На правах рукописи

Калачёв Ярослав Борисович

Автоматизированный контроль качества текстов проектной документации на предприятиях топливно-энергетического комплекса

Специальность: 05.13.12 – Системы автоматизации проектирования (в промышленности)

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург – 2015

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего профессионального образования «Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ)

Научный руководитель:	Кандидат технических наук, доцент, доцент кафедры «Информационные технологии и автоматизированные системы» Московского института электроники и математики НИУ ВШЭ Клышинский Эдуард Станиславович
Официальные оппоненты:	Доктор технических наук, профессор, директор по развитию бизнеса ООО «Витте Консалтинг» (ГК ЗАО «Ай-Теко», г. Москва) Сарафанов Альберт Викторович
	Кандидат технических наук, доцент кафедры Государственного казённого образовательного учреждения высшего профессионального образования «Академия Федеральной службы охраны Российской Федерации» (г. Орел) Скурнович Алексей Валентинович
Ведущая организация:	Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Тверской государственный технический университет» (г. Тверь)

Защита состоится 25 июня 2015 г. в 15 часов 00 минут на заседании диссертационного совета Д 212.238.02 при Санкт-Петербургском государственном электротехническом университете «ЛЭТИ» им. В.И. Ульянова (Ленина) по адресу: 197376, г. Санкт-Петербург, ул. Профессора Попова, д. 5.

С диссертацией можно ознакомиться в библиотеке университета и на сайте www.eltech.ru.

Автореферат разослан 24 апреля 2015 г.

Ученый секретарь
диссертационного совета Д 212.238.02
к.т.н., доцент Сафьянников Н.М.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Россия является крупной энергетической державой, энергетика которой играет важную роль в поддержании надежной экономической ситуации и укреплении места России на международной арене. Топливно-энергетический комплекс (ТЭК) РФ отвечает за добычу и переработку топлива, производство и распределение электроэнергии. Особенностью ТЭК РФ является его прямая взаимосвязь с государственными структурами, сложная структура отраслей промышленности и географический разброс предприятий. Всё это требует постоянного проектирования и разработки новых методов и систем с учетом взаимосвязи различных отраслей.

Проектирование объекта или системы начинается с документа, описывающего назначение, условия эксплуатации и требования к выходным параметрам, называемым техническим заданием (ТЗ). В соответствии с работами Норенкова И.П.: «Разработку ТЗ на проектирование называют внешним проектированием, а реализацию ТЗ – внутренним проектированием». Внешнее проектирование системы представляет собой описание предметной области, формирование основного назначения системы и определение требований к системе и разработке ТЗ. Результатом внутреннего проектирования является набор проектной документации (ПД), отвечающей требованиям ТЗ.

Развитые стандарты разработки (например, CALS-технология) требуют оформления полной документации. Основные спецификации представлены проектной, технологической, производственной, маркетинговой, эксплуатационной документацией. На этапах утверждения эскизного проекта, утверждения технического проекта, разработки документации и приемочного испытания опытного образца (или испытаний программы) разрабатываются текстовые документы, такие как пояснительная записка (ПЗ), инструкция (И) и программа и методика испытаний (ПМ), которые должны использоваться для отслеживания общего прогресса и контроля качества. Проверка выполнения требований ТЗ в ходе проектирования помогает вовремя выявить несоответствия в разработке.

Особенностью документации в ТЭК, разрабатываемой в ходе внутреннего проектирования, является использование знаний из нескольких предметных областей. В ходе разработки ситуация мягче: для выполнения многопрофильных задач пишутся собственные монотематичные ТЗ и документация. В связи с этим для проверки ПД по большому проекту необходима работа группы специалистов, чьи области знаний полностью покрывают тематику проекта. На практике не всегда имеется возможность выделить необходимое количество ресурсов на проверку отчетов. Зачастую проверкой занимается один человек, проверяющий лишь общую тематику и разделы отчета, анализируя основные понятия и данные, что снижает качество приемки. В связи с этим возникают задачи разработки методов автоматизации приемки текстовой ПД, позволяющих на основе анализа ТЗ проверить полноту покрытия списка требований ТЗ в отчетной документации и сократить временные затраты на такую проверку. Эта задача решается путем разработки автоматизированной системы поддержки процессов приемки. Такая система должна выявлять блоки итоговой документации, тематически сходные с ТЗ, выполнять визуализацию данных, необходимых для приня-

тия решения, являться дополнительным инструментом первоначальной проверки, выявляющим грубые ошибки и нарушения в тексте и содержании текстовой ПД. Ее внедрение должно повысить качество сдаваемой ПД.

