The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.8, pp.374~381

eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.8.374

A Study of GitHub Documentation Repositories: What Makes GitHub Documentation Repository Popular?

Jung Il Kim†

ABSTRACT

Documentation repositories on GitHub are used to share information that is helpful in performing various tasks. Popular documentation repositories have an advantage in attracting contributors who can help manage and extend documentation repository. Therefore, it is important to understand the characteristic of documentation repositories helpful to obtain popularity for developing strategies attracting attention of users. This paper presents a study on GitHub documentation repositories. To conduct the study, we collected 566 documentation repositories from GitHub and manually categorized their topic into 30 topics. Based on the stargazer score of the collected documentation repositories, we divided the collected documentation repositories into popular and unpopular documentation repository groups and investigated the topics in the popular documentation group. Then we statistically examined the differences in README characteristics of the popular and unpopular documentation repository groups. As a result, we found that the studied documentation repositories have 23 popular topics. We also found that the popular and unpopular documentation repository groups have differences in 5 README characteristics. The result of our study indicates that what documentation repository become popular in GitHub.

Keywords : GitHub, Software Engineering, Repository, Software Repository Mining, Empirical Study

깃허브 문서 저장소들에 대한 연구: 무엇이 깃허브 문서 저장소를 유명하게 하는가?

김 정 일†

요 약

깃허브에서 문서 저장소들은 다양한 작업을 수행하는 데 도움이 되는 정보들을 공유하기 위해서 쓰인다. 인기 있는 문서 저장소는 저장소를 관리하고 확장하는 데 도움을 주는 기여자들을 끌어들이는 데 유리하다. 따라서 문서 저장소의 관점에서 사용자들의 관심을 받는 전략을 세우기 위해서 인기 문서 저장소의 특징을 자세히 이해하는 것이 중요하다. 그 특징을 알아보기 위해서 깃허브 문서 저장소를 연구했다. 깃허브에 있는 문서 저장소 566개를 무작위로 수집하고 수집한 문서 저장소들의 주제를 수동으로 분류했다. 별점을 토대로 문서 저장소들을 인기 문서 저장소 집단과 비인기 문서 저장소 집단으로 구분했다. 그런 다음 인기 문서 저장소 집단이 가진 주제들을 추출하고, 인기 문서 저장소 집단과 비인기 문서 저장소 집단이 가지는 README 파일 특징의 차이를 통계적으로 조사했다. 그 결과로 연구 대상 문서 저장소 집단에 23가지 인기 주제가 있다는 것을 찾았다. 또한 인기 문서 저장소와 비인기 문서 저장소 사이에 5가지 README 특징 차이가 있다는 것을 찾았다. 이 연구 결과는 깃허브에서 어떤 문서 저장소가 인기 문서 저장소가 될 수 있는지를 나타낸다.

키워드 : 깃허브, 소프트웨어 공학, 저장소, 소프트웨어 저장소 마이닝, 경험 연구

1. 서 론

깃허브(GitHub)는 세계에서 가장 규모가 큰 소프트웨어 개발자 웹 플랫폼이다. 지금까지 1억 명이 넘는 개발자들이 깃허브에 가입하고 있고, 4억 2천만 개가 넘는 저장소들이 깃허브에서 제공되고 있다[1]. 깃허브에서 저장소들은 다양한 목적을 위해서 만들어진다. 일반적으로 다른 사용자들과 소프트웨어 프로젝트 개발을 함께 하기 위해서 만들지만 쓸모 있는 자원에 대한 정보를 공유하기 위해서 저장소를 만들기도 한다. 후자의 목적을 위해서 만들어지는 저장소를 문서 저장소

※ 이 연구는 과학기술정보통신부의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(1711194613, RS-2023-00213733).
† 정 회 원 : 경북대학교 소프트웨어기술연구소 연구원
Manuscript Received : April 30, 2024
First Revision : July 31, 2024
Accepted : August 5, 2024
* Corresponding Author : Jung Il Kim(2009307043@knu.ac.kr)

1) https://github.com/about

(Documentation Repository: DR)라고 한다[1, 2]. 2014년 7월 11일에 Sindre Sorhus가 Awesome Lists라는 문서 저장소를 제시한 이후로 깃허브 사용자들은 문서 저장소를 “Awesome Lists”라고 하는 토픽[2]으로 분류하고 있다. Sindre Sorhus의 문서 저장소는 계속해서 업데이트되고 있으며 깃허브 사용자들에게 많은 관심을 받아서 2024년 4월 현재까지 73k가 넘는 사용자들이 그 저장소를 따르고(fllowing) 있다. Awesome Lists 토픽으로 분류된 문서 저장소들은 다양한 작업을 수행하는 데 도움이 되는 정보들을 공유하는 수단으로 유용하게 쓰인다[3, 4].

깃허브 저장소의 관점에서 사용자들에게 관심을 많이 받아서 유명해지는 것이 중요하다. 깃허브에서 저장소의 인기(Popularity)는 사용자들에게 받는 관심으로 결정된다 [5]. 깃허브 저장소는 별표(stargazer) 버튼을 통해서 사용자들의 관심을 받는다. 사용자들은 방문한 저장소가 자기에게 필요하거나 중요하다는 이유로 저장소의 별 버튼(Stargazer button)을 클릭해서 저장소에 별점수(stargazer score)을 준다. 별점수를 많이 받은 저장소는 인기 저장소로 분류된다. 인기 저장소는 깃허브 사용자들에게 알려질 기회가 많기 때문에 비인기 저장소보다 저장소 관리에 도움을 주는 기여자들을 끌어들이는 데 유리하다[3].

