The promises and perils of mining GitHub

Typ	Konferensartikel
Författare	Eirini Kalliamvakou
Författare	Georgios Gousios
Författare	Kelly Blincoe
Författare	Leif Singer
Författare	Daniel M. German
Författare	Daniela Damian
Webbadress	http://dl.acm.org/citation.cfm?doid=2597073.2597074
Ort	Hyderabad, India
Utgivare	ACM Press
Sidor	92-101
ISBN	978-1-4503-2863-0
Datum	2014
DOI	10.1145/2597073.2597074
Hämtad den	2019-09-27 14:53:59
Bibliotekskatalog	Crossref
Namn på konferens	the 11th Working Conference
Språk	en
Sammanfattning	With over 10 million git repositories, GitHub is becoming one of the most important source of software artifacts on the Internet. Researchers are starting to mine the information stored in GitHub's event logs, trying to understand how its users employ the site to collaborate on software. However, so far there have been no studies describing the quality and properties of the data available from GitHub. We document the results of an empirical study aimed at understanding the characteristics of the repositories in GitHub and how users take advantage of GitHub's main features—namely commits, pull requests, and issues. Our results indicate that, while GitHub is a rich source of data on software development, mining GitHub for research purposes should take various potential perils into consideration. We show, for example, that the majority of the projects are personal and inactive; that GitHub is also being used for free storage and as a Web hosting service; and that almost 40% of all pull requests do not appear as merged, even though they were. We provide a set of recommendations for software engineering researchers on how to approach the data in GitHub.
Protokolltitel	Proceedings of the 11th Working Conference on Mining Software Repositories - MSR 2014
Tillagd den	2019-09-27 14:53:59
Ändrad den	2019-10-07 12:49:27

Etiketter:

GitHib mining

Bifogade dokument

- Kalliamvakou m. fl. - 2014 - The promises and perils of mining GitHub.pdf

  Författarna diskuerar saker man bör tänka på när man utvinner information från GitHub:

  1. Ett projekt kan bestå av flera repron. Om en commit görs i en fork som sedan pullas till huvudreprot via en GitHubpullrequest syns inte de commits som gjordes till forken i huvudreprots commitlog. Slutsats: **Titta även på forkade repron när du analyserar commithistoriken för ett projekt.** Likaså, om någon gör en pullrequest, får en kommentar på denna, uppdaterar sitt bidrag och gör en ny pullrequest syns bara den sista pullrequesten.
  2. De flesta projekt har väldigt få commits och/eller är inaktiva: När artikeln skrevs hade 54% av projekten varit inaktiva det senaste halvåret. 32% av alla repron på GitHub hade bara varit aktiva under 1 dag.
  3. Många projekt innehåller inte mjukvara (=kod som man kan bygga verktyg av), utan används exempelvis för demos och tutorials, eller för att hosta websidor eller lagra data för eget bruk.
  4. 2/3 av alla projekt är personliga, d.v.s. används (commitas till) bara av 1 person.
  5. Väldigt få projekt (även av de icke-personliga) har pull-requests. Ändå var det 1700 projekt som fick >100 pull-requests under 2013, så det existerar data om man vill studera samarbete, men man ska vara medveten om att de flesta projekt inte är relevanta för detta.
  6. Om man mergear en fork med git (utanför GitHub) ser det ut på GitHub som att den är omergead. Författarna har ett antal tumregler som man kan använda för att hitta i alla fall en del av dessa fall. (Troligtvis inget vi behöver göra.)
  7. Många projekt har aktivitet (så som utveckling och ärendehantering) pågående utanför GitHub.

  Författarna lyfter dock fram att detta inte påverkar exempelvis analyser av språk och verktygstyper.

  GitHubs innehåll växer (växte) snabbt: När artikeln skrevs (juli 2013) hade 34% av alla GitHubprojekt skapats de senaste 6 månaderna!

  Verktyg för att utvinna metadata från GitHub:

  - GHTorrent (har recordat GitHubevents sedan 2012). Det finns en referens till en källa som förklarar varför den inte kan innhålla allt data.
  - Gitminer (för historik från specifika repron).

  Vidare innehåller http://www.githubarchive.org historik över GitHub-events.