Community by Design: Prioritizing the Factors that Drive Knowledge Use in Online Question & Answers Platforms

Babajide Osatuyi
The Pennsylvania State University
bosatuyi@psu.edu

Katia Passerini
St. John's University
passerik@stjohns.edu

Abstract

The question of how knowledge assets are utilized in the context of online communities is the primary impetus of this research. Using a multilevel approach, this paper investigates factors that influence the use of knowledge in an online question and answer platform (OQA). It focuses on three levels including informational, individual, and community, and reviews interactions across each level. The study tests the multilevel model with data from StackOverflow.com, a renowned online community for programmers to exchange knowledge assets, especially questions and answers about coding issues. Traditional hierarchical regression analysis proved insufficient to explicate the complexity associated with human decision-making processes with respect to asset utilization. However, a machine learning technique with a Chi-square automatic interaction detection algorithm provided a richer understanding of the relative importance of factors and their thresholds for influencing knowledge asset use.

1. Introduction

Researchers have shown that understanding knowledge flow between agents requires analyzing factors across multiple levels [1]. However, much of the current knowledge management research generally analyzes factors on one level at a time. Furthermore, while the focus of extant knowledge management literature is on how knowledge is created, shared and stored, limited research addresses the issue of how it used because the focus has been on identifying characteristics of knowledge or systems for knowledge management that facilitate creation and exchange, and actual use of the knowledge exchanged has taken a second seat. In this research, an answer in the online Q&A platform cannot be chosen without having been tested for its accuracy and ability to solve a specific coding problem. Because of this unique context that requires use and application before rating, this study is in a unique position to beyond the antecedent of knowledge exchange. Furthermore, this study concurrently analyses factors across three levels, while prior research focused on informational and individual level (embedding relationship and community metrics within the interaction among knowledge seekers and contributors) [1]. We analyze factors across three distinct levels (informational, individual and community).

In the online question and answer (OQA) community context, few researchers have examined how participants use knowledge assets available in the community platform. The main focus has been on how, when, and why users share or contribute information [2]. Studies focused on the psychological motivation for sharing knowledge assets concluded that users' membership levels have a considerable influence on their knowledge exchange behavior [2-5].

Although these studies provide useful insights into the factors that motivate the knowledge exchange, it is essential to understand the factors that influence the use of knowledge. That is, the actual use of the knowledge contributed [6].

The following sections illustrate the various aspects of this research. Section 2 presents the background for this study and its research questions. Section 3 explains the methodology used to gather the data and conduct analyses. Section 4 describes the results followed by a discussion of the results (Section 5), and conclusions and future work in the final sections.

2. Background

2.1. Multilevel Framework of Knowledge Utilization

The multilevel perspective on knowledge utilization draws upon the factors that influence knowledge transfer postulated by Szulanski's [7] and depicted in

URI: https://hdl.handle.net/10125/59989
ISBN: 978-0-9981331-2-6

Figure 1. The framework posits that the extent to which knowledge assets will be exchanged or used depends on factors across the three levels shown in the framework (informational, individual and community). This approach acknowledges that to explore the complexity associated with human decision-making processes, a multilevel approach is needed to understand how people make the decision to use knowledge assets. Persuasion theories [8] also follow this line of thinking by including factors across the three levels to explicate how people make knowledge exchange decisions. Persuasion theories could be useful framework in future research studies as they focus on driving decision making processes that lead to action, similarly to the action-oriented focus of knowledge management.

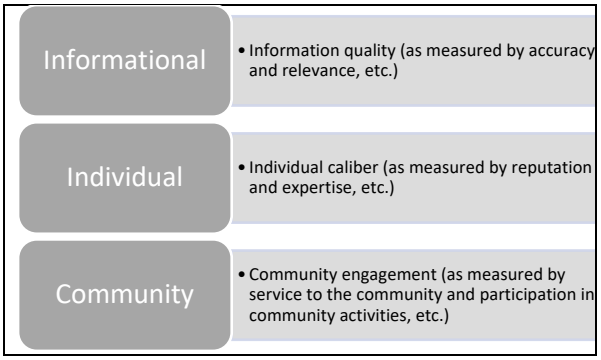

Figure 1. Multilevel Perspective of Knowledge Utilization

At the informational level, the decision to exchange and use knowledge assets depends on characteristics of the knowledge assets that signal quality indicators such as its accuracy and relevance. Studies have shown that the quality of information is an important determinant of knowledge exchange [9]. In the OQA community context, users evaluate the quality of a piece of information, after using it (i.e. the coding snippets) by assigning a quality rating or score.

