DOI 10.37882/2223-2966.2023.10.39

АНАЛИЗ ДАННЫХ ТЕХНИЧЕСКОЙ ДОКУМЕНТАЦИИ В ОБЛАСТИ ИТ С ИСПОЛЬЗОВАНИЕМ ПРЕДВАРИТЕЛЬНО ОБУЧЕННОЙ ЯЗЫКОВОЙ МОДЕЛИ ТРАНСФОРМЕРА

ANALYSIS OF TECHNICAL DOCUMENTATION DATA IN THE FIELD OF IT USING A PRE-TRAINED TRANSFORMER LANGUAGE MODEL

F. Shcherbanich
O. Romashkova

Summary. The article is devoted to the problems of development and analysis of technical documentation in the field of IT projects. With the help of a large language model of a generative pre-trained transformer, the sources of scientific literature were analyzed. The following problems were identified: incorrectness and incompleteness of documentation, lack of document writing skills among responsible persons, rapid obsolescence of documentation, neglect of documentation, ambiguity and unstructured presentation. The article also highlights and examines the ways of solving these problems proposed by the authors of the analyzed publications.

Keywords: technical documentation, IT projects, data analysis, language model, generative pre-trained transformer.

Щербанич Филипп Егорович
ФГБОУ ВО «Российская академия народного хозяйства и государственной службы при Президенте РФ (РАНХиГС)» г. Москва
scherbanich@gmail.com
Ромашкова Оксана Николаевна
Доктор технических наук, профессор, профессор, Российская академия народного хозяйства и государственной службы при Президенте РФ (РАНХиГС)» г. Москва
ox-rom@yandex.ru

Аннотация. Статья посвящена рассмотрению проблем разработки и анализа технической документации в области ИТ-проектов. С помощью большой языковой модели генеративного предобученного трансформера был проведен анализ источников научной литературы. Выявлены следующие проблемы: некорректность и неполнота документации, отсутствие навыка написания документов у ответственных лиц, быстрое устаревание документации, пренебрежение к оформлению документации, неоднозначность и неструктурированность изложения. В статье также выделены и рассмотрены предложенные авторами анализируемых публикаций пути решения данных проблем.

Ключевые слова: техническая документация, ИТ-проекты, анализ данных, языковая модель, генеративный предобученный трансформер.

Введение

Техническая документация программных продуктов является одним из основных средств хранения и передачи информации о разрабатываемой информационной системе и помогает обеспечивать взаимодействие между разработчиками системы, инженерами-тестировщиками, менеджерами и остальными пользователями [1]. Создание и поддержание технической документации в актуальном состоянии обычно связаны с рядом проблем и сложностей.

Идентификация и каталогизация проблем, а также соответствующих решений, предложенных в научных публикациях, позволит сформировать обобщенные выводы [2]. Эти выводы предоставят специалистам важные знания по оптимизации работы с технической документацией без необходимости детального анализа множественных источников литературы [3].

Методика исследования

В работе выполнен систематический анализ современных научных литературных источников 2019–2023 годов издания, в которых проводятся исследования, связанные с выявления проблем создания и систематизации технической документации ИТ-проектов, а также анализируются практики для решения выявленных проблем.

Для поиска подходящих публикаций был использован интернет-сервис Google Scholar, со строкой поиска: (*«software» OR «technical» OR «project» OR «code») AND «documentation» AND («quality» OR «problems» OR «issues»*). В результате было получено примерно 363 000 результатов, из были изучены первые 200 ответов (20 страниц поисковой выдачи), поскольку первые результаты поиска в Google Scholar являются наиболее релевантными. Из данного набора были выбраны 24 научных публикации, которые явились наиболее полезными для достижения целей настоящего исследования (*Aghajani E. Software documentation: automation and challenges: дис. — Università della Svizzera italiana, 2020; Rios N. et al. Hearing the voice of software practitioners on causes, effects, and practices to deal with documentation debt // Requirements Engineering: Foundation for Software Quality: 26th International Working Conference, REFSQ 2020, Pisa, Italy; Behutiye W. et al. Towards*

