RADA NAUKOWA DYSCYPLINY

INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

zaprasza na

OBRONĘ ROZPRAWY DOKTORSKIEJ

mgr. inż. Bartosza Wojciecha DOBRZYŃSKIEGO

która odbędzie się w dniu **25 września 2023 roku,** o godzinie **10:00** w trybie zdalnym

Temat rozprawy:

„Wielowymiarowa eksploracja repozytoriów programowych w zakresie raportów zgłoszeń oraz ich obsługi”

Promotor: prof. dr hab. inż. Janusz Sosnowski – Politechnika Warszawska

Recenzenci: dr hab. inż. Stanisław Jarząbek– Politechnika Białostocka

prof. dr hab. inż. Henryk Krawczyk – Politechnika Gdańska

dr hab. inż. Aneta Poniszewska-Marańda – Politechnika Łódzka

Obrona odbędzie się zdalnie na platformie MS Teams. Osoby zainteresowane uczestnictwem w obronie w formie zdalnej proszone są o zgłoszenie chęci uczestnictwa w formie elektronicznej na adres sekretarza komisji dr. hab. inż. Krzysztofa Cabaja, email : krzysztof.cabaj@pw.edu.pl do dnia 22.09.2023 r., godz.18:00.

Z rozprawą doktorską i recenzjami można zapoznać się w Czytelni Biblioteki Głównej Politechniki Warszawskiej, Warszawa, Plac Politechniki 1.

Streszczenie rozprawy doktorskiej i recenzje są zamieszczone na stronie internetowej: https://www.bip.pw.edu.pl/Postepowania-w-sprawie-nadania-stopnia-naukowego/Doktoraty/Wszczete-po-30-kwietnia-2019-r/Rada-Naukowa-Dyscypliny-Informatyka-Techniczna-i-Telekomunikacja/mgr-inz.-Bartosz-Wojciech-Dobrzynski

Przewodniczący Rady Naukowej Dyscypliny
Informatyka Techniczna i Telekomunikacja
Politechniki Warszawskiej
dr hab. inż. Jarosław Arabas, prof. uczelni

Streszczenie

Praca poświęcona jest badaniom nad wielowymiarową analizą i eksploracją procesu obsługi zgłoszeń rejestrowanych w repozytoriach programowych (np. *JIRA, Bugzilla*). Zgłoszenia te dotyczą m.in. błędów, nowych funkcjonalności, poprawy wydajności. Analiza dostępnej literatury oraz praktyczne doświadczenia autora pokazały potrzebę opracowania bardziej zaawansowanych i szczegółowych modeli eksploracji zawartości repozytoriów zaadaptowanych do specyfiki tych danych. Badania przedstawione w rozprawie dzielą się na dwa obszary: i) eksploracja zawartości informacyjnej programowych repozytoriów zgłoszeń, ii) wielowymiarowa analiza obsługi zgłoszeń.

W pierwszym obszarze prac autor zdefiniował szczegółowe profile dotyczące m.in.: statystyk dostępnych danych, aktywności aktorów i ich zaangażowania, korelacji pomiędzy atrybutami zgłoszenia i modyfikacjami kodu. Istotnym aspektem analiz jest eksploracja opisów zgłoszeń, która różni się od standardowych technik *text miningu.* Opracowane zostały wyrażenia regularne wspomagające analizę słownikową opisów oraz przetwarzanie wstępne tekstów zawartych w repozytoriach. Zostały one wykorzystane w autorskim algorytmie klasyfikacji raportów zgłoszeń. Przebadano wpływ konfiguracji danych na dokładność klasyfikacji.

Drugi obszar prac skupia się na wielowymiarowej analizie procesu obsługi zgłoszeń. Szczegółowe badania prowadzone są z użyciem autorskiej koncepcji grafowego modelu obsługi zgłoszeń IHG (ang. *Issue Handling Graph*). Zaproponowane profile wydajnościowe, czasowe, strukturalne stanów i ścieżek umożliwiają ocenę efektywności procesów obsługi zgłoszeń. Rozpatrzono różne typy, zgłoszeń oraz perspektywy obserwacji. Autor opracował oryginalne algorytmy wyszukiwania anomalii procesu obsługi zgłoszeń.

Przedstawiona metodyka analizy została zweryfikowana dla reprezentatywnych projektów *open source* oraz projektu komercyjnego. W pracy potwierdzona została użyteczność opracowanych wielowymiarowych analiz. Pozwalają one na wskazanie potencjalnych kierunków optymalizacji procesu raportowania i obsługi zgłoszeń oraz umożliwiają przeprowadzenie jego oceny w różnych momentach cyklu życia oprogramowania.

Słowa kluczowe: inżynieria oprogramowania, repozytoria programowe, monitorowanie procesu obsługi zgłoszeń, eksploracja danych.

Prof. Henryk Krawczyk
Politechnika Gdańska
e-mail: hkrawk@pg.edu.pl

Gdańsk, 21.07.2023

Recenzja rozprawy doktorskiej

mgra inż. Bartosza Wojciecha Dobrzyńskiego
zatytułowanej:
Wielowymiarowa eksploracja repozytoriów programowych w zakresie raportów zgłoszeń oraz ich obsługi
Promotor prof. dr hab. inż. Janusz Sosnowski

1. Problem badawczy i jego znaczenie

Tematyka rozprawy mgra inż. Bartosza Wojciecha Dobrzyńskiego dotyczy wielowymiarowej analizy i eksploracji danych rejestrowanych w repozytoriach programowych dostarczanych przez narzędzia klasy ITS (Issue Tracking Systems) dostępne w takich platformach jak JIRA, GitHub, czy Bugzilla lub Redmine. Narzędzia tego typu wspierają cały cykl życia projektu, więc mogą dotyczyć różnych aspektów inżynierii oprogramowania od tworzenia serwisu internetowego i raportowania błędów poprzez wspomaganie zarządzania projektami i nadzorowanie pracy zespołów aż do monitoringu procesów w całych organizacjach. Takie narzędzia są użyteczne i wręcz konieczne w czasach szybko zmieniającej się rzeczywistości biznesowej i błyskawicznego dostępu do informacji, które wymuszają konieczność podejmowania szybkich decyzji. Dlatego ich znaczenie stale wzrasta i stają się jednym z głównych elementów procesu wytwarzania.

Recenzowana rozprawa koncentruje się na określeniu aktywności projektantów oraz ich zaangażowaniu w proces wytwarzania, a także na wyznaczaniu korelacji pomiędzy atrybutami zgłoszenia i modyfikacjami tworzonego kodu systemu czy aplikacji. Doktorant rozpatruje statystyki charakterystyczne dla dostarczanych zgłoszeń oraz zajmuje się wielowymiarową analizą procesu obsługi zgłoszeń w celu zapewnienia wiarygodnego wyszukiwania możliwych anomalii. Poza tym stara się wskazać potencjalne kierunki optymalizacji procesu raportowania i obsługi zgłoszeń, by zminimalizować nakłady przeprowadzenia takiej oceny w wybranych etapach życia wytwarzanej aplikacji. Tego typu działania mają istotne znaczenie dla zapewnienia niezawodnego i ekonomicznego projektowania i testowania systemów i aplikacji informatycznych.

Po krótkim wprowadzeniu do analizowanej problematyki Doktorant formułuje tezę rozprawy doktorskiej oraz przedstawia konieczne zadania badawcze, których wyniki powinny potwierdzić jej zasadność. Podjęte przez Niego zadania dotyczą opisu i eksploracji zawartości informacyjnej repozytoriów zgłoszeń, zaproponowanie modelu grafowej obsługi zgłoszeń (IHG – Issue Handling Graph) oraz metod analizy zarejestrowanych w takich repozytoriach wielkości parametrów i komunikatów (tekstów). Ekstrakcja gromadzonych danych oraz przeprowadzone analizy dla tych danych wspierane są przez zbudowane przez Autora rozprawy narzędzie IssueAnalyzerTool oraz skrypty pomocnicze, które wywołują opracowane przez Autora rozprawy algorytmy takich analiz. W rozprawie wykorzystano też nowe modele klasyfikacji oraz analizy semantycznej, które wymagały odpowiedniej transformacji opisów tekstowych zawartych w zgłoszeniach.