깃허브에서 인기 저장소가 될 수 있는 방법을 찾기 위해서 인기 저장소와 비인기 저장소들을 조사한 연구들이 최근까지 몇 가지 있었다[3, 6, 7]. Borges 등[3]은 저장소의 분야에 따라서 저장소들이 얻는 별점수에 차이가 있다는 것을 보였다. Fan 등[6]은 인기 AI 학술 저장소와 비인기 AI 학술 저장소들 사이에 몇 가지 특징들의 차이가 있다는 것을 찾았다. Liu 등[7]은 잘 쓰여진 README 파일을 가지는 오픈 소스 자바 프로젝트 저장소들이 사용자들의 관심을 받는 데 유리할 수 있다는 증거를 찾았다. 이 앞선 연구들에서는 주로 소프트웨어 개발 저장소들과 AI 학술 저장소들의 인기를 분석하는 데 초점을 두었다. 그 연구 결과들로 어떤 문서 저장소들이 사용자들에게 인기를 얻을 수 있는지를 자세히 이해하는 데 한계가 있다.

이 논문에서는 깃허브 문서 저장소들을 대상으로 한 연구를 소개한다. 이 연구에서는 다음 두 가지 연구 질문(Research Question)을 탐구한다: (RQ1) 인기 문서 저장소들에 어떤 주제들(Topics)이 있는가? (RQ2) 인기 문서 저장소와 비인기 문서 저장소가 가지는 README 파일의 특징에 차이가 있는가? 이 연구를 위해서 깃허브에 있는 문서 저장소 566개를 연구 대상 문서 저장소 표본으로 수집했다. RQ1의 해답을 알아보기 위해서 연구 대상 문서 저장소들의 주제를 수동으로 분류하고 그 문서 저장소들이 얻은 별점을 토대로 인기 문서 저장소 집단과 비인기 문서 저장소 집단을 구분했다. 그런 다음 인기 문서 저장소 집단의 주제들을 추출했다. RQ2의 해답을 알아보기 위해서 연구 대상 문서 저장소들의 README 파일을 수집하고 수집한 README 파일들의 특징 6가지를 추출했다. 통계 분석을 적용해서 6가지 README 특징들에 대한 인기 문서 저장소 집단과 비인기 문서 저장소 집단의 차이를 분석했다. 이 연구의 결과로 문서 저장소 주제들 가운데 사용자들의 관심을 많이 받은 23가지 인기 주제가 있다는 것을 찾았다. 또한 인기 문서 저장소와 비인기 문서 저장소들 사이에 5가지 README 특징들에 대한 의미 있는 차이가 있다는 것을 찾았다.

이 연구의 기여는 깃허브에서 인기 저장소가 되는 데 중요한 요소를 밝힌 것이다. 이 연구 결과는 깃허브에서 어떤 문서 저장소가 인기 문서 저장소가 될 수 있는지를 나타낸다. 문서 저장소의 주제와 자원 정보의 수는 인기 문서 저장소가 되는 데에 중요한 요소가 된다. 따라서 깃허브에 인기 문서 저장소들이 많이 생기도록 하기 위해서 깃허브 조직(GitHub organization)은 문서 저장소에 대한 인기 주제 범주들을 제공할 필요가 있다. 또한 문서 저장소의 기여자들이 스스로 새로운 자원 정보를 추가하도록 권장할 수 있는 보상 시스템이 있으면 좋을 것 같다고 생각한다.

이 논문의 나머지 구성은 다음과 같다. 2장에서 자세한 연구 방법을 설명한다. 3장에서 이 연구로 얻은 결과들에 대한 논의를 전개한다. 4장에서 이 연구와 관련된 연구들을 소개한다. 5장에서 이 연구의 결론을 제시한다.

2. 연구 방법

2.1 전체 연구 방법

이 연구를 위한 모든 절차는 Fig. 1에 나와 있다. 첫 번째 절차는 깃허브에서 연구 대상이 되는 문서 저장소 표본을 수집하는 것이다. RQ1의 해답을 알아내기 위해서 하는 절차는 세 가지가 있다. 수집한 문서 저장소들의 주제를 분류하고, 인기 저장소와 비인기 저장소들을 결정하고, 인기 문서 저장소들의 주제들을 추출하는 것이다. RQ2의 해답을 알아내기 위해서 하는 절차는 세 가지가 있다. 수집한 문서 저장소들의 README 파일 수집하고, 수집한 README 파일들의 특징을 추출하고, 인기 저장소들과 비인기 저장소들의 README 파일의 특징 비교하는 것이다. 이어지는 부분 절 들에서 이 절차들에 대해서 자세히 설명한다.

2.2 데이터 수집

깃허브에 있는 문저 저장소들을 무작위로 수집했다. 깃허브에서 문서 저장소들을 수집하기 위해서 깃허브가 제공하는 검색 API[3](/search/repositories)를 썼다. 깃허브 검색 API는 질의 단어를 매개 변수를 요구한다. 문서 저장소의 이름과 설명 글에 자주 나오는 단어는 ‘awesome’이다. 이 단어를 질의 단어로 써서 문서 저장소들을 검색했다. 검색 결과에서 이 연

