Warszawa, 23 kwietnia, 2019

dr hab. Dominik Ślęzak, prof. UW
Instytut Informatyki
Uniwersytet Warszawski
ul. Banacha 2, 02-097 Warszawa
slezak@mimuw.edu.pl

RECENZJA ROZPRAWY DOKTORSKIEJ

mgr inż. Oskara Jarczyka

pt. „Empirical Analysis of Open-Source Software Development on GitHub"

Przedstawiona do recenzji rozprawa doktorska Pana mgr Oskara Jarczyka składa się z części wstępnej, zawierającej streszczenie, spis treści oraz siedemnastostronicowe wprowadzenie w tematykę, a także ze zbioru trzech publikacji naukowych współautorstwa doktoranta oraz krótkiego życiorysu. Całość (z wyjątkiem streszczenia w dwóch językach – polskim i angielskim) jest napisana w języku angielskim.

Istotność Badań

Opisane w rozprawie badania obejmują problematykę definiowania oraz mierzenia jakości w zespołach programistów tworzących oprogramowanie open-source. Jest to ważny kierunek związany z rosnącą popularnością zastosowań bazujących na komponentach open-source, co przekłada się także na wciąż powiększające się możliwości analizy danych dotyczących procesów tworzenia takiego oprogramowania, dostępnych między innymi na portalu GitHub. W istocie, autor rozprawy konsekwentnie odwołuje się w swoich pracach do badań empirycznych, formułując i weryfikując wszelkie hipotezy w odniesieniu do danych dostępnych w publicznej bazie GitHub Torrent. Opracowane miary mogą być pomocne dla oceny pracy zespołów programistów, na etapie doboru członków zespołu, jak również wspierają pewne znane w środowisku naukowców i praktyków obserwacje, np. dotyczące tak zwanych *zespołów chirurgicznych*. Rozprawa wymagała od autora przyswojenia rozległej wiedzy dziedzinowej dotyczącej zespołowego tworzenia oprogramowania open-source, umiejętnego zastosowania nowoczesnych metod sztucznej inteligencji i eksploracji dużych zbiorów danych, pewnych aspektów modelowania znanych z systemów wieloagentowych i sieci społecznościowych, a nawet rozwiązań znanych z systemów rekomendacyjnych. Wszystko to pozwala myśleć w przyszłości o zastosowaniu wiedzy zdobytej dzięki niniejszym badaniom również w innych problemach praktycznych, np. związanych z zarządzaniem dużymi organizacjami.

Zbiór Publikacji

Pośród trzech pub likacji wchod zących w skład rozprawy, można znaleźć dwie prace referowane na międzynarodowych konferencjach naukowych oraz jeden artykuł w uznanym międzynarodowym czasopiśmie naukowym (wszystkie te pozycje są widoczne na liście DBLP):

- **O. Jarczyk**, B. Gruszka, L. Bukowski, A. Wierzbicki: "On the Effectiveness of Emergent Task Allocation of Virtual Programmer Teams". Proceedings of the International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (WI-IAT), str. 369–376, 2014
- **O. Jarczyk**, B. Gruszka, S. Jaroszewicz, L. Bukowski, A. Wierzbicki: "Github Projects. Quality Analysis of Open-source Software". Proceedings of the 6th International Conference on Social Informatics, str. 80–94, 2014
- **O. Jarczyk**, S. Jaroszewicz, A. Wierzbicki, K. Pawlak, M. Jankowski-Lorek: "Surgical Teams on GitHub: Modeling Performance of GitHub Project Development Processes". Information and Software Technology 100, str. 32–46, 2018

Wszystkie te publikacje zostały napisane przy znaczącym wkładzie doktoranta jako ich pierwszego autora (choć wracam jeszcze do tego aspektu w dalszych uwagach i pytaniach do doktoranta).

Pierwsza z powyższych prac dotyczy dynamiki wirtualnych zespołów programistów, zarówno pod kątem rozwijania środowisk open-source, jak również przy realizacji projektów komercyjnych. Autorzy artykułu próbują zasymulować tę dynamikę bazując na danych dostępnych na portalu GitHub, biorąc pod uwagę fakt, że w przypadku czystej formuły open-source programiści samodzielnie decydują o przyłączeniu się do prac danego zespołu. Wykonane symulacje zdają się potwierdzać, że w takiej sytuacji ma sens strategia bieżącego przydziału zadań, a także że warto rozbudowywać mechanizmy wspierające proces zapraszania programistów o szczególnych zdolnościach na różnych etapach projektów.