Sposób przeprowadzania badań jak i zastosowana metoda analizy zgromadzonych danych w repozytoriach nie budzą zastrzeżeń. W rozprawie doktorskiej wykorzystano odpowiedni aparat matematyczny związany z eksploracją danych, analizą tekstów, teorią grafów oraz budową algorytmów. Nie mam więc wątpliwości, że recenzowana rozprawa ma charakter naukowy, a jej wyniki są wartościowe, bo przede wszystkim odnoszą się do praktycznych zastosowań. Należy też podkreślić, że zakres tematyczny rozprawy mieści się w dyscyplinie Informatyka Techniczna i Telekomunikacja.

2. Osiągniecia badawcze

W procesie wytwarzania oprogramowania istotną rolę odgrywa informacja o jego przebiegu oraz o postępie w realizacji kolejnych zadań. Źródłem takich informacji mogą być różnego typu zgłoszenia generowane przez członków zespołów projektowych (reporterów) i umieszczane w repozytoriach zgłoszeń. Format i treść tych zgłoszeń może być różnorodne w zależności od śledzonych parametrów procesu wytwarzania oraz od typu zgłaszanych sytuacji, a nawet od wykorzystywanych narzędzi wspomagających tego typu czynności. Tego typu zgłoszenia są przechowywane w tzw. repozytoriach zgłoszeń i wymagają odpowiedniej reakcji zespołu projektowego. Podejmowane reakcje również są rejestrowane i jest tworzona historia takich wszystkich działań. W ten sposób są gromadzone interesujące dane, które mogą być wykorzystane do analizy procesu wytwarzania aplikacji informatycznej. Z tego punktu widzenia istotna jest zakres rejestrowanych informacji jak i sam proces obsługi zgłoszeń, by wyłuskać i przeanalizować istotne dane mające wpływ na przebieg realizacji projektu. Rozważania Doktoranta dotyczyły właśnie analizy wybranych danych zawartych w repozytoriach zgłoszeń, jak też oceny jakości metody obsługi tych zgłoszeń. Wyjściem do Jego rozważań były praktyczne systemy wspomagające tego typu działania, a także ocena ich możliwości, a w dalszej perspektywie rozszerzenie ich funkcjonalności w celu gromadzenia bardziej adekwatnych danych i skutecznej ich analizy dotyczącej jakości procesu wytwarzania aplikacji i systemów informatycznych.

Do osiągnięć Doktoranta w obszarze analizy danych zawartych w repozytoriach oraz oceny metod obsługi zgłoszeń należy zaliczyć:

1. Opracowanie profili statystycznych zawartości repozytoriów mogących ujawniać niedoskonałości raportowania projektu, a także profili aktywności uczestników (aktorów) projektu tzn. wyróżnić członków zespołu o mniejszym zaangażowaniu w projekcie, bądź jak najbardziej aktywnych jego członków.
2. Przebadanie wpływu eksploracji opisów oraz komentarzy zgłoszeń, a w szczególności różnych konfiguracji klasyfikatorów oraz cech danych wejściowych na dokładność dokonywanej klasyfikacji (algorytm klasyfikacji - Alg. 4), co stanowi wyznacznik jakości opisów zgłoszeń w repozytorium.
3. Rozwinięcie metodyki wielowymiarowej analizy obsługi zgłoszeń umożliwiającej przeprowadzenie dokładnej ewaluacji procesu obsługi zgłoszeń w różnych perspektywach, np. ogólnej – uwzględniającej wszystkie typy zgłoszeń lub szczegółowej – koncentrując się na konkretach, takich jak typ zgłoszeń, grupy reporterów czy priorytet zgłoszeń. Proponowana metodyka postępowania bazuje na wprowadzonym grafowym modelu IHG oraz zestawie oryginalnych algorytmów, takich jak budowanie tego grafu (Alg. 7), filtracja wierzchołków i krawędzi grafu (Alg. 8, Alg. 9), a także generowania pliku stanów unikalnych ścieżek oraz wyszukiwania pętli (Alg.10 - Alg. 12). Dzięki tym algorytmom jest możliwe wyliczanie profili wydajnościowych, czasowych i strukturalnych, a także wyznaczenie możliwych anomalii. Pozwala to również na przeprowadzenie porównań, np. pomiędzy procesami obsługi dla

różnych priorytetów zgłoszeń czy różnych typów zgłoszeń, a nawet fragmentów realizowanych projektów.

4. Implementacja opracowanych algorytmów jako narzędzia IssueAnalyzerTool wykorzystanego do analizy wybranych sytuacji pojawiających się w procesie wytwarzania oprogramowania. Pozyskana w ten sposób wiedza może też zostać wykorzystana do optymalizacji prac związanych z obsługą zgłoszeń oraz poprawą konfiguracji systemu raportowania zgłoszeń (ITS). W szczególności umożliwia określanie „wąskich gardeł, zidentyfikowanie trudno dostępnych stanów oraz przejść między nimi, jak również wskazanie stanów o wysokim czasie obsługi, ścieżek z wieloma stanami i dużą liczbą komentarzy czy pętli stanów ścieżek. Co więcej świadomość tych anomalii daje szansę, w uzgodnieniu z całym zespołem projektowym, na podjęcie działań zaradczych minimalizujących wpływ takich sytuacji na przebieg procesu wytwarzania.

Są to istotne osiągnięcia badawcze Doktoranta, zrealizowane na bazie dostępnych praktycznych danych eksperymentalnych. Dane te pochodzą z wybranych projektów, których realizacja była raportowana przy wykorzystaniu narzędzi klasy ITS – JIRA. Dzięki tym narzędziom wyniki procesu gromadzenia danych udokumentowane zostały w odpowiednich repozytoriach. Jednak dostęp do nich jak i ich interpretacja nie jest łatwa. Zależy od kontekstu ich rejestracji (typu czy miejsca zgłoszenia), a także od wykorzystania indywidualnego żargonu, czy nieznanych skrótów w polach tekstowych, a także popełniania typowych ludzkich błędów. Poza tym istnieją też techniczne (wydajnościowe czy funkcjonalne) ograniczenia wykorzystywanych narzędzi co ogranicza zakres gromadzonych danych. Zatem konieczne było krytyczne spojrzenie Doktoranta na sposób gromadzenia takich danych oraz interpretację ich znaczenia, jak też sprawdzenie ich przydatności w całościowej i szczegółowej ocenie realizowanych projektów.

Doktorant przeanalizował zagregowane dane literaturowe dotyczące projektów open source, takich jak: Apache Casandra (cas), Apache Spark (spark), Apache Flink (flink), Apache Ignite (ignite), Mozilla Tunderbird (moz), Red Hat Enterprise Linux 8 (red), a także projektu P1, który jest projektem komercyjnym zrealizowanym z istotnym udziałem Doktoranta. Porównywał rozkłady wartości kategorii i typów atrybutów zgłoszeń, analizował korelacje między atrybutami i typami zgłoszeń, rozpatrywał zapewnienie optymalnej konfiguracji narzędzia, dokonał analizy aktywności użytkowników repozytoriów oraz ewaluację procesu obsługi zgłoszeń, w tym ocenę długu błędów (błędów istotnych, ale odłożonych do późniejszej analizy). Różne spostrzeżenia odniósł do konkretnych projektów co jest istotną wartością badawczą z punktu widzenia inżynierii oprogramowania.

Przedstawione wyniki tego typu analiz pośrednio odnoszą się i potwierdzają tezę rozprawy doktorskiej zdefiniowanej następująco: *Analiza i ocena procesu wytwarzania oprogramowania wymaga opracowania reprezentatywnych modeli oraz metod eksploracji danych z repozytoriów zgłoszeń i repozytoriów kodu. Uwzględnienie różnych poziomów ekstrakcji i agregacji informacji oraz perspektyw obserwacji, poszerza zakres przedmiotowy monitorowania projektu i ułatwia identyfikację niedoskonałości.* Pewne uwagi krytyczne dotyczące zakresu stosowalności tej tezy zostały podane w następnym rozdziale.

3. Uwagi krytyczne

Zauważalnym mankamentem rozprawy doktorskiej jest ograniczenie się jej Autora do wykorzystania możliwości istniejących narzędzi klasy ITS, bez próby zdefiniowania uogólnionego modelu procesu tworzenia, rejestracji i obsługi zgłoszeń związanych z problemami realizacji typowego projektu informatycznego. To umożliwiłoby porównanie zakresu badań zawartych w rozprawie z

niezbędnym zakresem prac możliwych do wykonania. Cenne natomiast jest wykorzystanie modelu grafowego, umożliwiającego takie uogólnione podejście w przypadku analizy obsługi zgłoszeń.