2) https://github.com/topics/awesome

3) https://docs.github.com/en/rest/search/search?apiVersion=2022-11-28

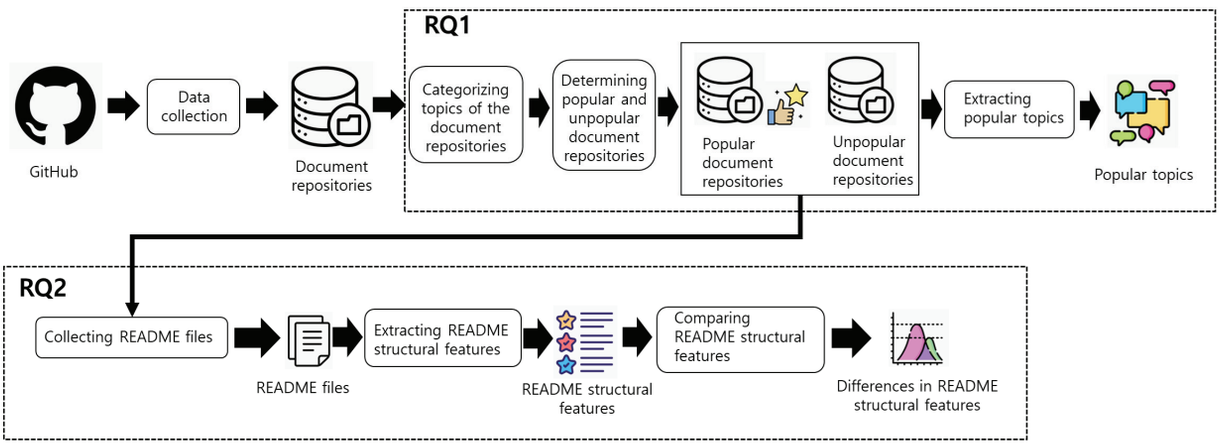

Fig. 1. The Overall Workflows of the Study

구에 적합하지 않은 저장소들이 여럿 포함된 것을 확인했다. 문서 저장소가 아닌 저장소들은 이 연구에 적합하지 않다. 또한 다른 문서 저장소들을 소개하는 문서 저장소와 README 문서가 영어가 아닌 다른 언어로 쓰여진 저장소들은 그 주제를 명확하게 결정할 수 없기 때문에 이 연구에 적합하지 않다. 검색 결과에 있는 저장소들의 README 파일 내용을 보고 그런 저장소들을 찾아서 제외했다. 이 데이터 수집 절차의 결과로 566개 문서 저장소들이 연구 대상 데이터로 남았다.

2.3 RQ1 : 인기 문서 저장소들에 어떤 주제들이 있는가?

연구 동기: 특정한 주제에 대한 유용한 정보를 공유하기 위해서 깃허브 사용자들은 문서 저장소를 만들어 관리한다. 문서 저장소가 가질 수 있는 '주제(Topic)'의 범위는 넓다. 이 연구는 깃허브 문서 저장소들 가운데 사용자들의 관심을 끄는 주제를 가진 인기 문서 저장소와 비인기 문서 저장소가 있다는 가정을 바탕에 둔다. 사용자의

관심을 끄는 문서 저장소의 주제를 이해하는 것은 깃허브 사용자들에게 많은 인기를 받을 수 있는 문서 저장소의 주제를 결정하는 데 도움이 될 수 있다. 따라서 이 RQ의 해답을 아는 것은 중요하다.

연구 방법: 연구 대상 문서 저장소들의 주제를 수동으로 분류했다. 이 논문의 저자와 컴퓨터 과학 전공자 대학원생 한 명이 이 분류 작업에 참여했다. 각 분류자는 연구 대상 문서 저장소를 283개씩 맡아서 각 문서 저장소의 README.md를 읽고 주제를 찾았다. 첫 번째 분류 작업이 끝나고 분류자들은 분류한 문서 저장소들을 서로 바꾸어서 다시 분류 작업을 했다. 같은 문서 저장소에 대해서 두 분류자들의 분류 판단이 같을 경우 그 분류 결과로 주제를 결정했고 두 분류자의 분류 판단이 다를 경우 서로 논의해서 적합한 주제를 결정했다. 이 분류 작업에서 연구 대상 문서 저장소들의 주제 30가지를 찾았다. 그 주제들은 Applications, Artificial Intelligence and Machine Learning, Back-End Development, Big Data, Business and Work, Cloud computing, Computer Science, Computer vision, Data science, Database, Development Environment, Editors, Engineering, Finance, Front-End Development, Gaming, Graphics, Hardware, Health, IoT, Learn, Media, Miscellaneous, Natural Language Processing, Networking, Platform, Programming Language, Security, Software Systems, Testing 들이다.

연구 대상 문서 저장소들을 인기 문서 저장소와 비인기 문서 저장소로 나누는 일에 별점수(stargazers_count)를 인기 척도로 썼다. 별점수는 저장소의 인기를 나타내는 척도이다. 저장소의 별점수는 저장소를 방문한 사용자가 저장소의 별표시 버튼(stargazer button)을 클릭하면 많아진다. 여러 연구들 [3, 4, 6-9]에서도 이 척도를 저장소의 인기 척도로 썼다. 별점수를 토대로 연구 대상 문서 저장소들에서 인기 문서 저장소 집단을 찾았다. 먼저 연구 대상 문서 저장소들이 가진 별점수 값을 기준으로 연구 대상 문서 저장소들을 내림 차순으로 정렬했다. 그런 다음 상위 N%에 속한 문서 저장소들을 인기 문서 저장소 집단으로 묶었다. 이 연구에서는 앞선 연구들 [6, 10, 11, 12]을 따라서 그 N 값을 20%로 결정했다. 이 N 값으로 찾은 인기 문서 저장소 집단에 분류된 문서 저장소 주제들이 포함된 횟수를 계산했다. 이 계산 결과에서 인기 문서 저장소 집단에 한 번 이상 포함된 주제들을 추출했다.

