dr hab. inż. Andrzej Paszkiewicz

Dowództwo Komponentu Wojsk Obrony Cyberprzestrzeni

ul. gen. broni Tadeusza Buka 1

05-119 Legionowo

Warszawa 11, listopada 2023 r.

„Nie osądzajcie innych, a nie będziecie osądzeni; bo jakim osądem sądzicie, takim zostaniecie osądzeni i jaką miarą mierzycie, taką i wam zostanie wymierzone..."

(Mt 7, 1-15)

RECENZJA ROZPRAWY DOKTORSKIEJ DLA RADY DYSCYPLINY NAUKOWEJ INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA WOJSKOWEJ AKADEMII TECHNICZNEJ

Tytuł rozprawy: Destylacja korpusu danych tekstowych w procesie fuzzingu z wykorzystaniem algorytmu genetycznego

Autor rozprawy kpt. mgr inż. Marcin Pachnik

Promotor: dr hab. inż. Kazimierz Worwa, prof. WAT

Promotor pomocniczy: płk dr inż. Rafał Kasprzyk

Wstęp

Praca została napisana w języku polskim i liczy 133 strony, włączając spis treści, wykaz stosowanych symboli i oznaczeń, słownik akronimów i pojęć, spis literatury, wykaz tabel oraz spis rysunków. Rozdział 1 (str. 15-18) zawiera wstępne wiadomości z zakresu tematyki rozprawy, w tym istotę pojęcia zawartego w tytule rozprawy zwanego fuzzingiem. Jest to metoda automatycznego wyszukiwania błędów i luk w zabezpieczeniach oprogramowania, której autorstwo przypisuje się ok. roku 1988 Bartowi Millerowi z Uniwersytetu Wisconsin. Starsi adepci informatyki przypominają

sobie, że w dawniejszych czasach dobrą metodą badania odporności oprogramowania było oddanie klawiatury w ręce dzieciom w trakcie działania programu. Żaden dorosły człowiek nie wymyśliłby takiej kombinacji klawiszy, która spowodowałaby zastopowanie działającego programu. Dzieciom często się to udawało. W dzisiejszych czasach, kiedy aplikacje informatyczne stosuje się powszechnie – cała nasza współczesna cywilizacja nie mogłaby bez nich poprawnie funkcjonować - zagrożenie jest o wiele większe niż kiedyś, gdy medium dostępu do programu stanowiła jedynie klawiatura, pomijając rzecz jasna znacznie niższy poziom dzietności, z jakim się w społeczeństwach rozwiniętych obecnie borykamy. Zatem tę tradycyjną metodę należy uznać za nieskuteczną. Odpowiedzią na istniejące zagrożenie jest metoda automatycznego badania odporności oprogramowania na błędy, przy czym nie tylko te, które w sposób niezamierzony się pojawiają. Współczesne metody fuzzingu, poprzez wygenerowanie różnych stanów oprogramowania często „prowokują" powstawanie problemów, które trudno przewidzieć. Przejrzenie całej przestrzeni tych stanów jest przy skomplikowanym, dużej objętości oprogramowaniu niemożliwe. Rola wykorzystania losowości, (którą, kiedyś generowały dzieci) wydaje się tu jak najbardziej stosowna do równomiernego ale wybiórczego testowania przestrzeni stanów oprogramowania. Autor rozprawy uznał, że genetyka jest do tego celu narzędziem jak najbardziej odpowiednim. Innym narzędziem mogą być algorytmy sztucznej inteligencji czy, być może, niektóre algorytmy przeszukiwania związane z wyżarzaniem kwantowym, w połączeniu z zastosowaniem (w przyszłości) prawdziwego komputera kwantowego.