Poza tym w rozprawie zdawkowo potraktowano opis analizowanych projektów jak i szczegóły implementacyjne narzędzia przeznaczonego do tworzenia repozytorium danych. Szersza charakterystyka analizowanych projektów oraz podanie szczegółów implementacyjnych proponowanego narzędzia wskazywałyby na możliwości dalszego rozwoju repozytorium zgłoszeń oraz dokładną ocenę możliwości i kompletności zastosowanej metody analizy.

Dla większej czytelności rozprawy doktorskiej przydałby się też podanie koncepcji i założeń dotyczących badań eksperymentalnych oraz metod analizy uzyskanych wyników. Bez próby dokonania takich uogólnień Doktorant w zasadzie koncentruje się na konkretnej analizie wybranych parametrów (ale czy najważniejszych?) opisujących niektóre aspekty procesu wytwórczego. Do takiej analizy zorientowanej na pojedyncze projekty, dostosowuje metody oceny bez uwzględnienia innych aspektów inżynierii oprogramowania również istotnych z punktu rejestracji nieodpowiednich zdarzeń czy podejrzanych sytuacji. Co prawda przy zaprezentowaniu algorytmów Doktorant skoncentrował się na konkretnej i trafnie dobranej strukturze danych, ale taka struktura może nie obejmować wszystkich możliwych przypadków. Konieczne byłoby więc potwierdzenie istnienia lub choćby zasugerowanie potrzeby opracowania standardu opisu zgłoszeń i danych zawartych w repozytoriach dotyczących procesu śledzenia i raportowania realizowanych projektów, co w przyszłości ułatwiłoby szerszą ich analizę. Z drugiej jednak strony zakres gromadzenia różnego typu zgłoszeń może być bardzo obszerny, podobnie różnorodność wymaganych metod do analizy wyników zawartych w repozytorium zgłoszeń, co wymagałoby znacznego wysiłku, przewyższającego zrealizowanie jednej rozprawy doktorskiej. Jednak brak dyskusji na temat kompletności repozytorium danych oraz oceny zakresu funkcjonalności systemów wspomagania procesów śledzenia czynności projektowych jest pewnym mankamentem tej rozprawy. Poza tym zauważa się brak w rozprawie wykazu najważniejszych oznaczeń wraz z krótkim ich opisem, co także zmniejsza czytelność recenzowanej rozprawy.

Jak wspomniano powyżej, czego Doktorant też jest świadomy, w praktyce inżynierii oprogramowania może pojawić się znacznie więcej różnego typu sytuacji istotnych dla przebiegu procesu wytwarzania aplikacji informatycznej niż te które uwzględniono w ocenianej rozprawie doktorskiej. Przykładem mogą być choćby takie przypadki jak: niskie kompetencje uczestników zespołów projektowych, ograniczone możliwości dostępnych narzędzi projektowania bądź nieadekwatny wybór metod wytwarzania czy zarządzania procesem projektowania, a także pojawiające się niesprzyjające okoliczności utrudniające działania. Mają one również istotny wpływ na przebieg i efekty realizowanego projektu. Jeśli więc zakres monitorowania występujących sytuacji może być znacznie szerszy niż ten, który został przedstawiony w ocenianej rozprawie doktorskiej, to ogólnie sformułowana teza rozprawy wydaje się być zbyt daleko idąca, powinna jedynie odnosić się do przypadków rozpatrzonych w tej rozprawie?

Co więcej w rozprawie nie podano jakie metodyki wytwarzania były wykorzystywane przy realizacji i analizie projektów informatycznych. Wspomniano jedynie o metodyce SCRUM, dla której zaprezentowane metody analizy repozytorium zgłoszeń dobrze sprawdziły się przy unikalnej ocenie efektywności sprintów. Interesujące byłoby odniesienie się do innych metodyk projektowania wykorzystujących również metody śledzenia i raportowania procesu wytwarzania aplikacji, zwłaszcza do takich jak DevOps czy ostatnio bardzo popularnej metody DevOps+AI, które to najczęściej są stosowane w środowiskach chmury obliczeniowej.

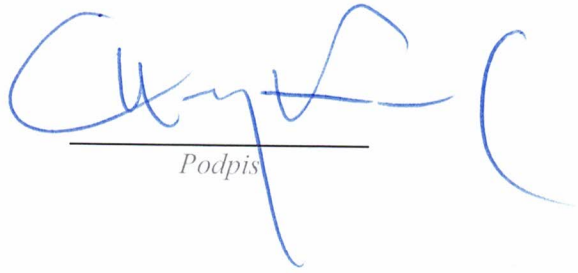
Na zakończenie opisu uwag krytycznych chciałbym jednak podkreślić, że mają one charakter dyskusyjny i nie przekreślają wartości i znaczenia badań i wyników szczegółowych zawartych w recenzowanej rozprawie doktorskiej.

4. Podsumowanie

Doktorant wykazał się dużymi umiejętnościami budowy i wykorzystania narzędzi śledzenia i raportowania procesów wytwarzania projektów informatycznych. W kilku rozdziałach swojej rozprawy odniósł się też do teoretycznych i praktycznych osiągnięć światowych dotyczących rozpatrywanej problematyki. To potwierdza dobry stan wiedzy Doktoranta w tym zakresie. Na podkreślenie zasługuje również właściwie dobrana literatura – 108 pozycji, w tym 4 z nich zostały przygotowane przez Doktoranta (współautorstwo, pozycje 10, 11, 14, 40). Najistotniejsza jego publikacja: *Analysing problem handling schemes in software projects. Information and Software Technology,* 2017.
Reasumując stwierdzam, co następuje:

1. Doktorant rozpatrzył bardzo istotny, ale zarazem bardzo trudny i złożony problem naukowy jakim jest usprawnienie śledzenia i raportowania procesu realizacji projektów informatycznych,
2. Opracował oryginalne i przydatne algorytmy opisu i analizy zarejestrowanych danych liczbowych i testowych gromadzonych w tzw. raportach zgłoszeń,
3. Dokonał oceny kilku reprezentatywnych przypadków (user cases) na podstawie dostępnych (open data) i własnych (komercyjnych) danych oraz zaproponował rozsądne usprawnienia procesu śledzenia i raportowania realizacji projektów informatycznych.

Biorąc powyższe pod uwagę oraz uwzględniając wymagania zdefiniowane przez odpowiednią Ustawę o stopniach naukowych i tytule naukowym, stwierdzam, że moja ocena rozprawy jest zdecydowanie pozytywna i proponuję dopuszczenie mgra inż. Bartosza Wojciecha Dobrzyńskiego do dalszych etapów przewodu doktorskiego.



Podpis

WYDZIAŁ INFORMATYKI
Politechniki Białostockiej
15-351 Białystok, ul. Wiejska 45A
tel. +48 85 746 90 50, fax +48 85 746 90 57
e-mail: wi.sekretariat@pb.edu.pl
www.wi.pb.edu.pl

Białystok 17 sierpnia 2023.

Prof. Ndzw. Stanisław Jarząbek
Politechnika Białostocka

Recenzja Rozprawy Doktorskiej mgr inż. Bartosza Dobrzyńskiego pt. : „Wielowymiarowa eksploracja repozytoriów programowych w zakresie raportów zgłoszeń oraz ich obsługi"

Uwagi ogólne

Systemy znane jako *Issue Tracjking Systems* (IST) takie jak JIRA czy Bugzilla umożliwiają śledzenie błędów i innych problemów pojawiających się w trakcie realizacji projektów programowych. W zakresie śledzenia błędów, IST gromadzą informacje o wykrytych błędach w repozytoriach w postaci zgłoszeń-raportów błędów, i udostępniają programistom narzędzia do ich analizy. Kluczowa rola tych systemów w zarządzaniu projektami polega na tym że duże systemy programowe rozwijane są przez zespoły programistów, a proces rozwoju programu może trwać miesiące albo lata. Błędy wykryte przez programistkę często nie mogą być przez nią poprawiane zaraz po ich wykryciu, choćby z uwagi na fakt że ktoś inny był odpowiedzialny za obszary programu w których błąd zaistniał. Aby umożliwić poprawę błędu w późniejszej fazie rozwoju programu i przez inną programistkę, niezbędne jest systematyczne udokumentowanie błędu w celu jego replikacji, i wyjaśnienia jego zaobserwowanych powiązań funkcjonalnościami programu i z innymi błędami. Efektywność naprawy błędów w dużym stopniu zależy od jakości zgłoszeń błędów, i jakości IST.