optimal quality requirement documentation in agile software development: A multiple case study //Journal of Systems and Software. — 2022; Wilsson L. Automating and increasing efficiency of component documentation maintenance: A case study. — 2022; Osorio K., Rosero J. L., Ch R. P. R. Technical Writer: A proposal to improve quality and documentation in the agile methodology" Scrum" //KnE Engineering. — 2020; Horvath A. et al. Understanding How Programmers Can Use Annotations on Documentation //Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. — 2022; Hull M. F. The Role of Technical Communicators in Open-Source Software: A Systematic Review. — 2021; Tang H., Nadi S. Evaluating Software Documentation Quality //2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR). — IEEE, 2023; Mathrani A., Wickramasinghe S., Jayamaha N. P. An evaluation of documentation requirements for ISO 9001 compliance in scrum projects //The TQM Journal. — 2022; Ajam G. Quality of Application Programming Interfaces Documentation and Topics Issues Exploration: Duc. — UNSW Sydney, 2019; Tan W. S., Wagner M., Treude C. Detecting Outdated Code Element References in Software Repository Documentation //arXiv preprint arXiv:2212.01479. — 2022; Shevchenko B. Issues in communication during architecture design in modern software engineering: A Systematic Literature Review. — 2023; Cadavid H., Andrikopoulos V., Avgeriou P. Documentation-as-Code for Interface Control Document Management in Systems of Systems: A Technical Action Research Study // European Conference on Software Architecture. — Cham: Springer International Publishing, 2022; Rasool A. A Review on Software Architecture Documentation in Agile Development //LC International Journal of STEM (ISSN: 2708-7123). — 2021. — T. 2. — № 1; Gilmore C. API Documentation for Users and Developers: Duc. — WORCESTER POLYTECHNIC INSTITUTE, 2022; Machado L. S. et al. How Online Forums Complement Task Documentation in Software Crowdsourcing //Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops. — 2020; Treude C., Middleton J., Atapattu T. Beyond accuracy: Assessing software documentation quality //Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. — 2020; Rebai S. Commits Analysis for Software Refactoring Documentation and Recommendation: Duc. — 2021; Szatmáry P., Schweitzer M. K. Exploring the Impact of Open-Source Software Documentation on Contributors' Motivation and Participation. — 2023; Aghajani E. et al. Software documentation issues unveiled //2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE). — IEEE, 2019; Slaughter A.E. et al. Continuous integration, in-code documentation, and automation for nuclear quality assurance conformance //Nuclear Technology. — 2021; Ansari M.J. An evaluation on Automated Technical Documentation Generator Tools //The University of Calgary. — 2022; Mendes L., Cerdeiral C., Santos G. Documentation Technical Debt: A Qualitative Study in a Software Development Organization //Proceedings of the XXXIII Brazilian Symposium on Software Engineering. — 2019; Jayasuriya D. B., Perera I. Ontology based software design documentation for design reasoning //2019 Moratuwa Engineering Research Conference (MERCon). — IEEE, 2019).

Для определения наиболее подходящих публикаций учитывались их заголовки, аннотации, вступления и заключения. Были выделены два основных раздела анализа, а именно обозначенные в каждой публикации проблемы и предложенные авторами пути решения этих проблем.

В качестве метода анализа материала научной литературы был использован комбинированный метод автоматизированного анализа с помощью большой языковой модели генеративного предобученного трансформера (GPT-4) с последующей ручной проверкой полученных результатов. Большие языковые модели представляют собой последовательности алгоритмов машинного обучения, способные анализировать, генерировать и понимать человеческий язык на необычайно высоком уровне. Они тренируются на обширных наборах данных, состоящих из миллиардов слов, что позволяет им обнаруживать сложные зависимости и нюансы в текстовых данных [4].

Применение больших языковых моделей в исследовательских задачах связано с несколькими ключевыми преимуществами. Во-первых, они способны быстро обрабатывать и анализировать огромные объемы текстовой информации, что значительно ускоряет процесс исследования. Во-вторых, они обеспечивают высокую точность в извлечении информации, что может быть особенно ценно при работе с научными текстами, где каждое слово и контекст его использования имеют значение.

