An Empirical Study of Self-Described Texts of Open Source Projects and Their Attention Levels

*Xincheng Wu[1], Hailin Shi[2]

*2021225020104@stu.scu.edu.cn[1], 2022225020069@stu.scu.edu.cn[2]

School of Business, SiChuan University, Chengdu, 610065, China[1,2]

Abstract. Open source, as a means of open innovation for enterprises, is becoming more and more important in the current business competition. In order to analyze the relationship between the text in the enterprise open source project and the project attention, this paper carries out text mining on the supporting text of the code, and analyzes the regression model through the two dimensions of the amount of knowledge and the way of knowledge presentation. The results of the study proved that the amount of original knowledge and the amount of introduced knowledge are positively correlated with user attention, and the knowledge classification in the presentation of the text can effectively improve user recognition, while there may be a U-shaped relationship between the degree of systematization of the text knowledge and user recognition. This study provides an example for open source project management to extract features through text mining and thus conduct empirical research.

Keywords: Corporations, GitHub, Open-source project, Text analysis

1 Introduction

Open source is becoming increasingly important in the current business competition as a means of open innovation for enterprises[1]. By publishing open-source projects, companies export two main types of knowledge, one is code knowledge, while the other is textual introductory knowledge about the code. An introductory document for an enterprise open source project that may be relevant to the project's concerns and serves as a starting point for the flow of enterprise knowledge through the community[2]. Knowledge management focuses on two important processes: knowledge creation and knowledge transfer. Currently, there is a lot of research on KM in the open-source community, but most of it is focused on a single aspect, and not much research has been done on the textual features of the project's supporting presentation.

Creating highly recognized projects helps to increase user engagement, win investment and expand influence for enterprise open source. A 2018 academic research survey of more than 700 developers found that three-quarters of developers decide whether to use a project or become a contributor to a project through the recognition of the open source project [3]. In addition, highly recognized projects have the opportunity to be promoted on various recommendation lists, analysts at certain VC firms review Trending Repos on GitHub for potential investment opportunities, and Runa Capital, which has invested in Nginx and MariaD, has begun to determine the growth potential of open-source startups by tracking project

BDEDM 2024, January 12-14, Ningbo, People's Republic of China

DOI 10.4108/eai.12-1-2024.2347288

recognition. A study by Fan et al. also illustrates that there is a correlation between the popularity of AI repositories and code, reproducibility, and document characteristics[4].

In summary, this paper provides relevant suggestions for enterprises to carry out knowledge management of open source projects through the relationship between text features and project attention. The innovations of the study are as follows:(a) This study considers text as knowledge, and deconstructs the knowledge dimension through text analysis to provide new ideas for research on knowledge flow and knowledge management. (b) This study focuses on the enterprise background elements in open source projects, discusses the relationship between their introductory texts and project concerns, and further refines the open source community participation behavior from the enterprise perspective. (c) There are fewer studies on enterprise-based open source projects, which to a certain extent fills the gap in the field.

2 Research hypothesis

2.1 Textual knowledge content

The enterprise's open source project participates in the social development of the community [5], and the more knowledge it has, the higher the degree of disclosure of the enterprise's knowledge, and the more knowledge it participates in the process of continuous and dynamic knowledge description[6], so that the attention it receives can help the enterprise to gain a competitive advantage. At the same time, enterprises emphasize the use of exclusivity mechanisms to prevent technological spillovers, and set up a large number of governance tools to control the flow of knowledge so that it serves their own interest claims. This behavioral logic is contrary to the core value orientation of open source communities, which makes enterprises face high barriers to participation in open source communities[6][7]. Users realize purposes such as learning knowledge and polishing skills through the knowledge within the open source projects mentioned above [9], and in the process the knowledge achieves flow and innovation.

Code is the core knowledge, which mainly affects users who are motivated to learn technology and can be involved in transformation and positive feedback. Text is the supporting knowledge, which is equally important as the interpretation of the code knowledge, and its content expands the knowledge boundary of the code and promotes the efficiency of knowledge flow. Therefore, in general, the richer the amount of knowledge in the text, the higher the utility to the community users, and the higher the chances of gaining their attention. In addition to its own knowledge content, the introductory text of an open-source project can also introduce knowledge outside the text through hyperlinks, expanding the knowledge of the project. This makes the user can not intuitively access the knowledge of the project, but it greatly expands the knowledge of the project and provides a new presentation channel. Therefore, the more knowledge that is referenced in the text promotes the project's attention. Therefore, hypothesis H1a&H1b is proposed:

H1a: Positive correlation between the amount of original knowledge of the text and user attention.