Объектом исследования являются техническое задание, выработанное в ходе внешнего проектирования, и текстовые документы проектной документации, в частности, пояснительная записка, программа и методика испытаний, а также инструкция, разрабатываемые в ходе внутреннего проектирования на предприятиях топливно-энергетического комплекса.

Предметом исследования является метод приемки проектной документации, разработанной на этапах внутреннего проектирования на предприятиях топливно-энергетического комплекса, основанный на определении полноты изложения требований, представленных в техническом задании.

Целью диссертационной работы является снижение временных затрат на проверку текстовой проектной документации на завершающих этапах внутреннего проектирования за счет автоматизации оценки степени ее соответствия техническому заданию. Для реализации поставленной цели необходимо решить следующую **общую задачу**: необходимо разработать автоматизированный метод определения степени покрытия списка требований технического задания в текстах проектных документов.

Для этого необходимо решить следующие частные научные и практические **задачи**:

1. разработка алгоритма выделения требований из текстов технического задания с использованием лексико-синтаксических шаблонов;
2. разработка метода поиска описаний выделенных требований в тексте проектного документа и метода определения степени покрытия списка требований технического задания;
3. разработка метода визуализации полученных результатов;
4. проектирование и разработка программного комплекса для анализа отчетных документов на полноту описания;
5. проведение вычислительных экспериментов по тестированию разработанного метода и определению границ применимости метода.

Методы исследования. При решении поставленных задач использовались основные методы теории формальных языков, компьютерной лингвистики, теории дискретной математики, алгоритмы и методы обработки данных, а также методы объектно-ориентированного программирования.

На защиту выносятся следующие основные положения.

1. Метод автоматического поиска в тексте технического задания описания требований, относящихся к различным предметным областям.
2. Метод автоматического поиска и визуального отображения найденных требований в текстах проектной документации, снижающий временные затраты на принятие решения экспертами предприятий ТЭК в ходе приемки текстов проектной документации.

Одной из задач экспертов, принимающих текстовую документацию, подготовленную в ходе внешнего проектирования системы, является проверка ее

соответствия поданному на вход этапа проектирования ТЗ. Увеличение объема отчетов приводит к нелинейному росту сложности приемки и вероятности пропуска ошибок. Существующие автоматические методы работы с текстовой документацией не позволяют быстро перестраиваться на новые предметные области, характерные для ТЭК. В связи с этим возникает необходимость разработки новых методов определения полноты текстовой проектной документации.

Из разрешения полученного противоречия вытекает **научная новизна диссертационной работы**: в диссертационной работе разработан новый метод определения степени покрытия списка требований из ТЗ в текстах проектной документации на завершающих этапах внутреннего проектирования, учитывающий особенности документов ТЭК. В отличие от существующих, предложенный метод использует в качестве входных данных только тексты технического задания и проектной документации и не нуждается в предоставлении дополнительных словарей и баз данных. Разработанный метод визуализации результатов проверки проектной документации позволяет сэкономить время лица, принимающего решения (ЛПР), на поиск и проверку выделенных требований в ходе рассмотрения и утверждения эскизного или технического проекта.

Практическая значимость результатов. Предложенный метод может применяться как часть системы безбумажного документооборота технической документации и использоваться ЛПР для принятия решения о приемке или доработке документации, разработанной в ходе внешнего проектирования. В соответствии с предложенным методом был разработан программный комплекс на языке C++, использовавшийся для оценки корректности получаемых результатов. Практика использования программного комплекса показала применимость разработанного метода к решению задач, возникающих в крупных организациях. Проведенные эксперименты и полученные результаты подтверждают точность и работоспособность метода при нахождении требований ТЗ в отчетном документе.

Соответствие паспорту научной специальности. Основная область исследования соответствует паспорту специальности 05.13.12 – «Системы автоматизации проектирования (по отраслям)», в частности, пункту 4 – «Разработка принципиально новых методов и средств взаимодействия проектировщик – система» и пункту 7 – «Разработка научных основ построения средств автоматизации документирования, безбумажного документооборота, процессов работы электронных архивов технической документации, взаимодействия с изготовителем и потребителем изделий».

Апробация работы. Основные положения и материалы диссертационной работы докладывались на научно-практическом семинаре «Новые информационные технологии в автоматизированных системах» (2012, 2013, Москва) вне основной программы, а также на ежегодной научно-технической конференции студентов, аспирантов и молодых специалистов МИЭМ НИУ ВШЭ (2012, 2013, 2014, Москва).