연구 결과: 각 주제와 관련된 인기 문서 저장소 수와 비인기 문서 저장소 수들은 Table 1에 나와 있다. Table 1에서 #. popular repo.와 #. unpopular repo.은 인기 문서 저장소의 수와 비인기 문서 저장소의 수를 각각 나타낸다. 113개 인기 문서 저장소들이 23개 주제들과 관련이 있다는 것을 확인했다. 그 23가지 주제들은 Front-End Development, Artificial Intelligence and Machine Learning, Programming Language, Development Environment, Platform, Security, Back-End Development, Miscellaneous, Business and Work, Media, Cloud computing, Learn, Applications, Big Data,

Table 1. The Categorized Topics in the Target Dataset

Topic category	#. popular repo.	#. unpopular repo.
Platform	9	26
Front-End Development	17	32
Development Environment	10	15
Data science	1	18
Programming Language	11	33
Cloud computing	3	5
Security	8	43
Media	3	5
Artificial Intelligence and Machine Learning	12	55
Computer Science	1	6
Big Data	2	4
Software Systems	2	5
Business and Work	4	11
Miscellaneous	6	71
Back-End Development	7	15
Computer vision	2	20
Natural Language Processing	1	3
Learn	3	9
Finance	2	11
Engineering	2	10
Applications	3	4
Editors	2	6
Hardware	2	1
Database	0	6
Testing	0	5
IoT	0	6
Gaming	0	8
Graphics	0	5
Networking	0	12
Health	0	3

Software Systems, Computer vision, Finance, Engineering, Editors, Hardware, Computer Science, Data Science, Natural Language Processing 들이다. 이 주제들 가운데 Front-End Development와 관련된 인기 문서 저장소들이 가장 많다.

요약하면, RQ1의 해답을 다음으로 정리한다.

> *30가지 문서 저장소 주제들 가운데 23가지 주제 Front-End Development, Artificial Intelligence and Machine Learning, Programming Language, Development Environment, Platform, Security, Back-End Development, Miscellaneous, Business and Work, Media, Cloud computing, Learn, Applications, Big Data, Software Systems, Computer vision, Finance, Engineering, Editors, Hardware, Computer Science, Data Science, Natural Language Processing 들이 인기가 있다.*

2.4 RQ2 : 인기 문서 저장소와 비인기 문서 저장소가 가지는 README 파일의 특징에 차이가 있는가?

연구 동기: README 파일은 저장소를 설명하는 문서 파일로써 저장소에 대한 방문자의 첫 인상에 영향을 끼친다[7]. 잘 쓰여진 README 파일은 방문자에게 좋은 인상을 주는 데 효과가 있다. 따라서 README 파일은 방문자에게 별점수를 받는 데에 중요하다. 소프트웨어 개발 프로젝트 저장소들을 대상으로 한 연구 등[6, 7, 13]에서 README 파일이 가지는 몇 가지 특징들이 인기를 받는 것과 연관될 수 있다는 결과를 보였다. 문서 저장소와 소프트웨어 개발 프로젝트 저장소의 목적은 서로 다르기 때문에 이 두 저장소들의 README 파일 형식도 다르게 구성될 수 있다. 앞선 연구들만으로 문서 저장소의 README 파일 특징과 인기 사이의 연관성을 이해하는데 부족하다. RQ2는 인기 문서 저장소가 되기 위해 어떻게 README 파일을 작성해야 하는지를 알아내는 데에 도움이 된다.

연구 방법: README 파일은 저장소를 설명하는 텍스트 문서이다. 텍스트 문서의 품질은 문서에 포함된 정보의 양으로 평가할 수 있다. 정보가 많은 문서일수록 좋은 문서가 된다. RQ2는 문서 저장소의 README 파일의 품질이 인기를 얻는 것과 관련이 있다는 가설에 토대로 둔다. 연구 대상 문서 저장소들의 README 파일에서 리스트(List), 리스트 아이템(List item), 테이블(Table), 그림(Image), 큰제목(Heads) 링크(Link)들로 문서 저장소와 관련된 내용들을 제공하는 경향이 있다는 것을 관찰했다. 앞선 연구 등[6, 7, 13]에서도 README 파일의 품질을 평가하기 위해 이 요소들을 고려했다. 이들 앞선 연구들을 따라서 이 6개 요소들을 분석 대상 README 특징들로 선택했다.

README 파일 가져오기 GitHub API(GET/repos/{owner}/{repo}/readme)를 써서 연구 대상 문서 저장소들의 README 파일을 수집했다. 이 GitHub API는 주어진 owner와 repo 매개변수 값을 가진 깃허브 저장소의 README 파일 내용을 응답으로 돌려준다. README 파일은 마크 다운 형식으로 쓰여진다. 수집한 README 파일에서 분석 대상 요소들을 체계적으로 추출하기 위해서 두 개 파이썬 라이브러리 markdown과 html.parser를 썼다. markdown과 html.parser은 각각 존 그루버 마크다운(John Gruber's Markdown)과 html 파서를 파이썬으로 구현한 라이브러리이다. 먼저 markdown을 써서 수집한 모든 README 파일들을 HTML 형식으로 변환했다. 그런 다음, html.parser로 변환된 HTML 형식에서 6개 README 특징들을 추출했다.