H1b: The amount of text citation knowledge is positively correlated with user attention.

2.2 Textual knowledge presentation

Knowledge transfer occurs when knowledge diffuses from one entity (e.g., an individual, group, organization) to other entities, and knowledge can be transferred purposefully or can occur unconsciously in other activities [10]. Knowledge transfer and knowledge flow can be considered as components of knowledge sharing, which can occur in the form of broadcasting[11], where knowledge transfer mostly refers to a unidirectional one, while knowledge flow is a reciprocal relationship, a type of knowledge sharing that occurs above the binary level [12].In open source projects, different ways of presenting textual knowledge can increase the reading experience and efficiency of users, which in turn promotes the exchange of knowledge between enterprises and users [12], and ultimately prompts the project to gain attention. At the same time, different ways of presenting textual knowledge have different effects on the manifestation of invisible knowledge of code knowledge in the project, and the presentation of textual knowledge that helps to manifest can enhance the efficiency of knowledge search in the community network, thus enhancing the user attention of the project. The text presentation ability of open source projects can be reflected in two aspects, one is the categorization of textual knowledge and the other is the construction of a system of textual knowledge. The classification of text knowledge helps users to identify whether the knowledge is suitable for them, thus saving the user's learning time, while the classification of knowledge in the text also facilitates the user's knowledge search, thus improving their programming ability, and making it easier to harvest the user's goodwill. Another is to build a system of text knowledge, open source projects presented through the text format, which can make the user more quickly and intuitively understand the full picture of the project knowledge, rapid knowledge search orientation, so the systematization of knowledge can also enhance the user's attention. Therefore, hypothesis H2a&H2b is proposed:

H2a: The degree of knowledge categorization is positively correlated with user attention.

H2b: The degree of knowledge systematization is positively correlated with user attention.

3 Data and method

3.1 Data

The data for this study is based on a self-constructed MongoDB database, which collects data on open source projects operated by Fortune 500 companies in GitHub through a python program, with a total volume of more than 25,000 entries, and the data collection date is December 1, 2023. We identify and manually collect the repositories of the open source projects of the Fortune 500 companies, specifically by searching for the corresponding company name through the User category in GitHub Search, and then comparing the company name, organization avatar, and repository homepage of the returned entries one by one, to ultimately determine the number of corresponding repositories. After obtaining the repository data, we collect all the project links under the repository by traversing the repository web links. Finally, we collect all project-related data from the project links. A final project contains the fields shown in Table 1, Descriptive statistics are shown in Table 2.

Table 1. Variable names, symbols, and meanings(Owner-drawing).

Variable Types	Variable symbol	Variable meaning
Explained Variable	UA	The user acceptance of OSP in GitHub
Explanatory Variables	OTK	Original textual knowledge
	TCK	Text citation knowledge
	DKD	Degree of knowledge disaggregation
	DKS	Degree of knowledge systematization
Control variables	Issues	Whether the company is listed or not
	Size	Project size(KB)
	Time	The year corresponding to the project

Table 2. Descriptive statistics of variables(Owner-drawing).

Variable	Number	Average	S. D.	Minimum	Maximum
UA(Stars)	1,954	55.33	118.7	0	999
UA(Forks)	1,971	36.03	87.15	0	912
OTK	2,000	164.2	215.6	2	3,366
TCK	2,000	6.221	7.771	0	157
DKD(Language)	2,000	3.994	9.315	0	159
DKD(Image)	2,000	0.207	0.936	0	16
DKS1	2,000	3.061	3.489	0	69
DKS2	2,000	1.549	5.105	0	181

3.2 Method

In order to effectively study the relationship between open source project introduction texts and attention, the NLP method used in this study analyzes the project introduction texts in a multidimensional way, including both lexical analysis and dictionary lookup analysis, with lexicality usually being used as the basis for text sentiment analysis. [13]found that most of the opinion words such as adjectives, adverbs, and verbs that affect the sentiment tendency of aspect words by performing lexical statistics on aspect-level sentiment analysis dataset, [14]input lexical labels directly into the model for sentiment classification, [15]utilized lexical embedding by designing a gating mechanism to filtering contextual information. In this paper, four lexical properties, real words, adjectives, adverbs and verbs are extracted as the basis for measuring the knowledge of the introductory text. In addition, the text was also analyzed based on word frequency keywords, based on statistical methods of word importance usually consists of word frequency statistics, such as word frequency (term frequency, T F) [16], word frequency-inverse document frequency (TF-IDF), mutual information, Frequency distribution, etc., and mainly follow the assumption that "words that appear frequently in the text are important, and therefore more likely to be associated with other important terms", statistical-based methods are easy to understand and simple to operate, and have a wide range of applications in the fields of bibliometrics, competitive intelligence, and scientific and technological innovation[17].