At the individual level, knowledge asset utilization is based on who contributed the knowledge asset. In other words, the focus here is on the reputation and expertise of who contributes information in the OQA community. This idea echoes persuasion theories [8, 10] that stipulate that characteristics of the information provider weighs highly on determining the extent to which information is accepted by members of a community. Knowledge assets contributed by OQA members with high reputation scores have a higher likelihood of being accepted compared to assets contributed by members with low reputation scores [2].

Finally, factors at the community level focus on contributors' efforts to engage in activities to make the community a better place [11]. Furthermore, research shows that participation in activities that are outside the *modus operandi* of most users e.g., curating posts to improve the quality of the posts or suggesting additional sources of information, helps community engagement [12]. In the OQA context, engagement in such community service makes users more informed and better equipped to solve questions posed on the platform.

Although the described multilevel framework provides three levels of factors that influence knowledge exchange (informational, individual and community), it leaves out the order of importance of those factors as they may change from one context to the next. This study seeks to examine the relative importance of factors across the levels.

Given that most earlier studies demonstrate that information quality contributes the highest to explaining the variability in exchange behaviors [1, 2], such factor is expected to have the highest influence in the determination of the extent to which knowledge will be utilized. Engagement in activities in the OQA community, including editing posts to improve readability and comprehension or revising questions to make it easier for potential contributors to better answer the questions, is expected to be the next most influential determinant of knowledge utilization because such activities increase the information quality. While individual reputation is important, a good answer could outperform the expertise of the individual contributor if his/her answer is not as accurate or relevant to the question at hand (in other words, quality outperforms individual reputation).

The above considerations lead us to the following research questions:

RQ1: Do informational, individual and community factors positively influence knowledge use? In what sequence?

3. Method

3.1. Data Collection

The data was gathered using a Web crawler designed by the first author from StackOverflow.com (SO) for a total of twelve weeks. SO was used for several reasons. First, it is a community that is extensively text-based, enabling the tracking of questions and answers. Second, it organizes solutions by the questions making it easy to mine the data for analysis. Third, the solution that is selected by the information seeker is marked to separate it from the other solutions provided by other contributors. Fourth, SO keeps track of each user's engagement in community services such as editing incorrect questions and answers. Therefore, all the

elements in Figure 1 are captured using StackOverflow.com data.

A total of 1191 answers provided by 889 contributors were selected from a random list of resolved questions on SO. These questions were asked by a total of 122 programmers experiencing some difficulty with their javascript coding project. SO provides unique identifiers for users and their contributions on the site, making it useful for tracking users' activities on the site.

3.2. Research Variables

The variables of interest in this study include answer selected as the best answer, the reputation of the contributor, posts edited by the contributor and finally, quality of the contribution. Table 1 depicts how each of the research variables are measured on StackOverflow.com and operationalized in this study. It is to be noted that earlier research focused on each component separately (i.e. [1] does not include community) whereas this research contributes this factor as an independent variable. To meaningfully operationalize the values on SO for analysis, the logarithm (log) transformation was applied to manage the effect of large numbers that are typical in online communities.

Table 1. Research variables and operationalization.

