СЕКЦІЯ 8
КІБЕРБЕЗПЕКА, КОМП’ЮТЕРНА ТА ПРОГРАМНА ІНЖЕНЕРІЯ

УДК 004.415.53:004.052.3

https://doi.org/10.68134/ah.2026-11-20.024

Федорняк Богдан Романович
студент, Карпатський національний університет імені Василя Стефаника

Науковий керівник: Костишин Л. П., кандидат фізико-математичних наук, доцент, доцент кафедри диференціальних рівнянь і прикладної математики, Карпатський національний університет імені Василя Стефаника

ОЦІНЮВАННЯ ЕФЕКТИВНОСТІ ІНСТРУМЕНТІВ АВТОМАТИЗОВАНОГО ТЕСТУВАННЯ ЗА ВПЛИВОМ НА НАДІЙНІСТЬ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ: ІНТЕГРАЛЬНИЙ ІНДЕКС RTAEI

Анотація. Інструменти автоматизованого тестування зазвичай обирають за зручністю, вартістю та сумісністю. Їх внесок у надійність програмного забезпечення оцінюють рідко. Мета роботи — розробити й перевірити методику кількісного оцінювання ефективності таких інструментів з орієнтацією на надійність. Запропоновано інтегральний індекс RTAEI — зважене середнє геометричне семи нормованих критеріїв, один з яких характеризує відносне зниження ймовірності відмови за моделлю росту надійності Гоела–Окумото; ваги визначаються ентропійним методом із нижньою межею ваги критерію надійності, а деградацію виявляють контрольними межами Шухарта. Методику реалізовано у вебзастосунку мовою Python і перевірено автоматизованими тестами, а також імітаційно на 600 наборах даних із відомим еталонним рейтингом та на вимірюваннях двох інструментів для вебзастосунку із засіяними дефектами. Індекс краще за рівні й експертні ваги відновлює рейтинг за надійністю і знижує інструменти з незбалансованим профілем; значну деградацію виявлено за два випуски у 77–80 % випадків. Частину переваги зумовлено нижньою межею ваги надійності, а реальний кейс має лише демонстраційний характер і не підтверджує гіпотез статистично.

Ключові слова: автоматизоване тестування, надійність програмного забезпечення, інтегральний індекс, ентропійні ваги, модель Гоела–Окумото, контрольні межі Шухарта.

Abstract. Test automation tools are usually selected by usability, cost and compatibility, while their contribution to software reliability is rarely quantified. The aim of this work is to develop and validate a method for quantifying the effectiveness of such tools with respect to reliability. We propose the RTAEI composite index, a weighted geometric mean of seven normalised criteria, one of which is the relative reduction of failure probability estimated with the Goel–Okumoto reliability growth model; weights are obtained by the entropy method with a lower bound on the reliability weight, and degradation is detected with Shewhart control limits. The method is implemented in software and validated by simulation on 600 datasets with a known reference ranking and on measurements of two tools against a web application with seeded defects. The index recovers the reliability-based ranking better than equal and expert weights and lowers the rank of unbalanced tools; substantial degradation was detected within two releases in 77–80 % of cases. The results support using RTAEI for reliability-oriented tool selection and monitoring, although part of the advantage stems from the lower bound on the reliability weight and the real-world case is only illustrative.

Keywords: test automation, software reliability, composite index, entropy weights, Goel–Okumoto model, Shewhart control limits.

Вступ

Автоматизоване тестування стало основним засобом забезпечення якості програмного забезпечення (ПЗ) в умовах коротких циклів випуску. Наявні підходи до вибору інструментів

(AHP, TOPSIS, зважена сума) спираються на експертні ваги й узагальнені показники [3], а кількісний зв'язок між інструментом і надійністю продукту досліджено недостатньо. Для оцінювання надійності застосовують моделі росту надійності, зокрема неоднорідний пуассонівський процес Гоела–Окумото [1] та S-подібну модель Ямади [2]. Окрема проблема — нестабільні тести (flaky tests), що спотворюють результати прогонів [5]. Додатково потребує обґрунтування вибір способу агрегування критеріїв у складених індексах [4]: за адитивної згортки сильні допоміжні показники компенсують слабку надійність.

Мета і завдання

Мета — розробити й експериментально перевірити методику оцінювання ефективності інструментів автоматизованого тестування за впливом на надійність ПЗ. Завдання: 1) обґрунтувати систему критеріїв і спосіб згортки; 2) реалізувати методику в програмному засобі; 3) перевірити її на даних із відомим еталоном і на реальних вимірюваннях; 4) оцінити здатність виявляти деградацію інструмента в часі.