Достоверность и обоснованность полученных результатов подтверждается:

- корректностью использования математического аппарата и методов испытаний;
- апробацией и публикациями основных результатов исследований;
- результатами внедрения разработанных методов и рекомендаций в практику.

Реализация и внедрение результатов. Описанный в данной работе метод был реализован в виде автоматизированной системы поддержки лиц, принимающих решение, по приемке отчетных документов. Разработанная программная система внедрена в использование в ФГБУ "САЦ Минэнерго России".

Публикации. Всего автором опубликовано 5 научных работ, из них 2 в журналах из перечня ВАК и одна в журнале, индексируемом в Scopus.

Структура и объем работы. Диссертация содержит 131 страницу основного текста, 29 рисунков, 8 таблиц. Список литературы состоит из 146 позиций.

СОДЕРЖАНИЕ РАБОТЫ

Во введении формулируются цели исследования, определяется научная новизна и практическая значимость результатов. Дается обоснование актуальности темы работы.

В первой главе рассматриваются задачи, возникающие при работе с такими полнотекстовыми документами, как ТЗ и отчетная документация.

Развитие ТЭК России ведет к постоянной разработке новых проектов, направленных на развитие и взаимосвязь отраслей. Как следствие, в ходе проектирования создаются новые документы, в том числе и рабочая документация, написанная в свободной форме на естественном языке. С увеличением количества текстовой информации растет и необходимость ее автоматической обработки и хранения. Данная задача решается с использованием систем PDM/PLM систем. Хотя современные достижения в области информационной поддержки процессов жизненного цикла изделий (ИПИ- или CALS-технологий), как и стандарты оформления документации (УСКД по ГОСТ 2.105-95), и создают основу для ручного контроля качества документации, они не позволяют решить задачу автоматизации определения полноты излагаемых в ней требований.

Особенностью ТЭК является разработка межотраслевых проектов. В этих случаях как ТЗ, поступающее на вход этапа проектирования системы, так и готовая ПД содержат в себе знания нескольких смежных предметных областей. Из-за этого для проверки документации по большому проекту необходима работа группы специалистов, чьи области знаний покрывают всю тематику проекта. На практике не всегда имеется возможность выделить столько ресурсов на проверку отчетов. При работе с документами ТЭК необходимо учитывать, что все документы имеют отраслевую принадлежность со своей собственной спецификой, терминологией и стандартами. Таким образом, при приемке ПД необходимо учитывать уникальные параметры каждой отрасли. Еще одной сложностью ТЭК является большой объем текстов ПД.

В отличие от приемки итоговой (например, эксплуатационной) документации, на этапе проектирования отсутствует готовая система. При проектировании разработанная документация является единственным результатом, а ошиб-

ки в ней приводят к ошибкам в проекте. Повышение качества ПД приводит к снижению затрат на разработку проекта. Как следствие невозможно корректное тестирование документации и существенно возрастают затраты на приемку.

Наиболее часто применяемым способом проверки технических документов на практике является ручной анализ с привлечением экспертов. Проверка многостраничного документа может занять у опытного эксперта, знающего предметную область и структуру документа, незначительное время. Однако любое изменение структуры или тематики приведет к увеличению трудозатрат эксперта. Как минимум, при повторной приемке исправленного документа эксперт должен снова ознакомиться со всем документом для того, чтобы найти все изменения в нем. В случае сложных проектных работ, в процессе которых порождается большой объем документации из различных областей знаний, для проверки выделяется группа совместно работающих экспертов. Однако увеличение количества экспертов ведет к еще большему увеличению трудозатрат в связи с необходимостью налаживания между ними взаимодействия и не всегда приводит к полному покрытию всех предметных областей.

Автоматическая обработка технической документации является одним из развивающихся направлений обработки текстов на естественном языке. Системы автоматизации используются на стадии проектирования для накопления в стандартизованной форме результатов труда разработчиков в установленном стандарте. Автоматизированные системы поддержки электронных моделей изделия обозначаются термином PDM – системы управления данными о продукте. На сегодняшний день существует множество популярных продуктов, упрощающих создание, ведение и проверку документации: Technical Guide Builder, Arbortext, Atlassian Confluence и др. Использование данных систем облегчает работу по созданию качественной документации при разработке систем и программных продуктов.