Level	Variable	Description	How Measured on StackOverflow.com	Operationalization in this study
	Answer	Answer selected by the seeker	Selected answer from a list of answers provided by contributors	0 for not selected, 1 for being selected
Informational (0)	Information quality	The quality of each answer provided to answer the question posted by a seeker	Number of "up" (positive) votes received by each response	Log transformed value of the number of votes
Individual (1)	Reputation of the contributor	A characteristic of the individual that provided an answer to the question of the seeker	The reputation score of the contributor	Log transformed value of the reputation score of the contributor
Community (2)	Posts edited by the contributor	The number of questions, answers and modifications made on the site to make information easier to understand	Posts edited score reported on the contributor's profile	Log transformed value of the posts edited by the contributor

4. Analysis

In consideration of the proposed multiple levels of the factors that influence the utilization of knowledge assets in an OQA, this study employs a hierarchical regression modeling (HRM) approach [13] for its analysis. The data in this research lends itself to the assumptions of HRM since each question receives a different number of answers and the answers are independent of each other. A random number of answered questions were selected for the analysis, making the data pooled cross-sectional.

Although HRM is suited for analyzing the data, this data also lends itself to decision tree analyses, an unsupervised machine learning technique. This technique is also considered because the sample size. Furthermore, decision tree analysis reveals the relative importance of contributing independent variables rather than simply indicating their significance in explaining the target outcome variable. IBM SPSS version 25 was used for all the analyses in this paper.

5. Results

First, we begin with the descriptive results to understand the distribution of the research variables. Table 2 summarizes the descriptive results and also

provides skewness and kurtosis values for assessing if the distribution of the data is suitable for the chosen analyses. The skewness values are within the acceptable range of zero and the kurtosis values for the research variables are lower than the expected value of 3 [14], indicating that the research variables are normal and independent. These results meet the assumptions of HRM, and therefore make the research variables suitable for HRM analysis.

Table 2. Descriptive statistics of predictor variables

Variables	N	Min.	Max.	Skewness		Kurtosis	
	Value	Value	Value	Value	Std. Err.	Value	Std. Err.
Information Quality (IQ)	1191	-6.910	9.090	-1.300	0.059	0.972	0.118
Contributor's Reputation (RE)	1191	0.000	13.740	-0.531	0.059	-0.192	0.118
Posts edited by the contributor (PE)	1722	-6.910	9.570	-0.471	0.059	-1.318	0.118

The results from the HRM analysis summarized in Table 3 indicate that all three variables are statistically significant predictors of knowledge asset utilization in the OQA context. An R^2 change analysis was also reported to evaluate the contribution of adding each predictor into the model. The significance of the result reported in the R^2 change column in Table 3 indicates that each of the variables tested in the model is an important predictor of knowledge asset utilization in the OQA context. To ensure that the order of introducing the predictors in the model did not affect the model estimation results, similar models were estimated by changing the order of introducing the predictors and the results were consistent.

Table 3. Hierarchical modeling estimation results

Model	Level		Exp (B)	Wald	Classification Accuracy	Nagelkerkie R^2	R^2 Change
0	0	Intercept	0.219***	404.79	82.0%		
1	0	Intercept	0.020***	276.43	85.2%	0.37	0.160***
	1	IQ	2.161***	173.06			
2	0	Intercept	0.000***	174.08	87.7%	0.47	0.421***
	1	IQ	1.634***	63.92			
	2	RE	1.728***	78.44			
3	0	Intercept	0.012***	30.78	89.3%	0.55	0.118***
	1	IQ	1.981***	75.52			
	2	RE	0.839	2.64			
	3	PE	1.682***	56.01			
4	0	Intercept	0.006***	238.79	89.0%	0.55	0.652***
	1	IQ	1.634***	72.43			
	2	RE_x_PE	1.036***	107.78			
		N	1191				

NOTE:PE: posts edited by contributor; RE: contributor's reputation; IQ: votes received per answer; ***significant at 0.001; **significant at 0.01; *significant at 0.05

The results show that the classification accuracy increased with the addition of new predictors, indicating improvement in the ability of the model to represent the decision made by users (to choose an answer from a list of answers provided by contributors).

The results from the HRM analysis suggest that the reputation of the contributor becomes unimportant when posts edited by the contributors is introduced into the model in Model 3. This means that posts edited (a community level factor) is a strong moderating factor of the contributor's reputation scores. Model 4 was developed to examine this moderation effect by creating a product interaction term between the contributor's reputation scores and posts edited. The results as shown in Table 3 suggest that the interaction is significant. This means that user's reputation in the online community and his/her participation in community activities are related.