Z uwagi na powyższe, systemy śledzenia błędów stały się nieodzownym elementem rozwoju programów i zarządzania programowymi projektami.

Skupiając się na temacie skuteczności raportowania błędów, rozprawa atakuje problem ważny dla praktyki programowania, z wieloma obszarami możliwych ulepszeń i stawiający przed badaczami trudne wyzwania. Od lat pracując z systemami śledzenia błędów, autor poczynił wiele krytycznych obserwacji dających mu znakomity start do podjęcia tego tematu w swoim doktoracie.

Kierując się tymi doświadczeniami i na podstawie analizy literatury, autor zidentyfikował potrzebę opracowania bardziej zaawansowanych i szczegółowych modeli eksploracji repozytoriów zgłoszeń błędów, uwzględniając ich specyfikę. W celu sformułowania ulepszonych rozwiązań, autor badał zawartość repozytoriów zgłoszeń i dokonał wielowymiarowej ich analizy.

Autor zaproponował nowatorskie metody analizy korelacji pomiędzy atrybutami zgłoszenia i modyfikacjami kodu prowadzące do oryginalnych algorytmów klasyfikacji zgłoszeń, i wyszukiwania anomalii procesu obsługi zgłoszeń.

W części eksperymentalnej, autor zweryfikował zaproponowane metody dla wybranych projektów *open source* i komercjalnego projektu. Eksperymenty wykazały że zaproponowane metody wskazują kierunki istotnych zmian procesu zgłoszeń prowadzących do bardziej efektywnej ich obsługi.

Zaproponowane metody są oryginalnego pomysłu autora i prowadzą do ważnych usprawnień praktyki programowania.

Analiza istniejącego stanu wiedzy i teza pracy

Dysertacja dotyczy systemów zarządzania zgłoszeniami projektowymi (błędów czy wprowadzania nowych funkcjonalności) takich jak JIRA, Bugzilla, GitHub Issues, Redmine, OTRS czy MantisBT, znanych ogólnie jako *Issue Tracking Systems* (IST). Systemy te umożliwiają zarządzeni zgłoszeniami, monitorowanie postępów prac poprzez zmianę statusów oraz kooperację i komunikację między aktorami biorącymi udział w pracach projektowych.

W oparciu o własne doświadczenia autora poszerzone o dogłębną analizę literatury, autor stwierdził, że zgłoszenia w repozytoriach projektowych są w dużej mierze opisowe, i słabo zrestrukturalizowane. Ten brak formalizacji utrudnia ich badanie tradycyjnymi metodami eksploracji danych i analiz tekstowych. Autor wskazuje literaturę sygnalizującą te problemy.

Większość badań do analizy ścieżek obsługi zgłoszeń wykorzystuje system analizy grafów *Problem Handling Graphs* (PHG). Autor przytacza bogata literatura na ten temat. W dysertacji, autor zaproponował istotne rozszerzenia modelu IHG prowadzące do lepszych wyników analizy zgłoszeń.

Autor omawia w dysertacji szereg innych gałęzi badań mających powiązanie z jego głównym tematem pracy. Dotyczą one modeli wzrostu niezawodności oprogramowania, *Software Reliability Growth models*, ewaluacji i usprawnieniu procesów wytwarzania i monitorowania oprogramowania, analizie ocen wartości informacyjnych dodawanych zgłoszeń w repozytoriach błędów, i prace analizujące wyniki ankiet, w których zespoły projektowe oceniały wartości informacji zawartych w zgłoszeniach błędów. Równie szczegółowo omawia autor specyficzne techniki *text mining*, używane w analizie zgłoszeń i ich obsługi. Swoje wnioski autor opiera o bogatą literaturę.

Większość opublikowanych metod analizy zgłoszeń pomija – na co słusznie wskazuje autor - wstępną analizę danych tekstowych lub przeprowadza ją na ogólnym poziomie, z brakiem odniesienia do konkretnych projektów. Techniki opisane w literaturze mogą być użyte głównie do predykcji przypisania odpowiedniego zgłoszenia w danym projekcie. Nie są jednak odpowiednio zaadresowane trudności w zastosowaniu tych technik w innych projektach.

Teza pracy: Autor formułuje tezą swojej pracy następująco: „Analiza i ocena procesu wytwarzania oprogramowania wymaga opracowania reprezentatywnych modeli oraz metod eksploracji danych z repozytoriów zgłoszeń i repozytoriów kodu. Uwzględnienie różnych poziomów ekstrakcji i agregacji informacji oraz perspektyw obserwacji, poszerza zakres przedmiotowy monitorowania projektu i ułatwia identyfikację niedoskonałości.".

Precyzując, tezą pracy jest że wielostronna analiza danych obejmująca wiele różnych repozytoriów zgłoszeń wraz ze zwiększonym zakresem badanych projektów możliwe będzie zrozumienie różnic pomiędzy repozytoriami projektowymi, obejmujących także zróżnicowany styl zgłoszeń. Na bazie tego zrozumienia, techniki zaproponowane przez autora, głównie model IHG (rozszerzanie modelu PHG) umożliwią analizę repozytoriów zgłoszeń prowadzącą do poprawienia efektywności procesów raportowania i obsługi zgłoszeń.

Praca obejmuje sformułowanie technik/algorytmów wspomagających analizę repozytoriów dla poparcia tezy, i eksperymentalna weryfikacje całości proponowanej metody w oparciu o *open source* i komercjalne projekty.

Autor dokonał wyczerpującej analizy istniejących rozwiązań skrótowo we Wprowadzeniu, Sekcja 1.2. w sposób rozszerzony w Sekcjach 3.1 i 4.1.

Jakość zaproponowanych rozwiązań i ich opisu

Potrzeba ulepszonych metod raportowania zgłoszeń i ich obsługi jest dobrze umotywowana, zarówno na bazie doświadczeń autora jak i problemów dyskutowanych w literaturze.

Zarówno całościowe podejście autora do tematu, jak i konkretne techniki i algorytmy jasno wynikają z tych potrzeb i otwierają drogę do efektywnego raportowania zgłoszeń i ich obsługi. Dotyczy to w szczególności:

- Wielowymiarowej analizy różnorodnych repozytoriów w celu zdefiniowania metod pasujących do ich zmiennej specyfiki
- Modeli danych i schematów eksploracji, wspartych metrykami ukierunkowanymi na ekstrakcje cech charakterystycznych (składniowych, semantycznych, czasowych i statystycznych) repozytoriów śledzenia problemów, odniesionych do różnych perspektyw obserwacji,
- Śledzenia efektywności obsługi zgłoszonych problemów, uwzględniając szeroki zakres zależności względem opracowanego modelu stanowego grafu (IHG),
- Wykrywania niedoskonałości repozytoriów programowych oraz procesu obsługi zgłoszeń.
- Badania profili obsługi różnych typów zgłoszeń
- Detekcji anomalii oraz analizy jakości repozytoriów programowych

W badaniach autor zaadaptował i rozszerzył szereg statystyk i algorytmów eksploracji tekstu, uwzględniając specyfikę repozytoriów programowych. Badając efektywność procesu obsługi zgłoszeń, autor posłużył się autorskim modelem grafów (IHG), ułatwiającym śledzenie przepływów obsługi zgłoszeń, aktywności aktorów, schematów obsługi różnych kategorii zgłoszeń etc.

Zarówno ogólne podejście autora do tematu jak i poszczególne techniki/algorytmy zostały klarownie opisane.

Praktyka jest najważniejszym sprawdzianem dla nowatorskich rozwiązań inżynierii oprogramowania. Obszerna weryfikacja metod zaproponowanych przez autora potwierdza ich wysoka przydatność.

Eksperymentalna ewaluacja zaproponowanych rozwiązań

Autor uzasadnił proponowane metody na gruncie teorii w sposób jasny i przekonywujący. Ostatecznym testem dla metod jest zawsze praktyka. Autor dokonał obszernej eksperymentacji przy użyciu serwerów w celu ewaluacji proponowanych metod. Ewaluacja została wykonana poprawnie. Skala eksperymentów jest odpowiednio dostosowana do skali problemu, tak aby jej wyniki bez zastrzeżeń uznać za poprawne.