Specifically at the variable level, one scholars have reduced the dimension of seven indicators, namely watch, star, fork, commit, contributors, total submitted discussion topics and total PR quantity, to obtain a comprehensive quantitative indicator of project success [18]. In this study, the number of stars expresses the users' recognition of open source projects, so it is more suitable as an explanatory variable in this study. OTK uses the sum of the number of verbs, nouns, adjectives and adverbs in the text as a criterion, which is done by filtering the deactivated words

after jieba segmentation, and then using nltk to perform lexical statistics, filtering out the number of target lexemes and summing up the number of lexemes. text citation knowledge was characterized using hyperlinks to web pages introduced by text into non-projects. Degree of knowledge disaggregation (DKD) was characterized using the number of keywords and the number of hyperlinks to images in programming languages, where the number of keywords in programming languages was compared with the number of hyperlinks to images by jieba disaggregation with the 2022 The number of keywords for programming languages is derived by matching the jieba participle with the top 30 popular programming languages on GitHub in 2022, a variable that has a strong relationship with code knowledge, and the number of image hyperlinks is derived by the number of times a regular expression matches a link to a web page in the text with the img keyword. Degree of knowledge systematization (DKS) characterizes the number of formatting symbols used, and the number of readme documents for open source projects is mostly based on MKI. Most of the readme documents of open source projects are written based on Markdown format, and the use of # can classify the title of the text content, the specific method is to present the readme text segmentation, and segment by segment, extract the first word of each segment, and record the number of times # occurs. Issues is the number of issues discussed, which indicates the degree of collaboration between the enterprise and the user's knowledge, and Size is the amount of code in the project that characterizes the project size, which is closely related to the amount of knowledge in the code. Size is the code volume of the project characterizing the size of the project, which is closely related to the amount of code knowledge. Times denotes the operation time of the project, which is included in the regression model to overcome the time effect. The finalized empirical model is in equation (1):

$$UA = \beta_1 OTK + \beta_2 TCK + \beta_3 DKD + \beta_4 DKS + \beta_i Xi + \varepsilon_i \quad (1)$$

4 Empirical results and analysis

In order to project the introduction of the text and recognition between the about, this paper to build the benchmark model for the model 1, model 1 indicates that the explanatory variables for the field of UA is Star, using six explanatory variables and Size and Times two control variables. Model 2 and model 3 for robustness test, model 2 in the model one based on the addition of Issues control variables, model three in the model two based on the replacement of UA's table pin field for Forks. after determining the regression model, the use of Stata 16 into the regression analysis, the results are shown in Table 3:

Table 3. The result of mixed cross-section regression(Owner-drawing).

Variable	Modle1	Modle2	Modle3
OTK	0.064***	0.053***	0.001
	(0.021)	(0.020)	(0.014)
TCK	1.010***	0.993***	0.619**
	(0.387)	(0.366)	(0.283)
DKD(Language)	1.265***	1.193***	0.461**
	(0.379)	(0.359)	(0.231)
DKD(Image)	9.874***	9.028***	3.510***
	(2.887)	(2.730)	(0.604)

DKS2	2.240** (1.011)	2.573*** (0.956)	3.510*** (0.604)
DKS3	-1.492** (0.639)	-1.429** (0.604)	-1.025** (0.430)
Issues	-	√	√
Size	√	√	√
Times	√	√	√
N	1954	1954	1954
R2	0.137	0.229	0.267

