 2021, generating a dataset of 114,563 question entries. Employing an enhanced distil-BERT pretraining model (Sanh et al. 2020), we facilitate a data-driven approach to identifying duplicates. Each question's structure, including the question and its tags, underpins the construction of knowledge networks via complex networks and a Folksonomy-based knowledge organization model (Peters and Stock 2007). The adjacency matrices were built on tag co-occurrence networks (Feicheng and Yating 2014), leading to a dynamic exploration of these networks via time-series analysis. Knowledge evolution measured by depth and breadth. Influence of duplicates on knowledge evolution investigated via regression analysis with duplication score as independent variable and knowledge evolution indicators as dependent variables.



86th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – 31, 2023 |London, United Kingdom.

Figure 1. Knowledge Network Construction

FINDINGS

Knowledge Network Evolution: Depth and Breadth Explained

Knowledge evolution signifies the spatiotemporal dynamics of knowledge transmission and growth (Pontis and Blandford 2015). This study focuses on the network topology (Momennejad 2021), with depth reflecting knowledge density and intricacy, denoting internal interactions and transactions during evolution. Conversely, breadth represents diversity, illustrating the expandability of network structure, continuity, and capacity for expansion.

Specifically, the depth of the knowledge network is characterized by measures such as the clustering coefficient, closeness centrality, eigenvector centrality, degree, and the extent of the largest connected subgraph. Alternatively, the breadth of the knowledge network is represented through the average shortest path length, assortativity coefficient, betweenness centrality, and the number of major branches (Zou et al. 2019).

The Role of Duplicate Questions in Knowledge Network Evolution

This study scrutinizes the role of duplicate questions in the evolution of iOS domain knowledge network. Our findings (Table 1) establish that duplicate questions significantly enhance the depth and breadth of this evolution.

Categories	Variables	Standardized Coefficients	T	R^2	Adjusted R^2	F
Knowledge Depth	Clustering Coefficient	1.8999***	16.7170	0.6462	0.6439	279.4576
	Closeness Centrality	0.2378*	1.6679	0.0179	0.0114	2.7818
	Eigenvector Centrality	18.4443***	8.6781	0.3299	0.3255	75.3088
	Degree	-4.9459***	-7.4647	0.2670	0.2622	55.7213
	Largest Connected Subgraph	-31.6719***	-8.9664	0.3445	0.3402	80.3971
Knowledge Breadth	Average Shortest Path Length	-1.5369***	-5.2994	0.1551	0.1496	28.0841
	Assortativity Coefficient	-2.9332***	-7.3955	0.2633	0.2585	54.6934
	Betweenness Centrality	28.0378***	8.3149	0.3112	0.3067	69.1384
	Number of Major Branches	-25.1586***	-8.4677	0.3191	0.3146	71.7026

*Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$*

Table 1. Impact of Duplicate Questions on the Knowledge Network Evolution

Influence on knowledge evolution depth: significant positive findings associated with the clustering coefficient suggest that duplicate questions augment knowledge depth evolution by fostering inter-nodal connections. Further, they positively impact closeness and eigenvector centrality, thereby reducing network distance and promoting ties with significant questions. This implies **duplicate questions may enhance knowledge depth by linking new and existing knowledge**. However, an excessive proliferation of duplicates negatively affects the degree metric and largest connected subgraph, implying the trade-off between connectivity enhancement and potential information redundancy.

Influence on knowledge evolution breadth: duplicate questions negatively correlate with average shortest path length and assortativity coefficient (Mulders et al. 2020), suggesting an expansion of the network diameter and increased connection of dissimilar degree nodes. This indicates that high-degree nodes, such as popular questions, are more inclined to connect with low-degree nodes, thus potentially expanding the knowledge spectrum. **Duplicates** also exhibit a positive impact on betweenness centrality, **serving as 'bridges' to facilitate knowledge linkage and dissemination.** Interestingly, the negative correlation with the number of major branches reveals the role of duplicates in consolidating sub-communities, thereby fortifying knowledge evolution depth and curtailing fragmentation.