인기 문서 저장소 집단과 비인기 문서 저장소 집단들이 가지는 6가지 README 특징들의 차이를 확인하기 위해서 크루스칼-왈리스(Kruskal-Wallis) 통계 검정[14, 15]를 적용했다. 크루스칼-왈리스 통계 검정은 비모수 검정으로 크기가 다른 두 개 이상의 독립 표본들 간 분포를 비교하는 데 쓰인다. 각 분석 대상 특징에 대한 검정에서 귀무 가설은 인기 문서 저장

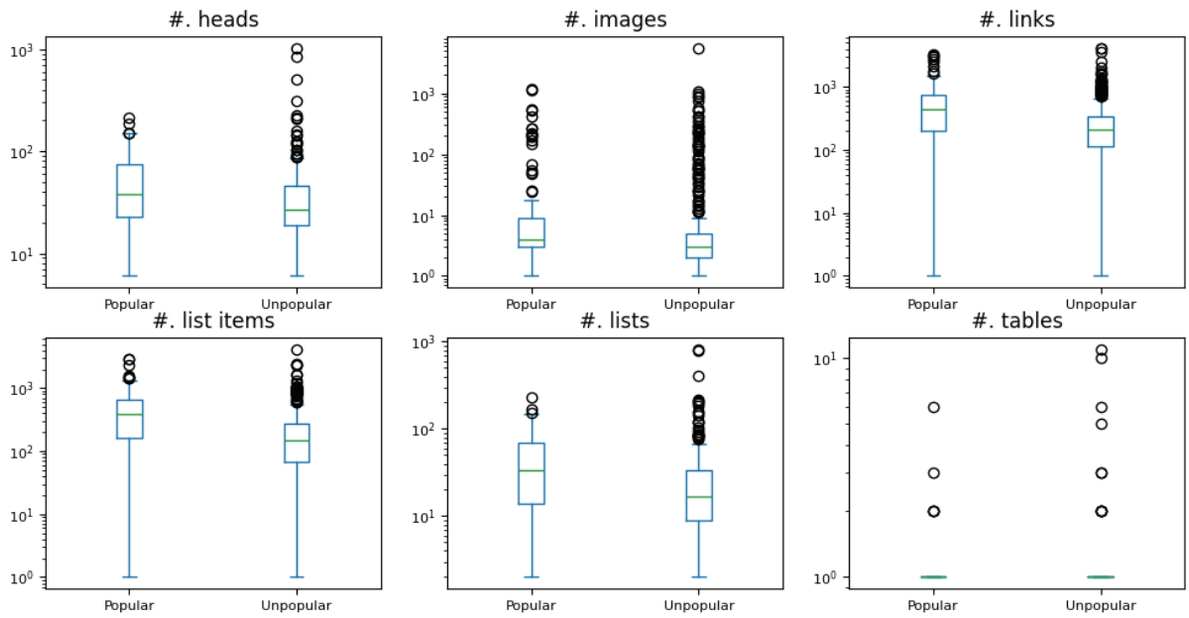

Fig. 2. The Distribution of README Features of the Popular and Unpopular Documentation Groups

소와 비인기 문서 저장소들 사이에 차이가 없음을 대립 가설은 인기 문서 저장소와 비인기 문서 저장소들 사이에 차이가 있음을 각각 나타낸다. 이 가설 검정에서 유의 수준을 1% (p-value 〈 0.01)로 설정했다. 인기 문서 저장소 집단과 비인기 문서 저장소 집단이 가지는 README 특징들의 차이를 계산하기 위해서 코헨'd(Cohen's d) 척도[16]을 썼다. 코헨'd 척도는 두 샘플 분포들이 가지는 평균의 차이를 측정하는 데 사용된다. 이 코헨 척도 결과에서 0.2 이하 값은 작은 차이, 0.2 이상 그리고 0.8 이하 값은 중간 차이 그리고 0.8 이상 값은 큰 차이로 각각 해석했다[17, 18].

연구 결과: Fig, 2는 연구 대상 인기 문서 저장소 집단과 비인기 문서 저장소 집단들이 가지는 6가지 README 특징들의 분포를 보여준다. 이 그림에서 각 특징 값을 나타내는 y축을 로그(Log) 규모로 표시하고 있다. 테이블 수를 제외한 모든 README 특징들에 대해서 인기 문서 저장소 집단과 비인기 문서 저장소 집단들 사이에 차이가 있다는 것을 볼 수 있다.

Table 2는 인기 문서 저장소 집단과 비인기 문서 저장소 집단이 가지는 README 특징들에 대한 크루스칼-왈리스 통계 검정과 코헨'd 계산 결과를 보여준다. 이 통계 검정 결과에서 테이블 수를 제외한 다섯 가지 특징들이 유의 수준 1% 보다 작은 것을 확인했다. Table 2에서 이 결과들을 굵은 글씨로 표시하고 있다. 이 결과는 인기 문서 저장소 집단과 비인기 문서 저장소 집단들 사이에서 이 특징들의 차이가 통계적으로 유의미하다는 것을 의미한다. 테이블 수는 설정한 유의 수준보다 더 큰 것을 확인했다. 이는 그 두 집단들 사이에 이 특징의 차이가 통계적으로 유의미하지 않다는 것을 나타낸다. 코헨'd 계산 결과에서 인기 문서 저장소 집단과 비인기 문서 저장소 집단들 사이에서 링크 수, 리스트 수, 리스트 요소 수, 큰제목 수에 대한 특징들에 중간 정도 차이가 있다는 것과 표 수와 그림 수에 작은 차이가 있다는 것을 확인했다.

Table 2. The Result of the Statistical Tests

feature	p-value	effect d(size)
Num. links	**5.51E-09**	**0.60(M)**
Num. tables	0.194815	0.01(S)
Num. lists	**2.13E-06**	0.29(M)
Num. list items	**4.16E-09**	**0.67(M)**
Num. images	**1.92E-05**	0.04(S)
Num. heads	**3.54E-05**	0.21(M)

요약하면 RQ2의 해답을 다음과 같이 정리할 수 있다.

> *인기 문서 저장소와 비인기 문서 저장소의 README 특징들에 차이가 있다. 특히 6개 README 특징들 가운데 링크 수, 리스트 수, 리스트 요소 수, 큰제목 수에 차이가 있다.*

3. 고 찰

3.1 연구 결과의 의미

이 연구에서는 깃허브 문서 저장소의 인기에 영향을 끼치는 요소들이 몇 가지 있다는 것을 밝혔다. RQ1의 결과로 문서 저장소가 가지는 주제들 가운데 사용자들에게 인기를 많이 얻는 특정한 주제들이 있다는 것을 알았다. 거기서 더 나아가 RQ2의 결과로 인기 문서 저장소와 비인기 문서 저장소의 README 특징에 차이가 있다는 것을 찾았다.