Although the results from the HRM show that factors across all three levels presented in this study are important predictors of knowledge asset utilization in an OQA community, HRM does not answer the question of the order of importance of the predictors. In other words, should users focus on building reputation before engaging in community activities or the other way around? To examine this question, we employ an unsupervised approach to understanding knowledge utilization (Figures 2-3).

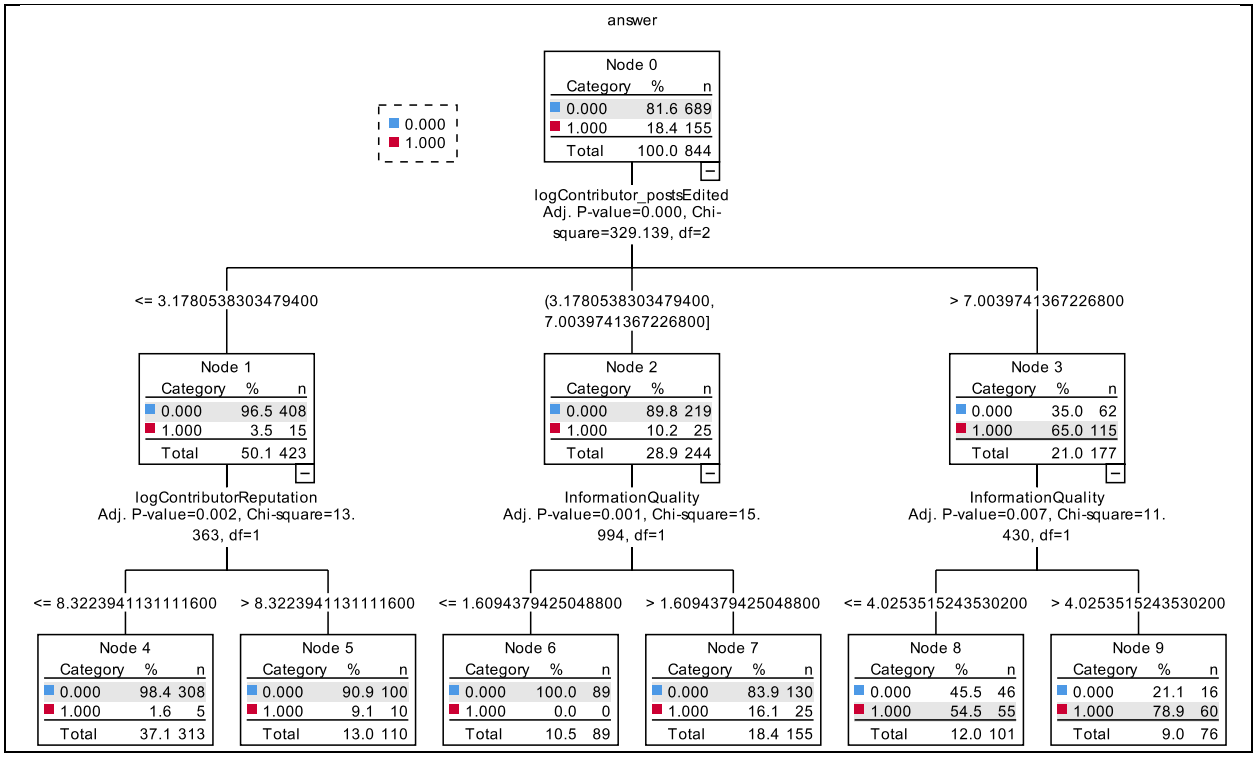

Figure 2. Decision Tree: Training Set Results

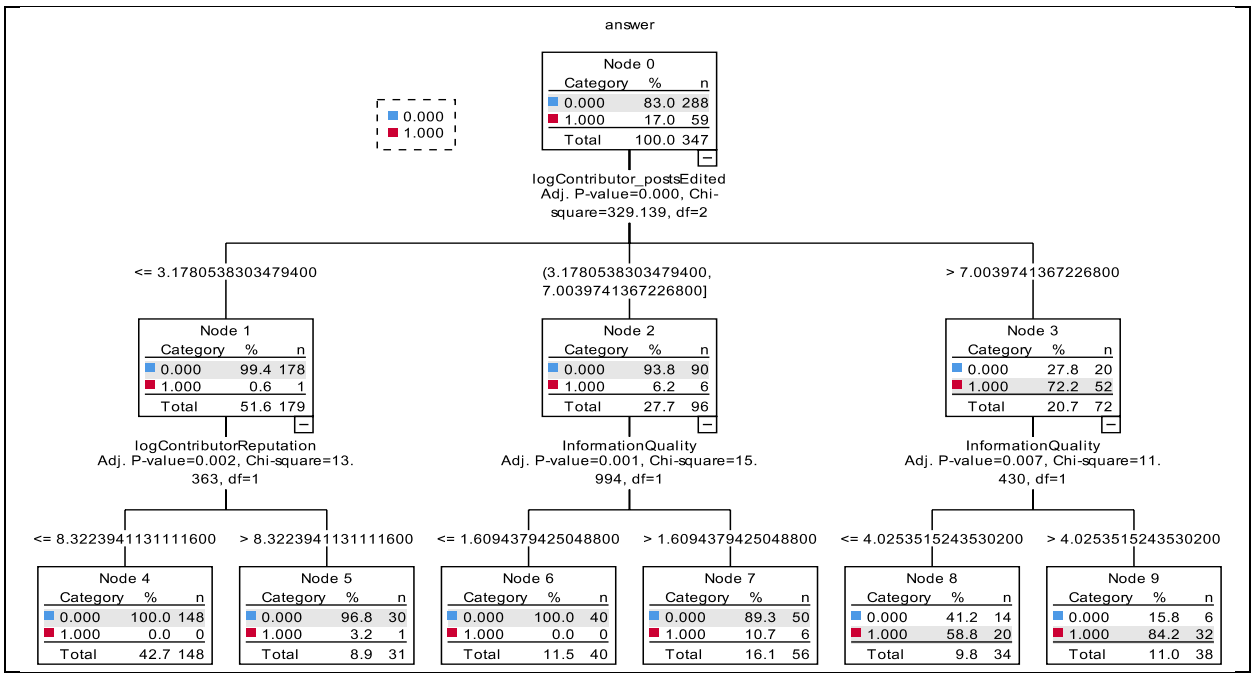

Figure 3. Decision Tree: Test Sample Results

Chi-square automatic interaction detection (CHAID) algorithm in IBM SPSS was used to build the decision tree model. This algorithm was chosen rather than the popular binary decision tree algorithms including classification and regression trees (C&RT) and QUEST (Quick, Unbiased, Efficient Statistical Trees) that only split nodes into two because it splits nodes into more than two nodes as needed to avoid overfitting the model. CHAID algorithm determines the importance of a predictor variable based on the strength of its interaction with the dependent variable.

For all the nodes in Figures 2 and 3, "zero" represents answers that were not chosen and "one" stands for an answer that was selected as the best answer. The optimal sample for training the decision tree (Figure 2) was about 69% of the original dataset and the rest was used to test the accuracy of the model (Figure 3). The decision tree analysis results indicate that the classification accuracy of the final model in Figure 3 is 92.2%. The high classification accuracy of the test model indicates that it is a good model for explaining knowledge asset utilization in an online question and answer site.

The CHAID algorithm labels nodes in the decision tree diagram in the order in which the split was done. It is important to note the p-value of the split node to avoid overfitting[1]. When the p-value is less than 0.05, it means that the split was done on a significant predictor at the split threshold. Interpreting the decision tree in Figure 3, suggests that the posts edited by the contributor (a community level factor) is the most important factor followed by the reputation of the contributor (an individual level factor) and lastly followed by the quality of the information (an informational level factor). Nodes 4 through 9 are the final leaves of the decision tree that enable understanding how the factors are related to influence users' decision to use knowledge assets.