Методи

Індекс RTAEI обчислюється як зважене середнє геометричне семи критеріїв $x_1 \ldots x_7$ на шкалі [0; 1] з нижньою межею значень $\varepsilon = 0{,}05$: $\mathrm{RTAEI} = (x_1')^{w_1} \cdot \ldots \cdot (x_7')^{w_7}$, де $x_j' = \max(x_j; \varepsilon)$, w_j — ваги, $\Sigma w_j = 1$. Критерії x_1–x_6 — ефективність виявлення дефектів, повнота покриття, швидкість зворотного зв'язку, стабільність тестів, низька вартість супроводу, інтеграція з CI/CD (часові показники відповідають характеристикам якості за ISO/IEC 25022 і 25023 [8]); x_7 — відносне зниження ймовірності відмови за моделлю [1, 2]. Ваги визначаються ентропійним методом за вікном останніх випусків; критерій x_6 має фіксовану вагу 0,10, а вага x_7 не менша за $w_{min} = 0{,}30$. Для виявлення деградації використано контрольні межі Шухарта [7] на ряді RTAEI за випусками. Імітаційне дослідження організовано за схемою ADEMP [6]: три профілі даних (збалансований, «маскування» слабкої надійності, слабкий допоміжний критерій), по 200 повторень, ступінь узгодженості з еталоном — коефіцієнт рангової кореляції Спірмена. Реальний кейс — застосунок зі списком задач із 12 засіяними дефектами, шість випусків, Playwright і Puppeteer, 120 запусків. Програмний засіб (Python, Streamlit, SQLite) перевірено 38 автоматизованими тестами.

Результати та обговорення

Результати узгодженості з еталоном наведено в таблиці 1.

Таблиця 1

Коефіцієнт Спірмена з еталонним рейтингом за надійністю (середнє за 200 повторень)

Схема	Збалансований	Маскування	Слабкий допоміжний
RTAEI (p = 0)	0,77	0,59	0,63
Зважена сума, ті самі ваги	0,76	0,54	0,69
Рівні ваги, p = 0	0,63	0,28	0,41
Ваги AHP	0,71	0,42	0,63

RTAEI перевершує рівні ваги в усіх профілях і ваги AHP у двох профілях; відносно зваженої суми з тими самими вагами його перевага статистично підтверджена лише в профілі «маскування», а в профілі зі слабким допоміжним критерієм індекс поступається. Інструменти з незбалансованим профілем отримують нижчі місця, ніж за зваженої суми (на 0,09–0,21 місця). Значна частина переваги зумовлена нижньою межею ваги надійності: за $w_{min} = 0{,}15$ результат помітно гірший. Контрольні межі виявляють деградацію на 0,4 і більше за два

випуски у 77–80 % випадків (для 0,2 — 50 %) за частоти хибних попереджень 6,8 %. На реальних вимірюваннях конвеєр працює коректно: індекс різко знижується, коли зміна локаторів вимагає переписати всі сценарії. Кейс має два інструменти й засіяні дефекти, тому є демонстрацією, а не статистичним підтвердженням. Дані статистичного експерименту — модельні, тому висновки потребують перевірки на проєктах із природними дефектами.

Висновки

Запропоновано індекс RTAEI, що поєднує тестові метрики з прогнозним критерієм надійності та обмежує компенсацію слабкої надійності сильними допоміжними показниками. Методику реалізовано в програмному засобі й перевірено на модельних і реальних вимірюваннях. Подальші дослідження — перевірка на реальних проєктах із природними дефектами та порівняння з байєсівськими моделями росту надійності.

Список використаних джерел

1. Goel A. L., Okumoto K. Time-Dependent Error-Detection Rate Model for Software Reliability and Other Performance Measures. IEEE Transactions on Reliability. 1979. Vol. R-28, No. 3. P. 206–211. DOI: 10.1109/TR.1979.5220566
2. Yamada S., Ohba M., Osaki S. S. S-Shaped Reliability Growth Modeling for Software Error Detection. IEEE Transactions on Reliability. 1983. Vol. R-32, No. 5. P. 475–484. DOI: 10.1109/tr.1983.5221735
3. Garousi V., Mäntylä M. V. When and what to automate in software testing? A multi-vocal literature review. Information and Software Technology. 2016. Vol. 76. P. 92–117. DOI: 10.1016/j.infsof.2016.04.015
4. Greco S., Ishizaka A., Tasiou M., Torrisi G. On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness. Social Indicators Research. 2019. Vol. 141, No. 1. P. 61–94. DOI: 10.1007/s11205-017-1832-9
5. Luo Q., Hariri F., Eloussi L., Marinov D. An Empirical Analysis of Flaky Tests. Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE 2014). 2014. P. 643–653. DOI: 10.1145/2635868.2635920
6. Morris T. P., White I. R., Crowther M. J. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019. Vol. 38, No. 11. P. 2074–2102. DOI: 10.1002/sim.8086
7. Shewhart W. A. Economic Control of Quality of Manufactured Product. D. Van Nostrand. 1931. Vol. 9, No. 2. P. 364–389. DOI: 10.1002/j.1538-7305.1930.tb00373.x
8. Kuz M., Bezghachniuk Y., Horodenko T., Nezamai B., Oliinyk A. Methods for quality assessment of standardized software time indicators. Scientific Journal of the Ternopil National Technical University. 2026. Vol. 121, No. 1. P. 88–98. DOI: 10.33108/visnyk_tntu2026.01.088