RQ1의 결과는 깃허브에서 인기를 얻은 문서 저장소 주제

들을 나타낸다. 이 결과는 깃허브 조직한테 쓸모있다. 예를 들어, 깃허브 조직은 이 결과를 토대로 사용자들에게 인기 있는 문서 저장소 주제 범주 목록을 정리해서 제공할 수 있다. RQ2의 결과는 문서 저장소가 인기 저장소가 되기 위해서 README 파일에 많은 정보를 포함시키는 것이 중요하다는 것을 나타낸다. 따라서 인기 문서 저장소가 되기 위해서 문서 저장소의 주제와 관련된 다양한 자원을 활발하게 제공할 필요가 있다. 문서 저장소의 관리자는 이 일을 체계적으로 할 수 있는 방법을 제공하고 기여자들에게 그 방법을 잘 지키고 따르는 것을 권장해야 한다. 어떤 방법이 문서 저장소의 확장에 효과가 있는지 알아보는 것도 중요할 것 같다.

더 나아가 RQ1과 RQ2의 결과는 문서 저장소가 얻을 인기를 예측하는 모델[4]을 설계하는 데 쓰일 수 있다. 이 예측 모델은 문서 저장소의 주제와 README 파일 특징들을 설명 변수로 쓰고 별점수를 예측 변수로 써서 주어진 문서 저장소의 설명 변수들을 바탕으로 얻게 되는 별점수를 미리 예상한다. 별점수를 많이 받을 수 있게 문서 저장소를 개선하기 위해 이 예측 모델은 도움이 될 것이다.

3.2 문서 저장소의 인기와 다른 요소들의 상관 관계

문서 저장소의 요소들 가운데 기여자 수(Number of contributors)와 활동 시간(Active time)은 문서 저장소의 별점수와 상관 관계가 있을 수 있다. 인기 문서 저장소에는 기여하는 사람들이 많이 모일 수 있고, 오랜 시간 운영된 문서 저장소는 사용자들에게 별점수를 받을 수 있는 기회가 많을 수 있기 때문이다. 이 요소들과 별점수 사이에 상관 관계가 있는지 알아보기 위해서 상관 관계 분석을 했다. 이 상관 관계 분석에서 문서 저장소의 활동 시간으로 문서 저장소가 마지막으로 갱신된 날짜(updated date)에서 처음 생긴 날짜(created date)를 뺀 날의 수를 썼다.

문서 저장소가 얻은 감시자 수(Watchers count)와 포크 횟수(Fork count)도 별점수처럼 인기 척도가 될 수 있다. 이 두 요소들과 별점수 사이에 상관 관계가 없다면 문서 저장소의 인기 척도로 이 두 요소들을 함께 고려할 필요가 있다. 이 두 요소들과 별점수의 상관 관계를 알아보기 위해서 상관 관계 분석을 했다. Table 3에 이 상관 관계 분석으로 얻은 결과가 나와있다. 감시자 수와 포크 횟수는 완벽한 양성 관계, 아주 강한 양성 관계가 각각 있다는 것을 확인했다. 이 결과는 감시자 수와 포크 횟수는 별점수와 거의 완전하게 중복된다는 것을 나타낸다. 기여자

수와 활동 시간은 중간 양성, 아주 작은 양성 관계가 각각 있다는 것을 확인했다. 이 결과는 기여자 수와 활동 시간은 별점을 많이 받는 것에 큰 영향을 끼치지 않는다는 것을 나타낸다.

Table 3. The Result of Correlation Analysis of Number of Contributors, Active Days, Watchers Count and Fork Count with Popularity

Feature	Correlation
#. contributors	0.51
Days of life	0.18
Watchers count	1
Fork count	0.9

3.3 인기 문서 저장소와 비인기 문서 저장소의 또 다른 차이점들

RQ2의 결과로 인기 문서 저장소와 비인기 문서 저장소들 사이에 몇 가지 README 특징들에 양적 차이가 있다는 것을 알았다. 더 나아가 인기 문서 저장소와 비인기 문서 저장소들 사이에 다른 차이점이 있는지를 알아보기 위해서 간단한 질적 분석을 했다. 이 질적 분석에서 Front-End Development 주제와 관련된 인기 문서 저장소 “vuejs/awesome-vue”와 비인기 문서 저장소 “Urigo/awesome-meteor”를 각각 분석 대상으로 선택했다. 분석 대상으로 선택된 인기 문서 저장소는 7만 1천이 넘는 별점수를 받았고 비인기 문서 저장소는 1400이 넘는 별점수를 받았다. 이 두 저장소의 깃허브 페이지 내용들을 비교해서 다음 2가지 차이점이 있다는 것을 찾았다. 첫 번째는 분석 대상 인기 문서 저장소는 두 기업(Open Collective와 TIDELIFT)에게 지원을 받고 있지만 비인기 문서 저장소는 기업 지원이 없다는 점이다. 기업 지원을 받는 문서 저장소는 기업 지원이 없는 문서 저장소보다 더 믿을 만하다. 또한 기업 지원이 있는 문서 저장소에 많은 기여를 하면 기업에게 보상을 받는 것을 기대할 수 있다. 이는 문서 저장소의 기여자들이 기여 활동을 활발하게 하는 동기가 될 수 있다. 두 번째는 분석 대상 인기 문서 저장소가 비인기 문서 저장소보다 더 자세한 기여 안내서(CONTRIBUTING.md)를 제공한다는 점이다. 기여 안내서는 문서 저장소에 기여하는 방법을 설명하기 위해서 제공된다. 자세하게 쓰인 기여 안내서는 기여자들이 문서 저장소에 기여하는 방법을 이해하는 데 큰 도움이 된다. 이로 인해 문서 저장소에 기여자들을 끌어들이는 일을 수월하게 하는 효과가 있을 수 있다. 이 두 가지 차이점이 인기 문서 저장소와 비인기 문서 저장소에 기여하는 기여자 수와 제공되는 자원 정보의 수에 큰 차이가 생기게 할 수 있다고 생각한다. 그 결과로 많은 사람들에게 별점수를 얻을 수 있는 결과로 이어질 수 있다고 생각한다.