The leaves provide insights into asset utilization as follows. First, the threshold of factors across one level determines what factors from the other levels are relevant to inform the decision-making process. This means that the order of importance of the predictors is such that engagement in community activities such as editing posts is the most important predictor, and the level of engagement determines if the contributor's reputation or the quality of the contributor's information is important. Second, the split conditions indicate thresholds that are necessary for a seeker to consider contributions in an OQA site. Node 5

[1] Using an algorithm such as CHAID that splits nodes into more than two nodes prevents a forced fitting of the model based on the supplied predictors for the model. In an exploratory analysis, the same decision tree was modeled using a forced binary split and the result showed that the first split was done despite the insignificance of the split variable, signifying an overfitted model.

indicates that 0.28[2]% of the time, a seeker is likely to select the answer from a contributor when the order of magnitude of posts edited by the contributor is less or equal to 10^3 and the contributor's reputation is greater than $10^{8.3}$. Node 7 shows that 1.72% of the time, a seeker is likely to select the answer from a contributor when the order of magnitude of the posts edited by the contributor is between $10^{3.2}$ and 10^7 and the quality of information is greater than $10^{1.6}$.

Node 8 indicates that 5.76% of the time, a seeker is likely to select the answer from a contributor when the order of magnitude of the posts edited by the contributor is greater than 10^7 and the quality of information is less than or equal to $10^{4.0}$. Lastly, node 9 shows that 9.26% of the time, a seeker is likely to select the answer from a contributor when the order of magnitude of the posts edited by the contributor is greater than 10^7 and the quality of information is greater than $10^{4.0}$. This explains the relative importance of the factors.

6. Discussion

The focus of this study is twofold. The first is to establish that a multilevel approach to understanding knowledge asset utilization in online communities is useful and richer than the single level approach in the current research stream. The three levels posited include informational, individual and community levels. The second is to uncover the relative importance of factors across the levels. The results demonstrate that in determining knowledge utilization in online communities, the community level factor is the most important followed by individual and finally informational level factors.

Although hierarchical regression modeling is an appealing approach for understanding multilevel models, this study shows that supplementing the analysis with machine learning techniques enables us to uncover additional insights from the same data set. For example, using the decision tree approach, this study clarified the order of influence of each factor considered, beyond the limits of HRM which only classified the strength of the relationship but not the order.

The findings from both analyses have research and practical implications. First, the research focuses on actual knowledge utilization, which is largely understudied in the knowledge management literature because the focus of such literature has been limited to creation, sharing and storing more than use. The nature of the online Q&A platform in this study (SO) required that code be used before it can be evaluated. Second, it draws attention to the need to consider factors from multiple levels rather than one level when examining knowledge use. Third, it establishes an order in which factors across the levels should be considered to understand knowledge utilization behaviors in OQAs.

From a practice standpoint, the finding that community engagement turns out to be the most crucial factor is insightful. Take an organization for instance, the person involved in several working groups, perhaps with a lower reputation or rank compared to the president of the company, is likely to know more about the company than the president. Consequently, questions on specifics in the organization will be better answered by such an individual rather than a top manager who only gets a high-level report. In the context of online communities, engagement in activities that promote the ease of use of content on the site will only build the user's competence through the actual and continued use of the site.

It is interesting that the importance of the contributor's reputation is quickly replaced by their engagement in community activities in the overall model. This aligns with the meaningful framing concept that promotes the idea of engaging in activities that benefit people other than oneself [11]. More importantly, the decision trees reveal that the quality of information in addition to the community engagement of the contributor constitute a stronger predictor of knowledge asset utilization than the combination of the reputation of the contributor and community engagement.

Finally, there are several implications for the design of enterprise online communities. Online communities need to provide its members with information about each user's engagement in improving the community in addition to the reputation of the user. The majority of online communities focus on activities needed for users to build their reputation, rather than providing the opportunity for them to help with sustaining the community. Hence, online community designers need to provide means to engage in activities such as editing posts of others, removing spam messages or questions and answers among others that will be visible to other users. In other words, the focus could be on achieving a community

[2] 0.28% is 3.2% of 8.9% as indicated on Node 5. The same metric will be used to report the likelihood of the events in the leaves.

participation score as opposed to an individual score. After all, the Internet is a highly democratized and social system based on the power of network interactions rather than the celebration of individual high achievers.