Однако, несмотря на автоматизированные возможности больших языковых моделей, ручная проверка результатов остается критически важной. С помощью данной проверки могут быть найдены неточности или ошибки, допущенные языковой моделью [5]. Такой комбинированный подход гарантирует наилучшую точность и надежность результатов, сочетая преимущества современных технологий с экспертным анализом исследователя.

Для выполнения исследования с использованием большой языковой модели, в первую очередь, в нее необходимо передать исходный материал, который нужно проанализировать, и на основе которого языковая модель будет предсказывать ответ [6]. В качестве такого исходного материала были использованы научные публикации, подлежащие анализу. Кроме того, к данному тексту также была применена специфическая подсказ-

ка: «*Analyze the text. Write point by point first the problems associated with the creation and support of technical documentation, which are described in the article, then write point by point the solutions that the author suggested to improve the technical documentation*». С помощью данной подсказки большой языковой моделью были выделены все необходимые для дальнейшей ручной обработки данные из текстов определенных литературных источников.

Результаты анализа найденной научной литературы

В результате проведенного анализа с использованием большой языковой модели генеративного предобученного трансформера, для каждой научной публикации был получен текст — список проблем и решений, определенных авторами анализируемых текстов. Данный текст был переведен на русский язык, после чего была выполнена дополнительная ручная проверка полученного результата.

На основе анализа проблем, описанных в рассмотренных 24-х научных публикациях, можно выделить следующие основные проблемы в разработки и использования технической документации:

1. **Некорректность и неполнота документации**. Неполнота и неоднозначность документации затрудняют понимание проекта как разработчиками, так и пользователями. Ошибки в документации могут вызвать проблемы в дальнейшей разработке проекта.
2. **Отсутствие у технических специалистов навыков написания документации**. Создание документации требует больших трудозатрат, и далеко не все разработчики обладают навыками эффективного создания документации, что влияет на ее качество и полезность.
3. **Устаревание документации**. Многие методы гибкой разработки, включая быстрое прототипирование и изменение требований, часто являются причиной быстрого устаревания документации. Проблемы, связанные с актуализацией информации, являются частой причиной, из-за которой техническая документация теряет свою пользу.
4. **Пренебрежение документацией.** При быстрой разработке и реализации проекта документация часто отставляется на второй план, что приводит к низкому качеству и неактуальности документов.
5. **Неоднозначность и неструктурированность документации.** Отсутствие ясности в документации и ее неструктурированность могут вызывать серьезные проблемы в понимании и использовании документов разработчиками и пользователями, что в конечном итоге может снизить ценность такой документации.

Проанализированные проблемы явно способствуют получению негативного опыта от работы с документацией ее пользователями, а также низкой эффективности работы команд разработчиков и других технических специалистов. В то же время, написание и поддержка документации требуют больших усилий и времени, которые могут быть направлены на разработку. Именно поэтому поиск решений указанных проблем требует особого внимания.

Рассмотрим основные предложения по улучшению качества документации и процесса ее создания, предложенные авторами рассмотренных публикаций. В ходе анализа данных решений можно сделать несколько основных выводов:

1. **Следует автоматизировать создание и обновление документации.** Это ключевой вывод, который повторяется во многих статьях. Использование инструментов для автоматического создания документации позволит существенно экономить время при наполнении документов, при этом результат работы систем автоматизации всегда предсказуем и соответствует ожиданиям пользователей, связанным с созданием документации продукта.
2. **Обучение команды методикам создания документации, а также участие всего рабочего коллектива в процессе документирования.** Развитие навыков и знаний сотрудников в области составления документации может существенно повысить ее качество и снизить трудозатраты на ее написание.
3. **Контроль и поддержание актуальности документации.** Регулярная проверка документации поможет гарантировать ее точность и высокое качество, а значит и удовлетворенность ее пользователей.
4. **Стандартизация процесса документирования.** Установление единых стандартов и правил для составления документации может помочь придать ей последовательность и качество, которое не ухудшается по прошествии времени.
5. **Включение технических писателей в процесс работы.** Важность этого шага подчеркивается многими публикациями. Хотя задача таких специалистов в основном заключается в том, чтобы создавать документацию, они также могут вносить ценные предложения по улучшению и оптимизации процесса документирования.
6. **Использование эффективных платформ и инструментов для хранения и обмена документацией.** С помощью подобных платформ существенно улучшается опыт взаимодействия пользователей с документами.