Oryginalność i użyteczność zaproponowanych rozwiązań

W badaniach autor zaadaptował i rozszerzył szereg statystyk i algorytmów eksploracji tekstu, uwzględniając specyfikę repozytoriów programowych, co znacząco odróżnia je od klasycznych metod eksploracji danych.

Rozprawa doktorska przedstawiona przez autora pokrywa szerszy zakres analizowanych danych niż to miało miejsce w poprzednich rozwiązaniach opublikowanych zarówno przez

autora jak i innych badaczy. Co więcej, niektóre z proponowanych metod nie ograniczają się jedynie do problemu adresowanego w dysertacji, ale zostały sformułowane tak aby umożliwić ich wykorzystanie w innych kontekstach. Na przykład, algorytmy przetwarzania tekstu i identyfikacji słowników mogą być pomocne w wyszukiwaniu podobieństw lub w klasteryzacji zgłoszeń. W odróżnieniu od poprzednich rozwiązań, które były wypracowane dla konkretnych dzienników, metody zaproponowane przez autora rokują wypracowanie wyspecjalizowanych algorytmów dopasowanych do szczególnych typów dzienników na bazie proponowanych rozwiązań.

Autor opublikował wyniki badań w prestiżowych wydawnictwach i wygłosił komunikaty o nich na międzynarodowych konferencjach. Na szczególne wyróżnienie zasługują publikacje 45, 47-49 (odnośniki do spisu Literatury na końcu rozprawy).

Metody wypracowane przez autora otwierają możliwości dalszych badań. Myślę, że publikacje autora które dokumentują te metody będą szeroko cytowane. Ponadto pewne fragmenty (np. dotyczące opracowanych algorytmów) zasługują na kolejne publikacje.

Słabe strony pracy

1) Nie jest dla mnie w pełni jasne czy problem eksploracji zawartości repozytoriów zgłoszeń był rozpatrywany w pracy magisterskiej, ewentualnie w jakim zakresie. Podobnie słabo wypunktowano różnice wcześniejszego modelu PHG i opisywanego w pracy doktorskiej modelu IHG.
2) Nowatorstwo podejścia i specyficznych metod jest omówiona w pracy, w opisie metod i sekcjach *Dyskusja* kończących główne Rozdziały dysertacji. Myślę, że właściwe byłoby omówienie nowatorstwa rozwiązań w osobnym Rozdziale *Nowatorstwo i wkład przedstawionych rozwiązań*, w sposób bardziej skondensowany niż jest to omówione w sekcjach Dyskusja i w końcowym Rozdziale Podsumowanie. W tym samym Rozdziale przedstawiłbym sumaryczne porównanie zalet przedstawionych rozwiązań w porównaniu z rozwiązaniami wcześniej proponowanymi w literaturze i ich ewentualnych ograniczeń. Autor mógłby tu również omówić sposób w jaki jego praca otwiera nowe kierunki badań nad zgłoszeniami.
3) W motywacji (str 10), warto jest wskazać słabości PHG jako metody analizy ścieżek obsługi problemów które doprowadziły do rozszerzeń tego modelu zaproponowanych przez autora.
4) Rozdział 2 Teza i cel pracy: W obecnej formie, rozdział zawiera sporo materiału który tu nie pasuje i powinien znaleźć się gdzie indziej. Konkretnie, tekst począwszy od „Sformułowanie problemu badawczego ..." na stronie 14 powinien być przeniesiony do wprowadzenia i/lub motywacji pracy. Omówienie struktury pracy (str. 17) powinno znaleźć się we wprowadzeniu. Natomiast w Rozdziale 2 dodałbym parę uwag na temat nowatorskich technik przy pomocy których autor zamierza osiągnąć cel pracy. Proponuje tytuł: *Teza, cel i zakres pracy.*
5) Czytając prace często miałem wrażenie déjà vu, że pewne sekwencje materiału już się wcześniej pojawiły. Przykładem są tu Sekcje 1.2 i 3.1. Myślę, że praca by zyskała na klarowności gdyby materiał był lepiej zorganizowany i przedstawiony bez zbędnych powtórzeń.
6) Str. 10: Poprawić niejasne sformułowanie: „Zgłaszane problemy często wymagają analizy przed rozwiązaniem, jednak ostatecznie może okazać się, że nie są to *prawdziwe błędy*" – nie jest jasne co autor rozumie przez *prawdziwe błędy* w kontekście następnych zdań.

Praca napisana jest dobrą polszczyzną i starannie edytowana, niemniej jednak uwzględnienie przedstawionych wyżej sugestii mogłoby zwiększyć jej czytelność.

Podsumowanie

Rozprawę charakteryzuje nowatorstwo proponowanych rozwiązań, wysoki potencjał praktycznych zastosowań i inspiracji dalszych badań nad wiarygodnością systemów programowych. Materiał jest dobrze zorganizowany i sama rozprawa napisana w sposób przystępny, z wystarczająca dozą detali i przykładów umożliwiających zrozumienie przedstawianych metod i ich krytyczną ocenę.

Biorąc powyższe pod uwagę oraz uwzględniając wymagania zdefiniowane przez odpowiednią Ustawę o stopniach i tytułach naukowych, stwierdzam, że moja ocena rozprawy jest zdecydowanie pozytywna i proponuję dopuszczenie magistra Bartosza Dobrzyńskiego do dalszych etapów przewodu doktorskiego.

X Stanisław Jarząbek

Stanislaw Jarzabek

Łódź, dnia 04.08.2023

dr hab. inż. Aneta Poniszewska-Marańda
Instytut Informatyki
Politechnika Łódzka

RECENZJA ROZPRAWY DOKTORSKIEJ DLA RADY NAUKOWEJ DYSCYPLINY „INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA" POLITECHNIKI WARSZAWSKIEJ

Tytuł rozprawy: Wielowymiarowa eksploracja repozytoriów programowych w zakresie raportów zgłoszeń oraz ich obsługi
Autor rozprawy: mgr inż. Bartosz Wojciech Dobrzyński
Promotor rozprawy: prof. dr hab. inż. Janusz Sosnowski

Recenzja sporządzona na podstawie pisma Przewodniczącego Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja, dr hab. inż. Jarosława Arabasa, prof. uczelni, z dnia 12 czerwca 2023 r.

Ocena układu rozprawy doktorskiej, zawartość

Rozprawa zawiera sześć rozdziałów, bibliografię oraz załączniki. Posiada 141 stron, w tym 10 stron załączników. W rozprawie umieszczono 108 pozycji bibliograficznych, dobrze dobranych i aktualnych.

Rozdział pierwszy jest wprowadzeniem do pracy. Przedstawia motywację, jaka przyświecała Autorowi podczas realizacji pracy doktorskiej i pisania rozprawy oraz kontekst badań.
Praca dotyczy zagadnień związanych z procesem rozwoju i utrzymania oprogramowania, w aspekcie monitorowania tego procesu w obszarze prowadzenia projektów informatycznych. Monitorowanie tego procesu skupia się na wykorzystaniu współczesnych repozytoriów kodu oraz obsłudze zgłoszeń, celem zapewnienia wsparcia procesu wytwarzania oprogramowania.
Kontekst badań jest osadzony we współczesnej literaturze, dotyczącej poruszanej tematyki. Autor w ogólny sposób odnosi się do wybranych pozycji literatury, jednak w wystarczający sposób jak na Wprowadzenie do pracy, prezentując ich główne założenia, zalety i wady.

Rozdział drugi prezentuje tezę pracy oraz sformułowanie problemu badawczego. Teza pracy została określona następująco:
„Analiza i ocena procesu wytwarzania oprogramowania wymaga opracowania reprezentatywnych modeli oraz metod eksploracji danych z repozytoriów zgłoszeń i repozytoriów kodu. Uwzględnienie różnych poziomów ekstrakcji i agregacji informacji oraz perspektyw obserwacji, poszerza zakres przedmiotowy monitorowania projektu i ułatwia identyfikację niedoskonałości".
Z tezą powiązano zadania badawcze, dotyczące trzech zagadnień, które stanowią zakres niniejszej rozprawy.