3.4 연구의 한계점

외부 타당성 우려는 이 연구에서 사용한 문서 저장소 표본과 관련 있을 수 있다. 연구 대상 문서 저장소 표본을 수집하기 위해서 깃허브 검색 API를 썼다. 깃허브 검색 API는 주어진 조건을 따라 한 번에 가장 많이 1000개 저장소들에 대한 검색 결과를 돌려준다. 깃허브 검색 API로 깃허브에 있는 모든 저장소들을 검색하는 데 한계가 있다. 깃허브에 있는 문서 저장소들 가운데 연구 대상 문서 저장소 표본에 포함되지 않은 문서 저장소들이 있을 수 있다. 그러나 깃허브 검색 API는 앞선 연구 등[5, 7, 19]에서도 깃허브 저장소들을 찾는 데 쓰

였고, 그 연구들은 믿을 수 있는 결과를 얻었다. 따라서 이 외부 타당성 우려가 크지 않다고 생각한다.

내부 타당성 우려는 인기 저장소 집단과 비인기 저장소 집단을 결정하는 방법과 관련이 있다. 이 연구에서는 별점수로 정렬된 연구 대상 문서 저장소 집합에서 상위 20% 문서 저장소들을 인기 문서 저장소 집단으로 구성하고 나머지 문서 저장소들을 비인기 문서 저장소 집단으로 구성했다. 상위 N%를 조절하는 것으로 인기 문서 저장소 집단과 비인기 문서 저장소 집단이 다르게 구성될 수 있다. 별점수는 깃허브 저장소의 인기 척도로 자주 사용되어 왔지만 그 척도를 바탕으로 인기 저장소를 결정하는 정확한 기준 값은 아직 제시되지 못하고 있다. 앞선 연구 [6]에서는 인기 저장소 집단을 식별하는 데 상위 20%를 경계 값으로 쓰는 것이 제법 타당하는 것을 보였다. 이 연구에서도 그 경계 값을 바탕으로 인기 문서 저장소 집단을 결정했다. 따라서 이 내부 타당성 우려는 아주 작다고 생각한다.

4. 관련 연구

깃허브에서 어떤 저장소들이 사용자들에게 관심을 끄는지를 이해하기 위해서 최근까지 여러 연구들이 있었다. Weber 등[5]은 깃허브에 있는 파이썬 소프트웨어 프로젝트 저장소들의 코드 인기를 조사했다. 그 조사에서 파이썬 추상 구분 트리와 관련된 20개 특징들과 프로젝트 관리와 관련된 18개 특징들과 코드 인기 사이의 관련성을 분석했고 그 특징들과 코드 인기 사이에 연결점이 있다는 것을 관찰했다. Borges 등[3]은 2500개 깃허브 인기 저장소들을 조사했다. 프로그래밍 언어, 어플리케이션 영역 그리고 저장소 소유자에 따라서 저장소가 가지는 인기가 다르다는 것을 관찰했다. 또한 저장소가 포크된(Forked) 횟수가 저장소 인기와 강한 상관 관계가 있다는 것을 찾았다. 이어서 한 연구[4]에서 깃허브 저장소가 과거에 얻은 별점수 이력을 바탕으로 앞으로 얻을 수 있는 별점수를 예측하는 실험을 했다. 이 실험에서 다중 선형 회귀 모델을 써서 얻은 결과로 과거 6개월 별점수 이력을 학습 데이터로 사용해서 저장소가 얻을 별점수를 낮은 오류로 예측할 수 있다는 결과를 보였다. Fan 등[6]은 인공 지능 학술 저장소들의 인기에 대해서 연구했다. 인공 지능 학술 저장소의 논문 인용 횟수와 인공 지능 학술 저장소의 별점수 사이에 강한 상관 관계가 있다는 것을 관찰했다. 또한 README 파일과 관련된 10개 특징들 가운데 6개 특징들이 인공 지능 학술 저장소가 별점수를 얻는 데 영향을 미친다는 것을 보였다. Venigalla 등[13]은 저장소의 README 특징들과 저장소의 별점수 사이에 있는 상관 관계를 조사했다. 그 조사에서 인기 저장소들의 README 파일은 리스트, 이미지, 외부 소스 링크 요소들로 잘 구성된다는 것을 관찰했다. Liu 등[7]은 오픈 소스 자바 프로젝트 저장소들의 README 파일들을 비교했다. 통계 분석을 통해서 깃허브 안내서를 준수해서 작성된 README 파일들이 많은 별점수를 얻는다는 것을 보였다. 우리 연구는 이 앞선 연구들을 확장한다. 문서 저장소에 초점을 두고 문서 저장소의 인기와 관련된 주제 범주와 README 파일 특징을 찾았다. 이는 앞선 연구들에선 고려하지 않은 것이고 우리 연구와 앞선 연구들의 차이점이다.