W drugiej z powyższych prac, przeprowadzono badania nad zespołami zdecentralizowanymi, których członkowie porozumiewają się online. Również tutaj autorzy odnieśli się do portalu GitHub, traktując go jako przykład sieci społecznościowej. Dzięki zastosowaniu statystycznych technik analizy przeżywalności oraz regresji, opracowano i empirycznie zweryfikowano pewne miary jakości projektów open-source, jako potencjalnie skorelowane z cechami opisującymi członków grup programistycznych.

W trzeciej (i moim zdaniem najbardziej zaawansowanej) pracy, Pan Oskar Jarczyk wraz z współautorami kontynuuje powyższy kierunek poprzez rozwinięcie i zbadanie na przykładzie projektów zapisanych w repozytorium GitHub Torrent ciekawych miar charakteryzujących zespoły deweloperskie, odnoszących się do centralizacji działań, wewnętrznych procesów projektowych, jak również istnienia (potencjalnie niezależnych od samego projektu) powiązań społecznościowych pomiędzy programistami.

We wszystkich publikacjach daje się zauważyć głęboką wiedzę praktyczną autorów na temat procesów powstawania oprogramowania open-source i społecznych aspektów współzależności między członkami zespołów programistycznych. Z drugiej zaś strony, prace te cechuje umiejętność doboru narzędzi analizy dużych i złożonych źródeł danych, poprzez techniki statystycznego uczenia, eksploracji danych (np. z zastosowaniem hierarchicznej analizy skupień), jak i ekstrakcji struktur sieci społecznościowych.

Część Wstępna

Elementem spinającym zebrane publikacje, będącym niejakim przewodnikiem po dokonaniach Pana Oskara Jarczyka, jest wspomniane już dziewiętnastostronicowe wprowadzenie w tematykę rozprawy. Część ta składa się z sekcji opisujących problematykę rozprawy, przegląd literatury światowej, jasne wytłumaczenie innowacyjności i znaczenia zaprezentowanych przez doktoranta wyników dla rozwoju dziedziny, jak i przedstawienie niektórych możliwych kierunków dalszych badań. Szczególnie przydatne okazało się tu w mojej ocenie podzielenie Sekcji 1.2 na dwie podsekcje, poświęcone badaniom nad modelowaniem jakości projektów open-source oraz rozwojem i podziałem pracy programistów.

Zgodnie z opisem zawartym w rozprawie, prace podzielono na dwie części:

1. **Modelowanie miar jakości zespołów twórców oprogramowania open-source.**
2. **Analiza możliwych metod efektywnego podziału prac pomiędzy programistów.**

Na początek wprowadzono metody modelowania jakości bazujące na charakterystyce tworzenia się oraz działania zespołów na portalu GitHub. Następnie zdefiniowano miary jakości, które można stosować do oceny tych zespołów. Głównym celem na tym etapie było wyjaśnienie, które czynniki wpływają na to, jak zespoły te radzą sobie z informacjami pochodzącymi od użytkowników tworzonego oprogramowania. Zaobserwowano przy tym empirycznie cechy zespołów przyczyniające się do wysokiej jakości wsparcia użytkowników. Właśnie w tej części rozprawy dochodzimy do wspomnianej już, ciekawej obserwacji dotyczącej *zespołów chirurgicznych*, czyli grup programistów, którzy są dobrze zorganizowani, pracują wydajnie, a tym samym zapewniają cenne informacje zwrotne dla użytkowników.

Drugi główny aspekt naukowej aktywności Pana Oskara Jarczyka wiąże się z empirycznym badaniem strategii przydzielania pracy w zespołach programistów open-source. Doktorant bazuje w tym zakresie w szczególności na eksperymentach symulacyjnych i testowaniu różnych konfiguracji parametrów strategii. W oparciu o rzeczywiste dane historyczne z portalu GitHub, udało się zweryfikować, jak szybko różne metodyki alokacji zadań programistycznych prowadzą do zakończenia pracy nad projektami open-source. Wybrane strategie przydzielania zadań zostały przy tym porównane z wynikami otrzymanymi przy zastosowaniu centralnego planowania oraz strategii sieciowych. Uważam, że ten drugi aspekt rozprawy jest opisany przez doktoranta nieco mniej kompletnie. – A szkoda, ponieważ jest to kierunek bardzo interesujący, szczególnie biorąc pod uwagę wspomniany w Sekcji 1.4 pomysł reprezentowania programistów jako agentów bazujących na wielokryterialnych funkcjach celu.