6.1. Limitations and Future Research

There are some limitations about the study design that are useful to know before generalizing the results from this study. First, the study focused on questions about javascript. This decision was made to streamline the analysis but future studies can extend this by considering questions from a variety of topics to enable generalizability of the findings. Second, additional factors across the levels may be useful for measuring and understanding the complexity of the decision-making process when selecting answers in online question and answer sites as this model may not necessarily be comprehensive as it focused primarily on the factors identified by Szulanski [7].

7. Conclusion

The results discussed in this paper are preliminary analyses reported from an ongoing study to understand the dynamic creation, sharing and use of knowledge in online communities. The findings provide a roadmap to further investigate other conditions under which knowledge assets contributed are utilized. The results in this paper demonstrate that factors that influence knowledge asset utilization span across three levels including community, informational and individual levels. This paper found that the order of importance of the predictors are community, informational and individual respectively. This finding promotes a higher focus on community-driven design for online knowledge sharing systems, where the strength of community engagement is recognized as the driver as such community success, and as such, is included as an element of evaluation of the individuals. Some online communities are already moving in that direction. For example, ResearchGate reputation score is already a combination of the impact score of publications, but also a combination of users' engagement with the site (creating a project, Q&A, following others, comment on projects, etc.). By design, the community recognizes that it is built not only on the shoulder of giants (those with many publications with high impact factors) but also on the shoulders of many other members that continue to support its accuracy, relevance and goals by actively engaging in sharing, quoting, and re-using the online platform.

8. References

[1] R. Beck, Pahlke, I. and Seebach, C. "Knowledge exchange and symbolic action in social media-enabled electronic networks of practice: A multilevel perspective on knowledge seekers and contributors," *MIS Quarterly* 38:(4), 2014, pp. 1245-1270

[2] L. Khansa, Ma, X., Liginlal, D. and Kim, S. S. "Understanding members' active participation in online question-and-answer communities: A theory and empirical analysis," *Journal of Management Information Systems* 32:(2), 2015, pp. 162-203

[3] A. Durcikova and Gray, P. "How knowledge validation processes affect knowledge contribution," *Journal of Management Information Systems* 25:(4), 2009, pp. 81-108

[4] L. Zhao, Detlor, B. and Connelly, C. E. "Sharing Knowledge in Social Q&A Sites: The Unintended Consequences of Extrinsic Motivation," *Journal of Management Information Systems* 33:(1), 2016, pp. 70-100

[5] D. S. Staples and Webster, J. "Exploring the effects of trust, task interdependence and virtualness on knowledge sharing in teams," *Information Systems Journal* 18:(6), 2008, pp. 617-640

[6] H. A. Simon "Rational decision making in business organizations," *The American economic review* 69:(4), 1979, pp. 493-513

[7] G. Szulanski "Exploring internal stickiness: Impediments to the transfer of best practice within the firm," *Strategic management journal* 17:(S2), 1996, pp. 27-43

[8] C. I. Hovland, Janis, I. L. and Kelley, H. H. *Communication and persuasion; psychological studies of opinion change* Yale University Press, New Haven, CT, 1953

[9] H. H. Chang and Wu, L. H. "An examination of negative e-WOM adoption: Brand commitment as a moderator," *Decision Support Systems* 59:(2014, pp. 206-218

[10] R. E. Petty and Cacioppo, J. T. "The elaboration likelihood model of persuasion," *Advances in experimental social psychology* 19:(1986, pp. 123-205

[11] J. McGonigal. *Reality is broken: Why games make us better and how they can change the world* Penguin, 2011

[12] E. D. Mekler, Brühlmann, F., Opwis, K. and Tuch, A. N. "Disassembling gamification: the effects of points and meaning on user motivation and performance" CHI'13 extended abstracts on human factors in computing systems 1137-1142

[13] K. J. Klein and Kozlowski, S. W. *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* Jossey-Bass, 2000

[14] F. J. Gravetter and Wallnau, L. B. *Statistics for the behavioral sciences* Cengage Learning, 2016