

УТВЕРЖДАЮ
Директор Федерального
государственного учреждения
«Федеральный исследовательский
центр Институт прикладной
математики им. М.В. Келдыша
Российской академии наук»



_____ А.И. Аптекарев

5 апреля 2018 г.

ОТЗЫВ

ведущей организации на диссертационную работу Луцива Дмитрия Вадимовича на тему «Поиск неточных повторов в документации программного обеспечения», представленную на соискание ученой степени кандидата физико-математических наук по специальности 05.13.11 — «математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Практическая значимость и актуальность работы

Диссертационная работа Луцива Дмитрия Вадимовича посвящена вопросам, связанным с созданием и сопровождением документации программного обеспечения. Документация является важным аспектом создания и сопровождения программного обеспечения. Обсуждение принципов создания и сопровождения документации началось практически одновременно с созданием первых программ. Однако до сих пор не выработан единый подход к написанию программной документации. Для создания документаций программного обеспечения используются различные подходы, инструменты, методики. В результате даже в одном проекте могут использоваться разные подходы к созданию документации разными разработчиками.

Особо остро проблемы с документацией возникают при развитии и сопровождении программного обеспечения, когда после изменений исходного кода изменения документации не производятся или производятся, но не во всех частях документации. В результате документация не отражает актуального состояния программного обеспечения. Инструменты и методы, которые позволяют автоматизированно поддерживать документацию в самосогласованном и актуальном состоянии, крайне важны и найдут широкое применение.

Поэтому актуальность и практическая значимость диссертационной работы Луцива Д.В. не вызывает сомнений.

Основные научные результаты

Диссертация Луцива Д.В. состоит из введения, пяти глав, заключения, списка литературы и одного приложения.

Во **введении** обосновывается актуальность темы, определяются цели и задачи работы.

В **первой главе** приводится обзор подходов к структуризации документации программного обеспечения, описываются причины возникновения как точных, так и неточных повторов в программном обеспечении. Рассматриваются существующие подходы к поиску текстовых повторов. Дополнительно рассматриваются средства разметки электронных документов, которые могут использоваться в документации программного обеспечения.

Так же рассматривается документация нескольких доступных проектов для оценки наиболее употребительного размера документов, а так же для выбора набора документов, который в дальнейшем будет использоваться для оценки разработанных методов. Отдельно описаны известные методы и технологии, которые используются в диссертационной работе.

Во **второй главе** математически строго определяется понятие неточных повторов и приводится новый разработанный автором алгоритм поиска неточных повторов, основанный на сторонних методах поиска точных клонов, а так же приведено доказательство корректности предложенного алгоритма.

В работе дается два определения неточных повторов: одно основано на известном ранее неформальном определении, второе — предложенное автором более универсальное определение. Приведено утверждение о взаимосвязи этих определений.

В завершении главы приводятся результаты экспериментальных исследований реализации предложенного алгоритма поиска неточных повторов. Показано, что экспериментальная реализация алгоритма обрабатывает документы в среднем за 73 секунды, что подтверждает применимость алгоритма к реальной программой документации. Дополнительно приводятся оптимизации алгоритма, направленные на сокращение объема выдачи.

В **третьей главе** предложена новая методика интерактивного поиска неточных повторов, основанная на разработанном автором алгоритме поиска неточных повторов по образцу. Методика заключается в циклическом выполнении пользователем двух шагов. На первом шаге по автоматически построенной карте повторов пользователь выбирает часто встречающийся фрагмент текста, анализируя его семантический смысл. На втором шаге пользователь анализирует результат автоматического поиска неточных повторов выбранного фрагмента текста, при необходимости корректируя его. Данные шаги повторяются по необходимости. В результате формируется список неточных повторов в программной документации.

Для построения карты повторов используется известный инструмент Clone Miner. А для поиска неточных повторов выбранного фрагмента используется разработанный автором алгоритм поиска по образцу, подробно описанный в данной главе,

опирающийся на определения неточных повторов, данных во второй главе. В главе так же приводится подробное доказательство полноты предложенного алгоритма.

В завершении главы приводятся результаты экспериментальных исследований реализации предложенного алгоритма поиска по образцу. По результатам исследований сделан вывод, что прямолинейная реализация алгоритма не обладает необходимой скоростью работы. Поэтому предложены оптимизации, существенно сократившие время работы реализации до приемлемого времени выполнения одного поиска неточных повторов около 14 секунд, что позволяет применять разработанный алгоритм на практике. Доказано, что предложенные оптимизации не нарушают полноту алгоритма.

В **четвертой главе** описан новый метод улучшения программной документации на основе неточных повторов. В методе в начале используются алгоритмы поиска неточных повторов, рассмотренные в предыдущих главах. Затем найденные неточные повторы анализируются, приводятся к единому виду. После этого, если позволяет формат документации, по неточным повторам строятся текстовые шаблоны, которые заменяют повторы. В результате использования метода улучшается структура документации (одинаковые понятия описываются одинаковыми словами), документация приводится к единому стилю, снижается число орфографических и грамматических ошибок.

В **пятой главе** приводится описание разработанного автором программного инструмента Duplicate Finder, основанного на разработанных автором алгоритмах, методика и методах.

Достоверность основных положений и результатов работы

Все результаты, выводы и рекомендации диссертационной работы достаточно полно и хорошо аргументированы.

Диссертация имеет ясную и логичную структуру. Выбор направления исследования основан на тщательном анализе предыдущих работ и проблем в области создания и сопровождения программной документации. Это позволило прийти к постановке задачи, которая представляет интерес как с теоретической, так и с практической точки зрения.

В процессе работы диссертантом были использованы методики и подходы, разработанные предшественниками в области поиска точных повторов в программной документации, и предложены новые собственные понятия и алгоритмы поиска неточных повторов. Корректность и полнота алгоритмов доказана.