Rozdział trzeci pracy prezentuje metodykę analizy repozytoriów kodu. Opis metodyki poprzedzony jest analizą stanu wiedzy w badanej tematyce repozytoriów kodu. Następnie

prezentowana jest charakterystyka repozytoriów programowych, systemów zgłoszeń ITS (Issue Tracking Systems), w szczególności z punktu widzenia zastosowania narzędzi JIRA. W kolejnej części rozdziału zaprezentowano wielowymiarową eksplorację cech repozytoriów, której dokonano według następujących kryteriów: ogólne charakterystyki, aktywności aktorów, profile komentarzy dodawanych do zgłoszeń oraz korelacja atrybutów zgłoszeń. Eksploracja ta dotyczy trzech obszarów: (1) ogólne raporty statystyczne wraz z czasami aktywności, (2) strukturalne i semantyczne analizy podstawowych cech raportowanych pól, (3) określenie różnic repozytoriów oraz ich niedoskonałości. Ogólne charakterystyki pól repozytorium przedstawiono za pomocą następujących cech: współczynnik wypełnienia poszczególnych pól, rozkład przyjmowanych wartości, najrzadziej i najczęściej występująca wartość w polu, liczba możliwych wartości, cechy czasu dla zgłoszeń. W analizie korelacji atrybutów zgłoszeń modele rozkładu cech repozytoriów ITS mogą być wyznaczone niezależnie lub zależnie od innych pól lub innych połączonych repozytoriów, np. zagregowane dane mogą być uzależnione od filtrów, takich jak typ zgłoszenia, priorytet, istotność, rola reportera.

Rozdział czwarty pracy przedstawia analizę tekstową raportów zgłoszeń opartą na wielowymiarowej eksploracji tekstów dostępnych w repozytoriach zgłoszeń, m.in. w tzw. polach opisowych, takich jak tytuł, opis, komentarze zgłoszenia. Podobnie, jak w rozdziale trzecim, analiza tekstowa poprzedzona jest analizą stanu wiedzy w badanej tematyce, dotyczącej eksploracji tekstów w systemach śledzenia zgłoszeń (ITS) oraz systemach kontroli wersji (SVC). Następnie dokonano analizy słownikowej opisów błędów zgłoszeń w repozytoriach. Autor przenalizował dostępne teksty i doszedł do wniosku, że słowa występujące w zgłoszeniach są zbiorem słów języka naturalnego (należącego do tezaurusa) oraz słów niebędących częścią tezaurusa, które mogą reprezentować między innymi żargon IT/projektowy, nazwy funkcji aplikacji/systemu, odniesienia do kodu aplikacji/systemu, referencje/linki do załączników, referencje/linki do zewnętrznych zasobów, fragmenty kodu, logi aplikacji. Ponadto, tekst może zawierać błędy w pisowni, np. tzw. literówki lub zapożyczenia z języków obcych, co jest bardzo częste w projektach IT. Autor wyróżnił cztery słowniki: słownik słów należących do tezaurusa (NLW), słownik słów projektowych (FW), słownik słów niesklasyfikowanych (NCW), słownik słów zgodnych z notacją *CamelCase* (CCW). Słownik NLW w większości przypadków zawiera elementy występujące w słowniku tezaurus. Autor opracował zbiór wyrażeń regularnych identyfikujących badane teksty. Wyrażenia te zostały użyte w implementacji algorytmu przetwarzania tekstu podlegającego metodom klasyfikacji, który został zaprezentowany w kolejnym podrozdziale. Następnie przedstawiono dwa algorytmy w postaci pseudokodu, obrazującego funkcję generowania słowników NLW, FW, CCW, NCW oraz funkcję pomocniczą, nazwaną „is_valid_tag".
W kolejnej części rozdziału zaprezentowano autorski algorytm klasyfikacji zadań dodawanych do repozytorium zgłoszeń. Jest on wykonywany w dwóch fazach: (1) wstępne przetwarzanie tekstu, (2) klasyfikacja. Wstępne przetwarzanie tekstu obejmuje następujące kroki: (1) pobranie zgłoszeń z repozytorium poprzez API JIRA, (2) stworzenie zbioru oryginalnych encji tekstów oznaczonych identyfikatorem zgłoszenia lub komentarza, (3) redukcja tekstu poprzez usunięcie słów przystankowych oraz cyfr niepołączonych z innymi znakami, (4) transformacja oryginalnych encji tekstów z wykorzystaniem wyrażeń regularnych oraz wyznaczenie opracowanych cech tekstowych. Te kroki algorytmu zaprezentowano w postaci pseudokodu.
Zaproponowany algorytm klasyfikacji tekstów oparty jest na wyborze najlepszego klasyfikatora, który na wejściu przyjmuje listę algorytmów do walidacji, zbiór trenujący i listę cech, które mają zostać zweryfikowane. Wynikiem działania algorytmu jest klasyfikator (algorytm i zestaw cech) z najlepszym wynikiem precyzji oraz macierzą z wynikami wszystkich przeprowadzonych testów.

W kolejnej sekcji zaprezentowano wyniki eksperymentów przeprowadzone dla dwóch zadań klasyfikacji: typu zgłoszenia i kategorii komentarza oraz ich analizę. Ostatni podrozdział rozdziału czwartego zawiera dyskusję zagadnień poruszonych w tym rozdziale, w tym między innymi zagadnień klasyfikacji tekstów zgłoszeń w repozytoriach oraz przeprowadzonych eksperymentów dla zaproponowanych algorytmów.

Rozdział piąty pracy prezentuje wielowymiarową analizę procesu obsługi zgłoszeń w repozytoriach programowych. Analiza taka może być przeprowadzona na dwóch poziomach: (1) ogólnym, tzw. gruboziarnistym i (2) szczegółowym, tzw. drobnoziarnistym. W związku z tym w pierwszej części rozdziału zaprezentowano gruboziarnistą oceną efektywności procesu obsługi zgłoszeń według dwóch metryk: (1) czasu obsługi zgłoszeń, poprzez wyznaczone profile czasowe obsługi zgłoszeń oraz (2) zakres nierozwiązanych zgłoszeń, poprzez liczby nierozwiązanych zgłoszeń. Trzecia metryka to rozkład zmian kodu względem zgłoszeń. Ponadto, Autor zaprezentował pojęcie długu błędów i algorytm detekcji długu błędów wraz ze wzrostami w różnych okresach czasu. Algorytm ten bazuje na czterech danych wejściowych: data początku weryfikowanego okresu, data końca weryfikowanego okresu, interwał czasowy (miesiąc bądź sprint) i próg, powyżej którego liczba nierozwiązanych błędów będzie zliczana. Dodatkowo, opisano sposób minimalizacji ryzyka związanego z długiem błędów.
W drugiej części rozdziału zaprezentowano autorski model obsługi zgłoszeń w repozytoriach programowych. Jest to grafowy model, określony jako IHG, *Issue Handling Graph*. Model został zdefiniowany, opisany i zaprezentowany w formie graficznej oraz poparty przykładami na podstawie projektów informatycznych, które są rozpatrywane przez całą rozprawę, a które stanową punkt wyjścia do badań i analiz rozwiązań zaproponowanych przez Autora. W sekcji 5.2.1 przedstawiono szczegóły modelu IHG w formie opisu przepływu zgłoszeń w modelu. W sekcji 5.2.2. przedstawiono algorytmy analizy grafu IHG, wśród których wyróżniono trzy ich grupy: algorytmy przetwarzania danych, algorytmy agregacji danych, algorytmy pomocnicze. W ramach algorytmów przetwarzania danych zaprezentowano: algorytm pobierania danych, dotyczących poszczególnych zgłoszeń; algorytm budowania ścieżek z listy zgłoszeń JIRA. W ramach algorytmów agregacji danych zaprezentowano: algorytm budowy grafu IHG, który agreguje dane ścieżek wygenerowanych przez poprzedni algorytm. W ramach algorytmów pomocniczych zaprezentowano: algorytm wizualizacji grafu IHG, wygenerowanego przez algorytm budowy grafu, na który składa się algorytm filtracji wierzchołków, algorytm filtracji krawędzi, algorytm generowania pliku stanów, algorytm generowania pliku ze statystykami unikalnych ścieżek grafu, algorytm wyszukiwania i wyznaczanie anomalii (pętli) ścieżek.
Zaprezentowane algorytmy wspierają generowanie szeregu charakterystyk ścieżek oraz stanów grafu, które mogą być grupowane w statystyki ścieżek (czasowe, strukturalne, ilościowe, komentarzy zgłoszeń przechodzących przez ścieżkę) i statystyki stanów (czasowe, ilościowe,, strukturalne). Opracowane algorytmy mogą służyć do generowania grafu pełnego lub z ograniczeniami, np. dla określonego typu zgłoszenia, stanu, grupy użytkowników tworzących zgłoszenia lub czasu, w którym zostały one dodane do repozytorium.
W kolejnych sekcjach opisano charakterystyki profili stanów oraz ścieżek obsługi zgłoszeń opracowanych na podstawie modelu IHG – charakterystyki czasowe stanów grafu IHG w sekcji 5.2.3 oraz profile ścieżek obsługi zgłoszeń w sekcji 5.2.4, gdzie wyróżniono dwie klasy profili ścieżek: wydajnościowo-czasowe i strukturalne. Opisywane zagadnienia poparto przykładami konkretnych projektów programistycznych. Podczas generowania profili ścieżek zgłoszeń zaobserwowano pojawienie się anomalii w procesie obsługi zgłoszeń. W związku z tym zaproponowano autorską metodę wykrywania tych anomalii, opartą na użyciu specjalne zdefiniowanych wyrażeń regularnych. Kolejne sekcje zawierają analizę czasową obsługi zgłoszeń w ścieżkach oraz dyskusję zaproponowanych rozwiązań i otrzymanych wyników eksperymentów przeprowadzonych dla tych rozwiązań.