5. 결 론

인기 문서 저장소의 특징을 이해하기 위해서 문서 저장소들을 대상으로 하는 한 경험 연구를 수행했다. 이 연구를 하기 위해서 566개 깃허브 문서 저장소들을 무작위로 수집했다. 수집한 문서 저장소들의 주제 범주를 수동으로 찾고 별점수를 바탕으로 수집한 문서 저장소들을 인기 문서 저장소 집단과 비인기 문서 저장소 집단으로 구분했다. 인기 문서 저장소 집단에 대한 수동 분석을 통해서 23가지 인기 주제들이 인기가 있다는 것을 관찰했다. 문서 저장소가 제공하는 README 파일의 특징이 별점수를 받는 데 미치는 영향을 알아보기 위해서 인기 문서 저장소 집단과 비인기 문서 저장소 집단의 README 특징을 비교했다. 그 결과로 두 집단 사이에 5개 README 특징들에 명확한 차이가 있다는 것을 확인했다. 이 연구 결과는 문서 저장소의 주제와 README 특징이 사용자들에게 인기를 얻는 데 영향을 미친다는 것을 나타낸다.

인기 문서 저장소의 특징을 더 자세히 이해하기 위해서 추가 연구를 계획하고 있다. 먼저 연구 대상 문서 저장소들을 확장해서 연구 결과에 대한 일반화를 강화한다. 또한 기계 학습 모델을 활용해서 문서 저장소가 인기를 얻는 데 중요한 특징들이 무엇인지를 조사한다.

References

[1] E. Kalliamvakou, G. Gousios, K. Blincoe, L. Singer, D. M. German and D. Damian, “The promises and perils of mining github,” in *Proceedings of the 11th Working Conference on Mining Software Repositories*, pp.92-101, 2014.

[2] E. Kalliamvakou, G. Gousios, K. Blincoe, L. Singer, D. M. German, and D. Damian, “An in-depth study of the promises and perils of mining GitHub,” *Empirical Software Engineering*, Vol.21, pp.2035-2071, 2016.

[3] H. Borges, A. Hora, and M. T. Valente, “Understanding the factors that impact the popularity of GitHub repositories,” in *2016 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pp.334-344, 2016.

[4] H. Borges, A. Hora, and M. T. Valente, “Predicting the popularity of github repositories,” in *Proceedings of the The 12th International Conference on Predictive Models and Data Analytics in Software Engineering*, pp.1-10, 2016.

[5] S. Weber and J. Luo, "What makes an open source code popular on github?," in *2014 IEEE International Conference on Data Mining Workshop*, pp.851-855, 2014.

[6] Y. Fan, X. Xia, D. Lo, A. E. Hassan, and S. Li, "What makes a popular academic AI repository?," *Empirical Software Engineering*, Vol.26, pp.1-35, 2021.

[7] Y. Liu, E. Noei, and K. Lyons, "How ReadMe files are structured in open source Java projects," *Information and Software Technology*, Vol.148, pp.1-11, 2022.

[8] K. Aggarwal, A. Hindle, and E. Stroulia, "Co-evolution of project documentation and popularity within github," in *Proceedings of the 11th Working Conference on Mining Software Repositories*, pp.360-363, 2014.

[9] J. Zhu, M. Zhou, and A. Mockus, "Patterns of folder use and project popularity: A case study of GitHub repositories," in *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pp.1-4, 2014.

[10] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: Classification of skewed data," *ACM SIGKDD Explorations Newsletter*, Vol.6, No.1, pp.50-59, 2004.

[11] T. L. Alves, C. Ypma, and J. Visser, "Deriving metric thresholds from benchmark data," in *2010 IEEE International Conference on Software Maintenance*, pp.1-10, 2010.

[12] M. Yan, X. Xia, X. Zhang, D. Yang, and L. Xu, "Automating aggregation for software quality modeling," in *2017 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pp.529-533, 2017.

[13] A. S. M. Venigalla and S. Chimalakonda, "An empirical study on correlation between readme content and project popularity," *arXiv e-prints, arXiv-2206*, 2022.

[14] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, Vol.47, No.260, pp.583-621, 1952.

[15] G. W. Corder and D. I. Foreman, "Nonparametric statistics for non-statisticians," Hoboken: John Wiley & Sons. pp. 99-105. ISBN 9780470454619.

[16] A. B. Cantor, "Sample-size calculations for Cohen's kappa," *Psychol Methods*, Vol.1, No.150, 1996.

[17] M. Hess and J. Kromrey, "Robust confidence intervals for effect sizes: A comparative study of cohen's d and cliff's delta under non-normality and heterogeneous variances," in *the Annual Meeting of the American Educational Research Association*, pp.1-30, 2004.

[18] E. Noei, F. Zhang, S. Wang, and Y. Zou, "Towards prioritizing user-related issue reports of mobile applications," *Empirical Software Engineering*, Vol.24, pp.1964-1996, 2019.

[19] G. A. A. Prana, C. Treude, F. Thung, T. Atapattu, and D. Lo, "Categorizing the content of github readme files," *Empirical Software Engineering*, Vol.24, pp.1296-1327, 2019.

[20] F. Zanartu, C. Treude, B. Cartaxo, H. S. Borges, P. Moura, M. Wagner, and G. Pinto, "Automatically categorising github repositories by application domain," *arXiv preprint arXiv:2208.00269*.

김 정 일

https://orcid.org/0000-0001-6442-1152
e-mail : 2009307043@knu.ac.kr
2017년 경북대학교 컴퓨터학부(박사)
2017년 ~ 2022년 경북대학교 소프트웨어기술연구소 박사후연구원
2022년 ~ 2024년 경북대학교 자율군집소프트웨어연구센터 연구원
2024년 ~ 현 재 경북대학교 소프트웨어기술연구소 연구원
관심분야 : Software Engineering, Mining Software Repositories, Data Mining, Automated Software Engineering