Результаты работ неоднократно докладывались и обсуждались на российских и международных конференциях и семинарах. Основные результаты диссертации опубликованы в 8 научных работах, 5 из которых зарегистрированы в РИНЦ. 3 статьи опубликованы в журналах из «Перечня российских рецензируемых научных журналов, в которых должны быть опубликованы основные научные результаты диссертаций на соискание учёных степеней доктора и кандидата наук» и 3 статьи в изданиях, входящих в базы цитирования Scopus и Web of Science.

Практическая ценность и рекомендации по использованию результатов диссертационной работы

Результаты работы рекомендуются к использованию в Институте прикладной математики им. М.В. Келдыша РАН, Институте программных систем им. А.К. Айламазяна РАН, Институте системного программирования РАН и других организациях, и могут найти применение при разработке и сопровождении программного обеспечения, а также разработке программного инструментария для написания и сопровождения документации программного обеспечения.

Замечания

Диссертация имеет логичную структуру и написана понятным языком. Однако, в тексте имеются отдельные опечатки, неудачные формулировки и недоработки.

1. В различных частях текста используются разные обозначения фрагмента текста: то «d», то «fg», что затрудняет чтение.

2. На стр. 46 в определении 2.1.3. указано для « $\forall k \in \{1, \dots, N-1\}$ », хотя, очевидно, должно быть « $\forall k \in \{1, \dots, \#G_1\}$ ».

3. На стр. 47 в замечании 2.1.2 говорится, что «... $\langle VG, VG' \rangle = \dots$ также является вариативной группой в том случае, если она удовлетворяет определению 2.1.3», что является бессодержательным утверждением типа « $\langle VG, VG' \rangle$ является вариативной группой в том случае, если она удовлетворяет определению вариативной группы».

4. На стр. 49 второй раз встречаются замечания 2.1.1 и 2.1.2, которые должны иметь номера 2.1.3 и 2.1.4.

5. На стр. 49 приведено не совсем корректное замечание 2.1.2 (которое должно иметь номер 2.1.4). Если трактовать это замечание в широком смысле на любую последовательность символов, то оно не будет верным. Но если ввести некие дополнительные ограничения на текст документа и архетип, то оно, возможно, станет верным. Однако, в дальнейшем данное замечание в работе не используется, поэтому эта недоработка не отражается на общих результатах работы.

6. На стр. 51 приведено «очевидное» утверждение 2.1.1 об взаимосвязи двух определений неточных повторов (определения 2.1.9 и 2.1.11) без доказательства. Хотя это ключевое утверждение не является тривиальным из-за разных понятиях, используемых в определениях. Однако его несложно доказать самостоятельно.

7. На стр. 51 в определении 2.1.1 дважды описано одно и то же: « $\langle G_1, \dots, G_n, G'_1, \dots, G'_m \rangle$ или $\langle G_1, \dots, G_n, G'_1, \dots, G'_m \rangle$ », хотя, очевидно, должно быть « $\langle G_1, \dots, G_n, G'_1, \dots, G'_m \rangle$ или $\langle G'_1, \dots, G'_m, G_1, \dots, G_n \rangle$ ».

8. На стр. 52 опечатка в слове «неточных» (t → e).

9. На стр. 71 в доказательстве утверждения 3.3.5 перепутаны верхние индексы: «Но если $w_2' \subset w_2''$, то очевидно, что $\forall g \in G |g \cap w_2''| < |g \cap w_2'|$ », а должно быть «Но если $w_2'' \subset w_2'$, то очевидно, что $\forall g \in G |g \cap w_2'| < |g \cap w_2''|$ ».

10. К сожалению, автор лишь в реализации алгоритмов касается проблемы анализа документации, написанной с использованием какого-либо языка разметки (например, DocBook, JavaDoc или Markdown). Было бы интересно на основе проделанной работы разработать методы и алгоритмы поиска неточных повторов,

которые бы существенно опирались на разметку документации, что, возможно, повысило бы точность результатов. Однако это направление выходит за рамки поставленной задачи.

Заключение

Отмеченные недостатки не оказывают существенного влияния на общий уровень и основные результаты диссертации. Диссертационная работа Луцива Д.В. на тему «Поиск неточных повторов в документации программного обеспечения» является законченной научно-квалификационной работой, содержит новые научные и практические результаты, связанные с решением актуальной задачи, имеющей большое значение для развития методов разработки и реализации математического и программного обеспечения вычислительных машин, комплексов и компьютерных сетей. Название диссертации соответствует основному содержанию диссертации. Автореферат достаточно полно отражает содержание работы.

Таким образом, диссертационная работа соответствует паспорту специальности 05.13.11, удовлетворяет требованиям п.9 Положения о присуждении ученых степеней, утвержденного постановлением Правительства РФ от 24.09.2013 № 842 (ред. от 30.07.2014, от 21.04.2016), а сам автор заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 05.13.11 — «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Работа была заслушана и одобрена, отзыв был рассмотрен и утвержден на семинаре Отдела инструментального и прикладного программного обеспечения ИПМ им. М.В. Келдыша РАН, протокол № 2 от 13 марта 2018 г.

Отзыв подготовил
кандидат физико-математических наук
(специальности 05.13.11 — «Математическое
и программное обеспечение вычислительных
машин, комплексов и компьютерных сетей»),
старший научный сотрудник
Федерального государственного учреждения
«Федеральный исследовательский центр
Институт прикладной математики
им. М.В. Келдыша Российской академии наук».

125047, г. Москва, Миусская пл., д.4

Сайт: <http://www.keldysh.ru>

Тел.: +7 (499) 220-78-20

Эл. почта: yuklimov@keldysh.ru



Юрий Андреевич Климов