Применение систем электронного документооборота (СЭД) для решения поставленной в диссертации задачи возможно при проверке полноты списка документации и заключается в сравнении текущего списка документации по проекту со списком требуемых документов. Однако средства автоматизации контроля содержательной полноты документации в современных СЭД не предусмотрены. С другой стороны, на этапах проектирования (в особенности в области ТЭК) описанные системы работают с документами, оформленными в соответствии с требованиями ГОСТ, следовательно, возможность обработки и проверки более свободной отчетной документации исключается.

В целях повышения эффективности работы экспертов (оцениваемой во времени на приемку ПД, а также точности принимаемых ими решений) необходимо разработать автоматизированную систему, помогающую экспертам в предварительной оценке проверяемых отчетов. Система должна находить описание требований в ТЗ, на основе которого проводилось проектирование, по ним – описание требований в текстовой ПД. Вся информация должна визуализироваться для того, чтобы ускорить процесс принятия решений. Эксперт должен получать информацию о степени покрытия списка требований в отчете.

Задача обработки технической документации ставилась уже неоднократно. Так, группа под руководством Невзорова В.Н. разрабатывала систему «Лота» для анализа документов, описывающих логику функционирования системы, но сейчас ее развитие прекращено. Другая система разрабатывается в ВолГТУ под руководством Заболеевой-Зотовой А.В. и Орловой Ю.А. Она выделяет из ТЗ основные параметры разрабатываемой системы и заносить их для дальнейшего анализа в заранее подготовленную фреймовую структуру. Коммерческая система ABBYY Intelligent Search может быть настроена на выполнение поставленной в диссертации задачи, однако ее настройка требует существенных финансовых и временных затрат, выливающихся в самостоятельный проект.

Рассмотренные системы опираются на использование онтологий. Но разработка онтологий является длительным и сложным процессом. Если предприятие работает в нескольких предметных областях, разработка онтологии для них потребует значительных ресурсов. В связи с этим в диссертационной работе ставится задача разработки нового метода, позволяющего проводить анализ полноты ПД с использованием только имеющихся в распоряжении текстов с привлечением открытых и легкодоступных словарей и справочников. При этом на начальных этапах задача может решаться методами автоматической обработки текстов, а после того как будет разработана онтология предметной области, можно будет перейти к ее использованию. В связи с этим требуется разработка комбинированного метода, сочетающего в себе все плюсы использования методов автоматической обработки текстов и методов представления знаний.

Постановка задачи приводит нас к необходимости использования методов анализа текстов на естественном языке. При работе с полнотекстовыми документами можно выделить несколько наиболее распространенных задач, применимых в данном случае. Задачей систем антиплагиата является нахождение сходных фрагментов в разных документах. Результаты описаны в работах групп Si, Lancaster, Gipp, Васина А.Д. и др. Но так как не предполагается, что документация цитирует ТЗ, данные методы применимы лишь косвенно. Метод шинглов для определения нечетких дублей документов применяется в поисковых системах для уменьшения отклика на запрос и для поиска дубликатов страницы. Согласно работам Зеленкова Ю.Г и Сегаловича И.В., метод шинглов используется для сравнения небольших фрагментов текста (до 1000 знаков), но при сравнении документов большого объема точность метода понижается. В целом метод также не подходит для решения поставленной задачи.

Методы кластеризации и классификации позволяют автоматически рубрицировать документы для поиска и хранения в СЭД (см. работы Маннинга и др. и Песковой О.В.). Они не могут использоваться здесь напрямую, т.к. задача выделения тематических кластеров здесь не стоит. Из описанных работ нас интересуют методы определения тематической близости документов и их фрагментов. Для них обычно используется текст документа целиком, однако они применимы и при сравнении фрагментов, при условии их корректного выделения.

Так как документ включает термины разных предметных областей, снижается вероятность найти их в словаре. В связи с этим использование стандарт-

ных этапов анализа (графематического и морфологического) должно быть ограничено, а использование синтаксического анализа нецелесообразно.

В связи с тем, что объектом исследования является именно текстовая ПД, необходимо применение методов автоматической обработки текстов и выделение значимых частей. Из них подходят две группы методов: методы выделения терминов и многословных конструкций и методы определения тематической близости текстов. В работах зарубежных (Ahmad K. и Gillam L.) и отечественных (Киселева М.В., Кочеткова Н.А.) авторов были проработаны вопросы выделения тематических терминов текста. Также проработано применение таких мер выделения важных слов и словосочетаний, как tf*idf, MI, t-score, LSA и др. Синтаксические анализаторы словарных помет позволяют работать с синонимами, сокращениями и пояснениями, указанными в тексте. Подобные работы ведутся как для открытых интернет-словарей, таких как Викисловарь (Смирнов А.В. и Крижановская Н.Б.), так и для словарей, составленных по коллекции документов. Основной задачей подсистем морфологического анализа является нахождение нормальной формы рассматриваемого слова. Однако для их работы нужны внешние словари, разработка и настройка которых на предметную область занимает время. Существующие словари не содержат части специальной лексики, а методом пополнения таких словарей является ручной ввод. Автоматическое заполнение подобных словарей стало возможно с появлением свободных вычислительных мощностей современных компьютеров, но сталкивается с проблемой уникальности и особенности некоторых языков.

