

Streszczenie

Niniejsza praca przedstawia innowacyjne rozwiązanie problemu definiowania i mierzenia jakości w zespołach programistów tworzących wolne oprogramowanie (*ang.* open-source software - OSS). W badaniach nad programistami tworzącymi OSS było dotychczas kilka nietkniętych obszarów, które to stały się łatwiejsze do zbadania ze względu na rosnącą popularność oprogramowania typu OSS – a co za tym idzie, łatwiejszą dostępność danych.

Ta praca badawcza wykorzystuje owoce rewolucji technologicznej Big Data (*ang.* duże zbiory danych). Skorzystano z ogólnodostępnej bazy danych zawierającej informacje o bieżących i przeszłych projektach OSS portalu GitHub - obecnie największego portalu hostującego kod źródłowy OSS - gdzie programiści z całego świata współpracują nad projektami.

Duży nacisk kładzie się na empiryczną część badań, co oznacza, że wszystkie założenia są weryfikowane przez dane – nie bazujemy wyłącznie na wiedzy branżowej ani na założeniach zbudowanych z badań nad mniej licznymi społecznościami OSS. Nasze badania wykorzystują dane historyczne na temat wszystkich projektów open-source hostowanych na portalu GitHub, które zostały pozyskane z publicznej bazy danych „GitHub Torrent”.

Prace podzielono na dwie główne części, pierwsza polega na modelowaniu jakości zespołów OSS, a druga na analizie tego, jak efektywnie można dzielić pracę między programistów. Najpierw zdefiniowaliśmy sposób modelowania jakości w oparciu o charakterystykę tworzenia i działania zespołów GitHub. Następnym krokiem było stworzenie solidnych miar jakości, które można zastosować do każdego zespołu tworzącego wolne oprogramowanie. Jego podstawowym celem było wyjaśnienie, które czynniki wpływają na to, jak zespół OSS radzi sobie z informacją zwrotną od użytkowników oprogramowania jak i od innych programistów.

Finalnie, wynikiem naszych badań są zaobserwowane empirycznie cechy zespołów OSS przyczyniające się do wysokiej jakości wsparcia użytkowników. Dzięki temu wiemy, jak wspierać tworzenie takich zespołów i jakie czynniki mają na nie wpływ. Nasze najważniejsze odkrycia wspierają pojęcie Brook’a o istnieniu „zespołów chirurgicznych” programistów, którzy są dobrze zorganizowani, pracują wydajnie, a tym samym zapewniają wysokiej jakości informacje zwrotne dla użytkowników oprogramowania.

Druga część pracy poświęcona jest empirycznemu badaniu technik przydzielania pracy w zespołach OSS programistów, wspierane przez eksperymenty symulacyjne i testowanie różnych ustawień i strategii. W rezultacie, w oparciu o rzeczywiste dane historyczne portalu GitHub, przetestowaliśmy obliczeniowo, jak szybko różne strategie alokacji zadań prowadzą do zakończenia pracy nad projektami GitHub. Wybrane strategie przydzielania zadań zostały przetestowane i porównane z wynikami centralnego planowania i strategii sieciowych.

Badania stworzyły podstawowy zasób wiedzy (*ang.* know-how) dla każdego, kto jest zainteresowany pomiarem i wsparciem jakości w zespołach programistów open-source. Praca przyczynia się do rozwoju informatyki oraz informatyki społecznej, ponadto może być przydatna dla praktyków zarządzania organizacjami. Wreszcie, wnioski przedstawione w niniejszej pracy doktorskiej można zastosować do innych, podobnych wirtualnych zespołów, którego członkowie współpracują globalnie na całym świecie w celu tworzenia wysokiej jakości treści, tacy jak edytorzy Wikipedii.

Abstract

This thesis presents an approach to tackle the problem of defining and measuring quality in open-source software (OSS) teams of developers. There were some significant *terrae incognitae* in the OSS research, which are now easier to investigate due to rising popularity of OSS. This research work makes use of the benefits of the data revolution, namely of the publicly available big datasets holding information on current and past OSS projects located on the GitHub portal – currently the largest open-source hosting portal.

A significant emphasis is put on the empirical part of the research, which means we make sure that all assumptions are verified by data instead of using evangelic statements from the industry or relying strictly on past research. Our research utilises trace data holding information on all OSS projects hosted on the GitHub portal, which was acquired from the GitHub Torrent public data source.

Work was divided into two main parts, the first is modelling the quality of OSS teams, the second, analysing how work can be distributed efficiently among developers. We first defined how quality can be modelled based on characteristics of how GitHub teams are created and operate. The next step was to create a robust quality model which can be applied to every open-source team. Its fundamental goal was to explain which factors affect how an OSS team deals with direct feedback coming from both software users and other developers. The final results of our research were empirical evidence describing characteristics of OSS teams providing high quality user support. Thanks to that, we know how to foster creation of such teams and which factors strongly influence on such quality. Our most important finding supports Brook's famous notion of 'surgical teams' of developers, who are well structured, work efficiently and thus provide high quality feedback to software users.

The second part of the thesis is devoted to empirical investigation of work allocation techniques in the OSS teams of developers supported by doing simulation experiments and testing different settings and strategies. As a result, basing on real GitHub trace data, we tested computationally how fast different task allocation strategies lead to completion of the work on GitHub projects. Emergent task allocation strategies were tested and compared with performance of a central planer and network based strategies.

This research, providing the above mentioned contributions, is a baseline know-how for any party interested in measuring and supporting quality in open-source software teams of developers. Work contributes to the field of social informatics and computer science, and can potentially be useful for practitioners of business management. Last but not least, findings presented in this PhD thesis can be applied to some other, similar virtual 'cyber' teams of innovative people, who cooperate globally to create a high quality content such as teams of Wikipedia editors.