Rozdział pierwszy (str. 15-18) zawiera także cel i zakres pracy. W Rozdziale drugim (str. 19-38) przedstawiono w historię, podział i zastosowanie testów fuzzingowych. Znalazły się tu m.in. popularne i często wykorzystywane fuzzery i destylatory. Poruszony został także problem niebagatelny - poprawnego przygotowania zbioru danych testowych pod kątem automatycznego poszukiwania podatności. Autor nie omieszkał krótko zasygnalizować problem antyfuzzingu. Mamy z nim do czynienia w przypadku nieuczciwego dostawcy oprogramowania, który wbudował w software pewne podatności ale nie chce aby odbiorca mógł je wykryć. W tym celu stosuje techniki obfuskacyjne (zaciemniające), szyfrowanie dynamiczne kodu lub kompresję. Rozdział trzeci (str. 39-54) jest poświęcony algorytmom ewolucyjnym, w tym ich historii a także klasyfikacji tych algorytmów. Autor omawia w tym rozdziale tematykę kodowania chromosomów jak również operatorów genetycznych mutacji i krzyżowania. Opisana została rola zbieżności w kontekście zakończenia pracy algorytmu ewolucyjnego. Rozdział czwarty (str. 55-60) omawia tematykę zjawisk epigenetycznych i ich rolę w algorytmach ewolucyjnych. Stanowi on podbudowę pod sformułowanie problemu badawczego, które następuje w Rozdziale piątym (str. 62-80). Rozdział szósty (str. 82-114) przedstawia różne warianty przeprowadzonych badań. Z punktu widzenia konstrukcji pracy ma on charakter zasadniczy, ponieważ potwierdza w sposób doświadczalny przyjętą tezę. Wreszcie krótki Rozdział siódmy w sposób syntetyczny podsumowuje całość dysertacji.

Na uwagę zasługuje bibliografia przypisana do przedłożonej rozprawy. Zawiera ona 151 pozycji literatury, relewantnej z tematyką pracy, z różnego okresu rozwoju dziedzin związanych z algorytmami genetycznymi i testowaniem oprogramowania. Dwie spośród zacytowanych prac są autorstwa Doktoranta, przy czym jedna samodzielna a druga, której współautorami są Promotorzy. Cytowania pojawiają się we właściwych miejscach, gdzie Doktorant odnosi się do jakiegoś problemu już wcześniej poruszanego przez innych autorów. Reasumując kwestię cytowań – na sześćdziesięciu stronach odniesiono się do 151 prac, co daje imponującą „gęstość cytowania" wyrażającą się liczbą 2,517 na stronę. Niestety w jednym przypadku, wskazana strona internetowa otwiera się w nieczytelnym formacie. Być może wygenerowałem przez przypadek zbiór danych fuzzingowych, które „przewróciły" otwieraną stronę.

Na uwagę również zasługuje liczba rysunków i tabel ilustrujących poszczególne tematy omawiane w pracy lub wyniki obliczeń – w sumie 57 rysunków oraz 42 tabele.

Omówienie rozprawy

Jakie zagadnienie naukowe jest rozpatrzone w pracy (teza rozprawy) i czy zostało ono dostatecznie jasno sformułowane przez Autora? Jaki charakter ma rozprawa (teoretyczny, doświadczalny, inny)

Teza rozprawy została przedstawiona explicite: **„Wykorzystanie w procesie destylacji wielokryterialnego algorytmu genetycznego wzbogaconego o operator epigenetyczny oraz mechanizm sterowania zbieżnością pozwala na efektywną redukcję korpusu danych testowych w procesie fuzzingu"**. Tezę sformułowano wprawdzie w sposób zrozumiały ale nie w pełni precyzyjnie. Pewne obiekcje może budzić sformułowanie „efektywna redukcja", która sama w sobie nosi cechy pewnego rodzaju fuzzingu, krótko mówiąc wiąże się pewnym subiektywizmem. W procesie przeprowadzonych licznych doświadczeń i żmudnego gromadzenia wyników obliczeń, wykazano w sposób eksperymentalny, że nowa metoda w wielu przypadkach rzeczywiście radzi sobie lepiej z redukcją korpusu danych testowych niż wcześniej stosowane. Ma to wpływ też na skrócenie czasu obliczeń. Ponieważ praca nosi charakter wybitnie eksperymentalny, tego rodzaju dowód wyższości zaproponowanego rozwiązania, pod warunkiem losowości danych testowych jest do przyjęcia.

Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł (w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle) świadczący o dostatecznej wiedzy autora. Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonywujący?

Studia literaturowe przeprowadzone w rozprawie doktorskiej zostały przeprowadzone w sposób bardzo rzetelny. Zacytowano i omówiono nie tylko publikacje w sposób ściśle związany z rozważanymi zagadnieniami w recenzowanej pracy ale także prace ustalające pewien szerszy kontekst i background dotyczący obszaru zagadnień, którymi Doktorant się zajmował (zajmuje). Wydaje mi się, że jedyną rzeczą, na którą w przyszłości warto zwracać uwagę, jest podawanie liczby stron w cytowanych publikacjach, tam gdzie jest to oczywiście możliwe. To samo dotyczy umieszczania numerów DOI, które stało się w ostatnich latach dobrym zwyczajem publikacji naukowych. Nie mam innych krytycznych uwag dotyczących wniosków z przeprowadzonej analizy literaturowej.

Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy czy poziomu techniki reprezentowanych przez literaturę światową?

W pełni zgadzam się z Autorem rozprawy, uznając za Jego oryginalny wkład w tematykę zastosowania algorytmów genetycznych do usprawnienia wyszukiwania błędów oprogramowania, elementy przytoczone w podsumowaniu pracy. Są one następujące (patrz str. 117 dysertacji):

- Rozszerzenie destylacji korpusu danych testowych wykorzystywanych w procesie fuzzingu z problemu minimalnego ważonego pokrycia zbioru do problemu wielokryterialnego pokrycia zbioru;
- Wykorzystanie entropii uproszczonej jako nowego dodatkowego kryterium destylacji korpusu danych testowych;
- Zaproponowanie operatora epigenetycznego, który uwzględnia eliminację osobnika w procesie selekcji jako czynnik stresowy, niezbędny do jego wystąpienia tj. poprawne odwzorowanie zjawiska z ewolucji biologiczne;
- Zaproponowanie kodowania, które pozwala na implementację operatora epigenetycznego tak, aby był zgodny z rozwiązywanym problemem i twierdzeniem J. H. Hollanda (poz.[91] w spisie literatury zawartym w pracy);
- Rozszerzenie algorytmu VEGA o operator epigenetyczny oraz mechanizm sterowania zbieżnością;
- Doświadczalne ustalenie minimalnego skutecznego prawdopodobieństwa p_e wystąpienia operatora epigenetycznego oraz jego wariantu dla poszczególnych formatów plików podczas destylacji korpusu.

Wydaje mi się, że niezwykle trudną rzeczą byłoby ustalenie pozycji rozprawy w stosunku do najbardziej zaawansowanych badań prowadzonych na świecie. Można z dużą dozą pewności powiedzieć, że większość istotnych prac z tego zakresu stanowią badania, które posiadają oklauzulowany charakter, w szczególności badania na biologicznych łańcuchach DNA. W świetle ogólnie dostępnych źródeł, recenzowana praca wydaje się mieć silną pozycję (tym razem recenzent posłużył się fuzzingiem).

Czy autor wykazał umiejętność poprawnego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)?

Autor dość poprawnie posługuje się technicznym językiem polskim. Rozprawę tę czyta się (raczej zgłębia się) bez większych problemów i zacięć. Jak w każdej rozprawie, nawet po ostatecznych poprawkach pozostaje pewna ilość tzw. „głupich błędów", które w żaden sposób nie umniejszają jakości i wartości rozprawy. Również i w przypadku recenzowanej pracy znalazła się „dyżurna" reprezentacja niewielkich uchybień natury językowej. Przykładowo przytaczam niektóre z nich:

Str. 31, 8 wiersz od dołu

Jest: „Moonight" Powinno być: „Moonlight";

Str. 32, 9 wiersz od dołu

Jest: „oblicza rozwiązanie" Powinno być: „znajduje rozwiązanie";

Str. 37, 10 wiersz od góry

Jest: „są pakowane [64] i zaciemniane" Powinno być: „są kompresja [64] i zaciemnianie";

Str. 42, 10 wiersz od dołu

Jest: „funkcja ekwipotencjalna"	Powinno być: „funkcja eksponencjalna";
Str. 50, 10 wiersz od góry	
Jest: „rodzicielskiej."	Powinno być: „rodzicielskiej,";
Str. 56, wiersz 8 od dołu	
Jest: „takich ilość"	Powinno być: „takich jak ilość";
Str. 62, 4 wiersz od góry	
Jest: „z jednego z elementu"	Powinno być: „z jednego elementu";
Str. 64, wiersz 14 od dołu	
Jest: „dla co najmniej jednego."	Powinno być: „dla co najmniej jednego i.";

Innego rodzaju uchybienia dotyczą nieścisłości w sensie matematycznym i są one gatunkowo większym przewinieniem, np.:

Str. 42. opis wzoru (2) powinien wyglądać następująco

Długość ξ liczby $2^{\xi} - 1$ wyznaczana jest jako najmniejsza liczba naturalna ξ spełniająca nierówność

$$(v_R - v_L) \cdot 10^q \leq 2^{\xi} - 1$$

Str. 66 jako komentarz do wzoru na prawdopodobieństwo całkowite (20) powinien pojawić się zapis: „gdzie $B_i \cap B_j = \phi$ dla $i, j \in \{1,2, \dots, d, \dots, D\}, i \neq j$".