*** represent significant at 10%, 5% and 1% levels respectively

According to Table 3, OTK coefficient (0.064)>0 in model 1, coefficient (0.053)>0 in model 2, and both are significant at 1% level, and coefficient (0.001) is not significant in model 3, the results of model 1 model 2 support H1a, but failed the robustness test of model 3, the reason may be that fork is strongly correlated with the knowledge of the content of the code, and not with the knowledge of the text content. The coefficients of TCK in model 1 (1.010)>0, in model 2 (0.993)0 and both are significant at 1% level, and the coefficients of model 3 (0.619)>0 are significant at 5% level, the regression results are significantly positive and the robustness test is passed, H1b is supported. the coefficients of Language in model 1 (1.265)>0 for DKD, and model 2 (1.265)>0 for DKD, and model 3 (1.265)>0 for DKD. In model 2 (1.193)>0 are significant at 1% level, and in model 3 (0.461)>0 are significant at 5% level; Image coefficient of DKD is (9.874)>0 in model 1 and significant at 1% level; in model 2 (9.028)>0, in model 3 (0.461)>0 are significant at 1% level, and the regression results of DKD's Language and Image coefficients of the regression results are significantly positive and the robustness test is passed, H2a is supported. Regarding DKS, the coefficient of DKS2 in model 1 (2.240) >0, model 2 (2.573) >0 and model 3 (3.510) >0, and significant at 1% level, supports the hypothesis H3a; the coefficient of DKS3 in model 1 (-1.492) <0, model 2 (-1.429) <0 and model 3 (-1.025) <0, which is contrary to the original hypothesis H2b. DKS contrary to the original hypothesis may be due to the different degree of systematization of knowledge, DKS3 represents the number of use of third-level headings, indicating that too much detail in the construction of the knowledge system will instead lead to a decrease in user recognition, and that there may be a U-shape relationship between the degree of systematization of knowledge and the attention of the project. So the relationship between textual knowledge and attention for the final project is shown in Figure 1.

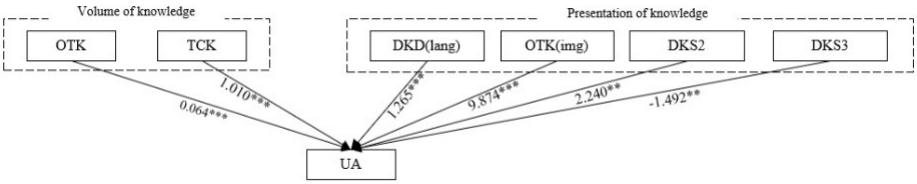

Figure 1. Regression results of textual knowledge and user attention.

5 Research conclusions and implications

5.1 Research conclusions

In this paper, we study the relationship between knowledge and attention in open source projects in the enterprise context, focusing on the mining and analysis of the introductory text accompanying the code knowledge, and summarizing the text features into two dimensions, namely, the amount of knowledge and the knowledge presentation, so as to carry out the empirical regression analysis. The results of the study are as follows: the two dimensions of textual knowledge volume, original knowledge volume and with the introduction of knowledge are positively correlated with user attention, indicating the effectiveness of the enterprise's knowledge disclosure in enhancing the recognition of the project. The presentation of the text, the classification of knowledge in the text can effectively enhance user recognition, while and the degree of text knowledge systematization and user recognition may have a U-shaped relationship. This study provides an example of extracting features through text mining and thus empirical research for open source project management.

5.2 Implications

The above conclusions firstly show that: first, Textual knowledge content is as important as code content knowledge, making it easier to understand the purpose of the code. Enterprises should take measures to improve the quality of project documentation, including clear project overviews, installation and usage instructions, functionality descriptions, contribution guides, screenshots and demos, etc. Secondly, the presentation of knowledge in text is as important as the content of the knowledge, and enterprises need to pay more attention to the presentation of knowledge in the text of the introduction. It is recommended that the relevant personnel in charge of the enterprise learn the coding format of the text presentation, and according to the characteristics and needs of the project, try to improve the aesthetics of the presentation text through the coding of the document, so that it is easier for users to read and understand. For example, the use of '#' prefix and list - shallow drunk, or will emphasize the use of text '**' included, etc.. Enterprises to text-related content for external links to the reference is also a good choice, in a point of knowledge is too much not convenient to intuitively show, give the corresponding web page link, so that users can easily realize the web page between the jump, but also to enhance the user's attention to an effective means. Third, the enterprise should enhance the text and code content linkage degree, provide sample code, running guide, etc., to facilitate the user's debugging and installation, and maximize the attraction of user participation in the open source community.

5.3 Limitations and future research

This article has some shortcomings and future research directions. This article mainly analyzes the two dimensions of text knowledge volume and text knowledge. In the future, first, the semantics of the introductory text can be further analyzed, and currently there is no language model for analyzing the introductory text of open source projects, and extracting semantic features, which is very helpful for facilitating the knowledge management of the enterprise and the knowledge search of the community users. Secondly, the role of some features of the text also needs to be mined and analyzed, such as the key issue of enterprise open source project management, open source license agreement in some projects will be presented differently, how

this affects the user's perception of the enterprise purpose and thus has an impact on the collaboration between the enterprise and the user. Third, there is a need to further systematize the degree of knowledge to collect data to further validate the analysis of its role with the project focus for analysis.