Szósty rozdział rozprawy zawiera podsumowanie pracy, opis osiągnięć pracy oraz perspektywy rozwoju tematu pracy w niedalekiej przyszłości.

Siódmy rozdział pracy prezentuje zastosowaną bibliografię, która zwiera 108 pozycji bibliograficznych.

Ostatnia część rozprawy zawiera załączniki do głównej części pracy, które prezentują odpowiednio: profile ścieżek obsługi zgłoszeń dla projektu Groovy, grafy modelu IHG dla projektu MongoDB, zestaw atrybutów ścieżek obsługi zgłoszeń oraz szczegółowe charakterystyki ścieżek obsługi zgłoszeń.

Ocena zastosowanego piśmiennictwa

Rozprawa zawiera 108 pozycji bibliograficznych, w tym 11 pozycji internetowych, które z reguły stanowią dokumentację techniczną wybranych narzędzi i technologii. Zdecydowana większość pozycji to artykuły naukowe, opublikowane w czasopismach naukowych lub w materiałach konferencyjnych konferencji naukowych w ostatnich kilku-kilkunastu latach. Wśród pozycji są również prace opublikowane przez Autora – w sumie 4 prace, których jest współautorem.

Autor przedstawił analizę źródeł literaturowych częściowo w rozdziale 2 – na temat repozytoriów programowych oraz częściowo w rozdziale 3 rozprawy – na temat eksploracji tekstów w systemach śledzenia zgłoszeń (ITS) i w systemach kontroli wersji (SVC).

Przeprowadzona analiza obecnego stanu wiedzy w poruszanej tematyce oraz istniejących rozwiązań wskazuje, że Autor poprawnie rozumie problemy, związane z tematem pracy. Posiada wiedzę na temat rozwiązań krajowych i światowych w poruszanym zakresie. Co prawda część istotnych zagadnień oraz problemów została jedynie zasygnalizowana, jednakże taka forma pozwala mieć nadzieję, że Autor posiada również wiedzę szczegółową na ich temat.

Wskazanie i ocena celu pracy

Autor sformułował tezę pracy i problem badawczy w rozdziale 2. rozprawy. W rozdziale tym zarysowano problemy, związane z procesem wytwarzania oprogramowania w aspekcie zgłoszeń błędów i nowych funkcji do realizowanych projektów.
Z tezą pracy powiązano zadania badawcze, dotyczące trzech obszarów: (1) eksploracji zawartości informacyjnej repozytoriów zgłoszeń (poprzez opracowanie charakterystyk strukturalnych, czasowych, zdefiniowanie profili różnych aspektów informacyjnych oraz opracowanie algorytmów analiz statystycznych i tekstowych), (2) opracowania rozszerzonego modelu grafowej obsługi zgłoszeń (IHG, *Issue Handling Graph*) oraz metod analizy zorientowanej na badanie profili obsługi różnych typów zgłoszeń i detekcji anomalii w repozytoriach, (3) weryfikacji opracowanej metodologii na danych z repozytoriów projektów typu open source oraz projektów komercyjnych.

Niemniej jednak cel pracy nie został jasno i precyzyjnie zidentyfikowany i zaprezentowany. Jedynie można domniemywać, że celem pracy jest zrealizowanie zdefiniowanych zadań badawczych.
Na stronach 18-19, w rozdziale 3, Autor definiuje cele badawcze w odniesieniu do repozytoriów projektowych oraz ich analizy. Cele te częściowo pokrywają się zadaniami badawczymi określnymi w rozdziale 2.

Wskazanie i ocena zastosowanych metod badawczych

Autor w swojej pracy stosuje następujące metody badawcze: analiza dostępnej literatury w temacie rozprawy, analiza przedmiotu badań, czyli repozytoriów programowych, obserwacja prowadzonych projektów informatycznych, metody statystyczne i ilościowe w analizach danych uzyskanych z repozytoriów programowych, prace eksperymentalne.

Uważam, że wszystkie zastosowane metody badawcze są poprawnie wybrane i zastosowane. Stanowią kompletny warsztat badawczy, niezbędny dla przeprowadzenia badań w ramach tematu rozprawy.

Autor opiera się również na własnym praktycznym doświadczeniu, wyniesionym z pracy nad projektami komercyjnymi. Takie doświadczenie stanowi niewątpliwie wartość dodaną w pracy nad poruszanym tematem oraz nad przygotowywaniem rozprawy, w szczególności w obszarze inżynierii oprogramowania.

Ocena części rozprawy dotyczącej omówienia wyników badań

Tematyka badań Autora obejmuje wielowymiarową analizą i eksplorację procesu obsługi zgłoszeń w repozytoriach programowych. Badania przeprowadzone w ramach realizacji pracy obejmują dwa powiązane ze sobą obszary: (1) eksploracja zawartości informacyjnej programowych repozytoriów zgłoszeń i (2) wielowymiarowa analiza obsługi zgłoszeń.

Wyniki badań zostały zaprezentowane przez Autora częściowo w rozdziale 3 i 4 oraz w rozdziale 5 rozprawy, odpowiednio na temat analizy repozytoriów programowych, analizy tekstowej raportów zgłoszeń oraz analizy obsługi zgłoszeń. Ponadto, dyskusja na temat uzyskanych wyników badań została zaprezentowana pod koniec rozdziału 4 i 5 oraz w podsumowaniu samej rozprawy.

Sposób prezentacji podejmowanych w pracy zagadnień oraz uzyskanych wyników badań jest na ogół jasny i zrozumiały, chociaż Autor stosuje dużo różnorodnych symboli na oznaczenie poszczególnych projektów oraz ich elementów. Wprowadzone symbole są ogólne, czasami nieintuicyjne, np. nazwy projektów używanych w badaniach. Ponadto, część zagadnień, a także wyników pracy Autora poruszona jest w rozprawie bardzo pobieżnie.

Niemniej jednak tekst rozprawy został przygotowany starannie. Strona redakcyjna nie budzi większych zastrzeżeń – moje uwagi zostały przedstawione w kolejnym punkcie recenzji, dotyczącym nieprawidłowości w rozprawie.

Praktyczne zastosowanie uzyskanych wyników badań

Autor w swoich badaniach łączył aspekty teoretyczne i praktyczne, Ponadto, jak sam wielokrotnie stwierdza w rozprawie, w trakcie prac opiera się na swoim doświadczeniu praktycznym, wyniesionym z wieloletniej pracy w tzw. przemyśle, przy realizacji komercyjnych projektów informatycznych.