In conclusion, duplicate questions play a dual role in knowledge network evolution. They enrich both the depth, by augmenting the density and intricacy of knowledge, and the breadth, by reinforcing the network connectivity.

CONCLUSION

This paper provides a unique examination of duplicate questions from a knowledge evolution perspective. We envision knowledge evolution as the temporal progression of a Folksonomy-embedded complex network, concentrating primarily on the role of duplicate questions in fostering knowledge evolution within these QACs.

Interestingly, we discern that duplicate questions have a positive effect on both the depth and breadth of knowledge evolution within these QACs. To quantify the specific role of duplicate questions in knowledge evolution depth and breadth, regression analysis was utilized. The outcomes corroborate the positive value of duplicate questions in knowledge evolution, thereby affirming their significance within knowledge communities.

This research has profound implications for the realm of duplicate questions within question-and-answer communities. It intends to deepen our understanding and measurement of the value inherent in duplicate questions. While extensive discussions and research focus on the detection, identification, or avoidance of duplicate questions, the potentially

beneficial value of duplicates has been largely neglected. This study thus illuminates the oft-overlooked merits of duplicate questions in the process of knowledge evolution.

REFERENCES

Ahmad, A., Feng, C., Ge, S., & Yousif, A. (2018). A survey on mining stack overflow: Question and answering (Q&A) community. *Data Technologies and Applications*, *52*(2), 190–247. https://doi.org/10.1108/DTA-07-2017-0054

Ellmann, M. (2019). Same-Same But Different: On Understanding Duplicates in Stack Overflow. *Informatik Spektrum*, *42*(4), 266–286. https://doi.org/10.1007/s00287-019-01185-y

Feicheng, M., & Yating, L. (2014). Utilising social network analysis to study the characteristics and functions of the co-occurrence network of online tags. *Online Information Review*, *38*(2), 232–247. https://doi.org/10.1108/OIR-11-2012-0124

Kale, M., Rayasam, A., Parik, R., & Dheram, P. (2022). Mining Duplicate Questions of Stack Overflow. In *ArXiv e-prints*. https://doi.org/10.48550/arXiv.2210.01637

Kumar, A., Ghadiyali, D., Chimalakonda, S., & Venigalla, A. S. M. (2023). SOCluster—Towards Answering Unanswered Questions on Stack Overflow via Answered Questions. *Proceedings of the 16th Innovations in Software Engineering Conference*, 1–5. https://doi.org/10.1145/3578527.3578544

Liu, X., Li, Y., Liu, F., Cai, Z., & Lim, E. T. K. (2019). Reinventing the Wheel: Explaining Question Duplication in Question Answering Communities. *Proceedings of the 40th International Conference on Information Systems (ICIS)*, 2970. https://research.cbs.dk/en/publications/reinventing-the-wheel-explaining-question-duplication-in-question

Momennejad, I. (2021). Collective minds: Social network topology shapes collective cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1843), 20200315. https://doi.org/10.1098/rstb.2020.0315

Mulders, D., de Bodt, C., Bjelland, J., Pentland, A., Verleysen, M., & de Montjoye, Y.-A. (2020). Inference of node attributes from social network assortativity. *Neural Computing and Applications*, *32*(24), 18023–18043. https://doi.org/10.1007/s00521-018-03967-z

Peters, I., & Stock, W. G. (2007). Folksonomy and information retrieval. *Proceedings of the American Society for Information Science and Technology*, *44*(1), 1–28. https://doi.org/10.1002/meet.1450440226

Pontis, S., & Blandford, A. (2015). Understanding "influence:" an exploratory study of academics' processes of knowledge construction through iterative and interactive information seeking. *Journal of the Association for Information Science and Technology*, *66*(8), 1576–1593. https://doi.org/10.1002/asi.23277

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (arXiv:1910.01108). arXiv. https://doi.org/10.48550/arXiv.1910.01108

Zou, Y., Donner, R. V., Marwan, N., Donges, J. F., & Kurths, J. (2019). Complex network approaches to nonlinear time series analysis. *Physics Reports*, *787*, 1–97. https://doi.org/10.1016/j.physrep.2018.10.005