В центре предложенных решений находится концепция создания культуры документации, когда документа-

ция рассматривается как полноправная и важная деталь процесса разработки. Участники процесса разработки должны ценить и улучшать процесс создания и совершенствования документации как средства обеспечения качества программных продуктов [7].

Все рассмотренные методы и подходы к решению проблем документации указывают на тесную связь между качеством документации и успешностью проекта. Отсутствие должной внимательности к документации может привести к проблемам в разработке, внедрении и поддержке функционала программного продукта, что в конечном итоге может оказать негативное влияние на развитие проекта.

Выводы

В результате проведенного исследования были выделены ключевые проблемы, связанные с потребностями пользователей и специалистов в области создания технической документации в процессе разработки и сопровождения ИТ-проектов, а также методы решения данных проблем. Кроме того, мы убедились в том, что большая языковая модель предобученного генеративного трансформера GPT-4 хорошо справляются с анализом больших текстов и выделением из них основных тезисов.

В результате анализа были выделены следующие основные проблемы, с которыми сталкиваются специалисты, работающие с технической документацией: некорректность и неполнота представляемой информации, недостаточный уровень профессионализма технических специалистов в области создания документации, проблемы своевременного обновления материалов и недооценка важности документации. Однако анализ показал, что имеются эффективные методы решения выявленных проблем: автоматизация, стандартизация, профессиональное обучение и интеграция технических писателей в команды разработки.

Кроме того, авторы научных публикаций зачастую подчеркивают то, что важность технической документации в некоторых организациях до сих пор остается недооцененной. Для повышения качества и актуальности документации проектов необходимо улучшать процессы ее создания, а также внедрять механизмы получения обратной связи от конечных пользователей.

Заключая, следует подчеркнуть, что, несмотря на важность документации в разработке программных продуктов, сегодня мы сталкиваемся с рядом сложностей, которые могут негативно влиять на ее качество и актуальность. Однако, исходя из проведенного анализа, есть все возможности для устранения этих трудностей, что, в конечном итоге, поспособствует повышению эффективности работы различных ИТ-организаций.

ЛИТЕРАТУРА

1. Chomal V.S., Saini J.R. Software template for evaluating and scoring software project documentations // International Journal of Computer Applications. — 2015. — Т. 116. — №. 1.
2. Gaidamaka, Y.V., Romashkova, O.N., Ponomareva, L.A., Vasilyuk, I.P. Application of information technology for the analysis of the rating of university // В сборнике: CEUR Workshop Proceedings 8. Сер. «ITTMM 2018 — Proceedings of the Selected Papers of the 8th International Conference «Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems»» 2018. С. 46–53.
3. Ромашкова О.Н., Пономарева Л.А., Василюк И.П. Применение инфокоммуникационных технологий для анализа показателей рейтинговой оценки вуза // В книге: Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем. Материалы Всероссийской конференции с международным участием. 2018. С. 65–68.
4. Brown T. et al. Language models are few-shot learners //Advances in neural information processing systems. — 2020. — Т. 33. — С. 1877–1901.
5. Muhlgay D. et al. Generating Benchmarks for Factuality Evaluation of Language Models //arXiv preprint arXiv:2307.06908. — 2023.
6. Ромашкова О.Н., Ермакова Т.Н. Применение инфокоммуникационных технологий для анализа показателей качества обучения образовательного комплекса // В сборнике: Технологии информационного общества. X Международная отраслевая научно-техническая конференция: сборник трудов. 2016. С. 388–389.
7. Ромашкова О.Н., Чискидов С.В. Методологии и технологии проектирования информационных систем // Учебно-методическое пособие / Часть 1. Москва, 2020.

Журнал «Современная наука: актуальные проблемы теории и практики»