W ramach realizacji pracy doktorskiej Autor przeanalizował istotną liczbę repozytoriów projektów, zarówno open source, jak i komercyjnych, celem zbadania zagadnień związanych z eksploracją powiązanych ze sobą procesów wytwarzania i utrzymania oprogramowania oraz

analizą istniejących w nich danych, W związku z tym opracował modele danych i schematy eksploracji, które zostały wsparte odpowiednimi metrykami, mającymi na celu wspieranie: (1) ekstrakcji cech charakterystycznych (składniowych, semantycznych, czasowych i statystycznych) repozytoriów śledzenia problemów, odniesionych do różnych perspektyw obserwacji, (2) śledzenie efektywności obsługi zgłoszonych problemów, (3) wykrywanie niedoskonałości repozytoriów programowych oraz procesu obsługi zgłoszeń.

Zaproponowany model IHG wspierany jest przez opracowane autorskie algorytmy analizy powiązane ze zdefiniowanymi profilami strukturalnymi i statystycznymi i ich miarami, obejmującymi różne aspekty procesów obsługi zgłoszeń. Ponadto, w trakcie analiz ścieżek obsługi zgłoszeń wielu projektów Autor zaobserwował anomalie, dla których opracował metody i schematy analiz.

Stworzona metoda została wykorzystana przez Autora w praktyce do zbadania i oceny procesu obsługi zgłoszeń w projekcie komercyjnym, a następnie od przeprowadzania w nim usprawnień. Według zamieszczonych przez Autora informacji, usprawnienia te dotyczyły optymalizacji złożoności zgłoszeń i optymalizacji procesu obsługi zgłoszeń. Ułatwiło to i przyśpieszyło realizację procesu obsługi zgłoszeń w repozytorium projektu.

Wyniki rozprawy mają zatem znaczenie dla dalszego rozwoju nauki w dziedzinie inżynierii oprogramowania oraz posiadają wartość aplikacyjną i praktyczną dla rzeczywiście realizowanych projektów informatycznych.

Ewentualne nieprawidłowości, które pojawiły się w rozprawie

Analiza rozprawy nasunęła mi kilka uwag krytycznych:

- Wykresy na rysunkach 1-4 i 10 nie posiadają oznaczenia nazw osi – powoduje to, że zaprezentowane wykresy są niezrozumiałe.
- Grafy przedstawione na rysunkach 5 i 11-14 są niestety bardzo mało czytelne.
- Niewłaściwe użycie słowa „funkcjonalność" zamiast słowa „funkcja" – na wielu stronach, w całej pracy.
- Niewłaściwe użycie słowa „sekcja" w stosunku do podrozdziału – na wielu stronach, w całej pracy, np. „sekcja 5.1." na str. 108.
- Zbyt ogólne i przez to niezrozumiałe nazwy niektórych rozdziałów i podrozdziałów oraz sekcji pracy.
- Forma niektórych zdań nasuwa wątpliwości stylistyczne.
- Standardowo tzw. „podpisy" dla tabel umieszczane są nad tabelami, natomiast w pracy znajdują się pod tabelami – na wielu stronach, w całej pracy.
- Co oznacza skrót „Alg"? oczywiście można się łatwo domyślić, ale następujące przykładowe sformułowania nie są eleganckie: „dlatego opracowałem Alg. 1 wspierający ten proces", „Alg. 2, zaimplementowany został z wykorzystaniem biblioteki", „Jako parametr wejściowy Alg. 1 przyjmuje", „przedstawiony w pseudokodzie Alg. 6", „(wynik działania Alg. 11)".
- Podobnie, co oznacza skrót „Tab"? – na przykład sformułowania „Statystyka zaprezentowana w Tab. 12 daje istotny wgląd", „budowanego za pomocą wyrażeń regularnych Tab. 11".
- Brak spisu rysunków i spisu tabel na końcu pracy. Brak indeksu stosowanych w pracy symboli.
- Numeracja rysunków, tabel, algorytmów jest niezgodna z ogólnie przyjętymi zasadami dla tzw. długich tekstów, jakim jest rozprawa doktorska.

- Praca powinna być napisana w formie bezosobowej, chociaż w przeważającej części.
- Częste stosowanie sformułowania „oparte o" – poprawnie powinno się stosować sformułowanie „oparte na".
- Błędy interpunkcyjne.
- Błędy językowe, np. „Model IHG umożliwia wyprowadzenie innych zagregowanych profili np. warianty ścieżek np. ze wspólnym stanem początkowym, wspólnym stanem początkowym i pośrednim, wspólnym stanem początkowym i końcowym itp.".

Przedstawione uwagi nie obniżają jednakże mojej pozytywnej oceny pracy.

Ocena, czy rozprawa stanowi oryginalne rozwiązanie problemu naukowego

Uważam, że teza rozprawy oraz obszar naukowy rozprawy zostały określone jasno i precyzyjnie. Praca ma charakter teoretyczno-implementacyjny. Część teoretyczna obejmuje:

- przeprowadzenie analizy dostępnych źródeł literaturowych na temat poruszanych w rozprawie zagadnień, dotyczących procesu eksploracji repozytoriów programowych w projektach informatycznych, w szczególności problemu analizy raportów zgłoszeń i ich obsługi,
- autorskie opracowanie profili statystycznych zawartości repozytoriów, profili aktywności aktorów (uczestników) projektu oraz wielowymiarowa eksploracja opisów i komentarzy zgłoszeń,
- zdefiniowanie ogólnych i szczegółowych charakterystyk procesu obsługi zgłoszeń,
- opracowanie autorskiej metody obsługi zgłoszeń, opartej na grafowym modelu IHG,
- opracowanie algorytmów analizy schematów obsługi zgłoszeń w kontekście zaproponowanej ich taksonomii.

Część implementacyjna obejmuje stworzenie narzędzia o nazwie *IssueAnalyzerTool*, które zostało użyte do przeprowadzenia analiz zaproponowanych rozwiązań teoretycznych na wybranym reprezentatywnym zbiorze repozytoriów programowych.

Tematyka rozprawy sytuuje ją w obszarze badawczym, związanym z poszukiwaniem efektywnych metod, dotyczących budowy systemów informatycznych, w szczególności eksploracji repozytoriów programowych celem szybkiego i efektywnego wyszukiwania i obsługi zgłoszeń na temat błędów i nowych funkcji w realizowanych projektach.

Uważam, że podjęcie tematu pracy doktorskiej w rozpatrywanym obszarze i zakresie jest celowe i w pełni uzasadnione potrzebą zapewniania i zwiększania jakości współcześnie tworzonego oprogramowania, a w szczególności dużych, złożonych oraz dynamicznych w swoim działaniu systemów informatycznych. Jest to w pełni uzasadnione ze względów teoretycznych, poznawczych i praktycznych na tle obecnego stanu wiedzy w rozpatrywanym obszarze.

Ocena, czy rozprawa prezentuje ogólną wiedzę teoretyczną kandydata w dyscyplinie oraz umiejętność samodzielnego prowadzenia pracy naukowej

W moim przekonaniu Autor w wystarczającym stopniu przeanalizował aktualny stan wiedzy w rozpatrywanym temacie, jasno sformułował problemy, a następnie w zadawalającym stopniu rozwiązał je w swojej pracy.

Oryginalnym rezultatem rozprawy, a tym samym samodzielnym i oryginalnym dorobkiem Autora według mojej oceny jest opracowanie metody wspomagającej budowę systemów informatycznych, która umożliwia wielowymiarową eksploatację repozytoriów programowych, w szczególności w zakresie raportów zgłoszeń i ich obsługi.

Dorobek Autora powiększa implementacja zaproponowanych rozwiązań poprzez stworzenie prototypu narzędzia o nazwie *IssueAnalyzerTool* do analizy zgłoszeń w repozytoriach w sposób zautomatyzowany. Zaproponowana metoda została ponadto zweryfikowana poprzez zastosowanie jej w rzeczywiście realizowanych projektach informatycznych.

Uważam, że przedstawiona mi do recenzji rozprawa prezentuje ogólną wiedzę teoretyczną kandydata w dyscyplinie informatyka techniczna i telekomunikacja oraz umiejętność samodzielnego prowadzenia pracy naukowej.

Wniosek końcowy

W mojej ocenie, Pan mgr Bartosz Dobrzyński w swojej rozprawie trafnie zdefiniował problem badawczy, a następnie go rozwiązał w odpowiednim zakresie. Pozytywnie oceniam wszystkie wymienione powyżej elementy pracy.

Uważam, że rozprawa spełnia wymagania stawiane rozprawom doktorskim przez obowiązujące przepisy o stopniach i tytułach naukowych. W związku z tym wnoszę o dopuszczenie jej do publicznej obrony.