Lewa strona wzoru (21) na tej samej stronie powinna mieć postać

$$\varepsilon = \frac{w_1}{U} \cdot \frac{(U - w_1)}{U - 1} + \cdots + \frac{w_d}{U} \cdot \frac{(U - w_d)}{U - 1} + \cdots + \frac{w_D}{U} \cdot \frac{(U - w_D)}{U - 1}$$

W wielu miejscach Autor zapomina o konwencji oznaczania zmiennych i parametrów użytych we wzorach i w ich opisie tymi samymi symbolami, np. symbol d – wyedytowany za pomocą narzędzia do tworzenia równań i symbol d, pomimo, że to ta sama litera oznaczają formalnie coś innego. Wprowadza to pewien bałagan tak jak jest to np. na str. 66 i wielu innych miejscach.

Str. 68, wiersz 9 od góry (opis entropii u – tego przypadku testowego) wartość entropii należy do zbioru dwupunktowego, tzn. jest $\varepsilon_u \in \{0,1\}$, podczas gdy powinno być $\varepsilon_u \in [0,1]$. Nie ma powodu aby wzmiankowana entropia miała być zbiorem o wartościach dyskretnych.

Jakie są słabe strony rozprawy i jej główne wady?

Praca, jak już wcześniej wzmiankowano, jest napisana na dobrym poziomie. Uzyskane wyniki doświadczalne są wartościowe i dobrze udokumentowane (tabele i wykresy). Istotnym mankamentem pracy jest natomiast mała w niej obecność formalnych metod opisu i modelowania matematycznego. Jeśli występują, to trochę na zasadzie „zła koniecznego". Autor nie posilił się np. o

dokonanie oceny złożoności obliczeniowej zastosowanego podejścia. Warto zaznaczyć, że już nawet niektóre elementy składowe algorytmów fuzzingu cechują się złożonością wyższą niż wielomianowa, np. problem pokrycia zbioru, czy jego „młodszy brat" – problem pokrycia wierzchołkowego grafu. Byłoby cenną rzeczą, gdyby Autorowi pracy udało się uzyskać wynik mówiący o tym, że da się coś zrobić z wykładnikiem funkcji złożoności problemu badawczego, np. doprowadzając tę złożoność do korzystnej z punktu widzenia obliczeń funkcji subwykładniczej (podwykładniczej). Brak podejścia ilościowego stawia recenzenta przed dylematem oceny wartości pracy na podstawie samej analizy tabel i wykresów zawartych w pracy. Brakuje także informacji, jak Autor radził sobie z wytwarzaniem losowości.

Jaka jest przydatność rozprawy dla nauk technicznych?

W ramach prac zrealizowanych na potrzeby rozprawy powstało oprogramowanie, pozwalające, jak się wydaje, znacznie obniżyć złożoność problemu wykrywania błędów w oprogramowaniu. Jest to bardzo istotne z punktu widzenia współczesnych potrzeb dzisiejszego ponad miarę zinformatyzowanego świata. Autor rozprawy poprzez swoje badania z dużym wykorzystaniem intuicji inżynierskiej uzyskał wartościowe, aplikacyjne wyniki, które moim zdaniem powinny doczekać się w najbliższym czasie analizy od strony ilościowej. Jeśli można użyć określenia z branży odzieżowej – praca jest dobrze skrojona pod kątem nauk technicznych.

Podsumowanie i ocena rozprawy

W swojej rozprawie doktorskiej pan kapitan magister inżynier Marcin Pachnik jasno sformułował i poprawnie zrealizował zagadnienie badawcze z zakresu testowania oprogramowania z wykorzystaniem algorytmów genetycznych. Jego rozprawę oceniam pozytywnie uważając, że spełnia ona wymagania stawiane przez obowiązującą USTAWĘ o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki i wnioskuję o jej dopuszczenie do publicznej obrony.

Andrzej Paszkiewicz

dr. hab. inż. Andrzej Paszkiewicz