References

[1] L. Corbo, S. Kraus, B. Vlačić, M. Dabić, A. Caputo, and M. M. Pellegrini, "Coopetition and innovation: A review and research agenda," *Technovation*, p. 102624, 2022.

[2] T. Wang, S. Wang, and T.-H. (Peter) Chen, "Study the correlation between the readme file of GitHub projects and their popularity," *J. Syst. Softw.*, vol. 205, p. 111806, Nov. 2023, doi: 10.1016/j.jss.2023.111806.

[3] H. Borges and M. T. Valente, "What's in a github star? understanding repository starring practices in a social coding platform," *J. Syst. Softw.*, vol. 146, pp. 112–129, 2018.

[4] Y. Fan, X. Xia, D. Lo, A. E. Hassan, and S. Li, "What Makes a Popular Academic AI Repository?," *Empir. Softw. Eng.*, vol. 26, no. 1, p. 2, Jan. 2021, doi: 10.1007/s10664-020-09916-6.

[5] Y. Yu, H. Wang, G. Yin, and T. Wang, "Reviewer recommendation for pull-requests in GitHub: What can we learn from code review and bug assignment?," *Inf. Softw. Technol.*, vol. 74, pp. 204–218, 2016.

[6] Y. Wang, Z. Fu, and P. Wu, "Tech-Framework for Semantic Knowledge Management in Conceptual Design," *Data Anal. Knowl. Discov.*, vol. 2, no. 02, pp. 2–10, 2018.

[7] J. WEI and G. Chen, "How to innovate with open source communities: Based on isomorphism — spawned cognitive legitimacy," *Stud. Sci. Sci.*, no. 10 vo 39, pp. 1860–1869, 2021, doi: 10.16192/j.cnki.1003-2053.20201211.002.

[8] Chen, Wei, Li, "Open Source Community: Research Context, Knowledge Framework and Research Prospects," *Foreign Econ. Manag.*, no. 02 vo 43, pp. 84–102, 2021, doi: 10.16538/j.cnki.fem.20200904.402.

[9] K. R. Lakhani and R. G. Wolf, "Why hackers do what they do: Understanding motivation and effort in free/open source software projects," *Open Source Softw. Proj. Sept. 2003*, 2003.

[10] K. D. Joshi, S. Sarker, and S. Sarker, "Knowledge transfer within information systems development teams: Examining the role of knowledge source attributes," *Decis. Support Syst.*, vol. 43, no. 2, pp. 322–335, 2007.

[11] W. Zhang and S. Watts, "Online communities as communities of practice: a case study," *J. Knowl. Manag.*, vol. 12, no. 4, pp. 55–71, 2008.

[12] Y. Pan, Y. C. Xu, X. Wang, C. Zhang, H. Ling, and J. Lin, "Integrating social networking support for dyadic knowledge exchange: A study in a virtual community of practice," *Inf. Manage.*, vol. 52, no. 1, pp. 61–70, 2015.

[13] T. Gu, H. Zhao, Z. He, M. Li, and D. Ying, "Integrating external knowledge into aspect-based sentiment analysis using graph neural network," *Knowl.-Based Syst.*, vol. 259, p. 110025, 2023.

[14] W.-H. Khong, L.-K. Soon, H.-N. Goh, and S.-C. Haw, "Leveraging part-of-speech tagging for sentiment analysis in short texts and regular texts," in *Semantic Technology: 8th Joint International Conference, JIST 2018, Awaji, Japan, November 26–28, 2018, Proceedings 8*, Springer, 2018, pp. 182–197.

[15] K. Shuang, M. Gu, R. Li, J. Loo, and S. Su, "Interactive POS-aware network for aspect-level sentiment classification," *Neurocomputing*, vol. 420, pp. 181–196, 2021.

[16] H. P. Luhn, “A statistical approach to mechanized encoding and searching of literary information,” *IBM J. Res. Dev.*, vol. 1, no. 4, pp. 309–317, 1957.
[17] X. Wang and F. Wang, “Keyword Extraction from a Paper’s Abstract Based on Semantic Text Graph,” *J. China Soc. Sci. Tech. Inf.*, vol. 40, no. 08, pp. 854–868, 2021.
[18] L. Wang, Z. Dong, and Q. Zhang, “Effect of Individual Characteristics of Open Source Software Project Initiator on Project Performance: Empirical Evidence from GitHub,” *Sci. Technol. Manag. Res.*, 2021.