Konkludując tę część recenzji chciałbym zauważyć, że pomysły oraz wnioski wyciągane przez doktoranta mają charakter bardzo dojrzały, zaś ich sformułowanie wymagało złożonego procesu analizy danych, dostępnych w portalu GitHub w sposób tylko częściowo ustrukturalizowany.

Biorąc zatem pod uwagę aktualny światowy stan badań w rozpatrywanej przez Pana mgr inż. Oskara Jarczyka dziedzinie, uważam, iż postawione cele oraz otrzymane wyniki są – pod względem zarówno naukowym, jak i praktycznym – w pełni godne rozprawy doktorskiej.

Uwagi

Przede wszystkim należy wysoko ocenić jakość publikacji składających się na rozprawę (co obrazują między innymi statystyki widoczne na profilu doktoranta na portalu ResearchGate), jak również dobry układ i przejrzystość części wstępnej. Z rozprawy można się dowiedzieć wielu wartościowych rzeczy, wyrobić sobie przydatne intuicje dotyczące istotności wyników. Z pewnością miało tu znaczenie bardzo wszechstronne doświadczenie Pana Oskara Jarczyka, opisane po części w Sekcji 3.2 rozprawy.

Z drugiej strony, być może warto zastanowić się, czy trzy artykuły to wystarczający materiał na doktorat, szczególnie biorąc pod uwagę to, że wszystkie prace są przygotowane przez wielu autorów, zaś indywidualny wkład doktoranta nie jest dość jasno sprecyzowany. Na szczęście na obronę Pana Oskara Jarczyka wpływa tu dodatkowo fakt, że na liście DBLP widać jeszcze inne pozycje, w tym kolejną pracę w czasopiśmie oraz kolejne prace konferencyjne o tematyce pokrewnej z tematyką rozprawy.

Pytania

Recenzję uzupełniam pytaniami, które mogą być szczególnie ważne dla autora rozprawy:

1. Zgadzam się z autorem, iż wypracowane wnioski można stosować w analizie innych wirtualnych zespołów, których członkowie współpracują przy tworzeniu treści wysokiej jakości. Można tu wspomnieć o procesie edytowania Wikipedii, czy też – jak to już zostało nadmienione – współpracy w organizacjach globalnych. Powstaje przy tym pytanie, jak dokładnie można przenosić opisane tu doświadczenia na tak różne dziedziny praktyczne, charakteryzujące się co prawda pewnymi cechami wspólnymi, ale też bardzo od siebie odmienne?
2. Jak już wspomniałem w poprzedniej sekcji, doceniam jakość publikacji zebranych w ramach tej rozprawy, jak też fakt, iż doktorant jest we wszystkich trzech przypadkach pierwszym autorem. W dzisiejszych czasach, praca zespołowa i interdyscyplinarna jest niezwykle istotna. Jednakże pozostaje zasygnalizowane już wcześniej pytanie, jaki był dokładnie wkład doktoranta w badania opisane w tych publikacjach? Na jakich ich aspektach się koncentrował?
3. W przeprowadzanych badaniach – co ilustruje przykładowo akapit „Sample selection from clean data" w Sekcji 1.3 rozprawy – widać niezwykle dużą, wspomnianą już pracę z danymi, ale również istotność ustawień parametrów przyjmowanych podczas tej pracy. Pojawia się zatem pytanie, jakie metody doboru parametrów stosowano w eksperymentach?
4. W Sekcji 1.3 autor pisze również o wykorzystaniu algorytmów genetycznych, jednakże ani w samej rozprawie, ani też we wchodzących w skład rozprawy publikacjach, nie ma zbyt wielu szczegółów na ten temat. Jak konkretnie wyglądało zastosowanie algorytmów genetycznych w badaniach związanych z problematyką niniejszej rozprawy doktorskiej?

Podsumowanie

Uważam, że opiniowana rozprawa doktorska Pana mgr inż. Oskara Jarczyka spełnia wymagania stawiane rozprawom doktorskim przez obowiązującą ustawę o stopniach i tytułach naukowych. Wnoszę więc o dopuszczenie jej do publicznej obrony.

dr hab. Dominik Ślęzak