Во второй главе описывается формальная основа метода.

Под качеством проектной документации (ПД) в данной работе будет пониматься наличие в ней описания всех требований, изложенных в ТЗ. То есть мы исходим не из полноты описания, а из процента упомянутых требований.

Исследование особенностей ТЗ и отчетных документов показало, что требования к результатам работы обычно описываются с использованием особых синтаксических конструкций. Принимая во внимание, что большинство документации ТЭК пишется с использованием отраслевых стандартов (к примеру, ОСТ 153-00.0-002-98), для анализа требований нужен не весь текст ТЗ. Кроме того, ТЗ содержит специфичные термины, используемые для описания проекта.

Определение полноты документа основывается на использовании значимых словосочетаний, которые показывают описание требований, и специализированных терминов, характеризующих текст ТЗ.

Представим текст как упорядоченное множество предложений: $\mathbf{t} = \langle \mathbf{s}_i \rangle$. Представим предложение как упорядоченное множество слов: $\mathbf{s}_i = \langle w_{ij} \rangle$. Под словосочетанием будем понимать упорядоченное множество слов: $\mathbf{c} = \langle w \rangle$.

Введем список словосочетаний-маркеров $M = \{\mathbf{c}\}$, вводящих требования к изделию, поставленные заказчиком. Маркеры выбираются экспертом с учетом тематики конкретного ТЗ. Так же введем множество $M_s = \{\mathbf{c}\}$, $M_s \subset M$, содержащее базовые маркеры, пригодные для любой тематики.

Предложение, в котором встречается маркер, расценивается как значимое. Т.е. предложение **s**, входящее в текст ТЗ, называется значимым, если $\exists \mathbf{c} \in M: \mathbf{c} \subset \mathbf{s}$. Несколько предложений, где одно или несколько из них являются

значимыми, называется значимым фрагментом: **f** = <**t**, *s*, *e*>, где **t** – текст, в который входит фрагмент, *s* – номер начального предложения фрагмента, *e* – номер последнего предложения фрагмента.

Эксперименты показали, что качество работы метода с документами ТЭК возрастает, если во фрагмент включается одно предложение до и после значимого предложения. Предыдущее предложение часто вводит определения или определяет общее направление и отрасль ТЭК, последующие расшифровывают требования и содержат значения конкретных атрибутов (например, напряжение или мощность). Параметры r_1 и r_2 алгоритма показывают размер значимого фрагмента вправо и влево от значимого предложения. Если ключевая фраза встречается в предложении, после которого идет перечисление, то выделяется это предложение и весь текст до конца перечисления.

Поскольку ТЭК России состоит из четырех основных отраслей и большого количества направлений, предварительное выделение специализированных терминов невозможно. При этом для повышения точности выделения значимых фрагментов текста необходимо выделять уникальные термины документа. Для этого используется мера странности. Пусть имеется набор текстов общей лексики (например, беллетристика), называемый контрастной коллекцией. Пусть также имеется набор текстов в заданной предметной области, называемый коллекцией предметной области. Тогда слова, редко встречающиеся в контрастной коллекции, но часто в коллекции предметной области, считаются терминами данной предметной области. Мера странности рассчитывается по формуле:

$$W = \frac{w_s/t_s}{w_g/t_g}, \quad (1)$$

где w_s – встречаемость слова в коллекции предметной области, w_g - встречаемость слова в контрастной коллекции, t_s –количество слов в коллекции предметной области, t_g –количество слов в контрастной коллекции. Терминами будем считать слова, для которых мера странности значительно больше единицы.

Для вычисления специализированных терминов документа производится двойная выборка кандидатов в термины, предложенная в работах Кочетковой Н.А. Кандидаты в термины приводятся к начальной форме и могут состоять только из существительных, прилагательных, причастий, порядковых числительных, предлогов и союза «и», а наречия и местоимения опускаются. Для них вычисляется мера странности по коллекции документов той же тематики, полученные кандидаты с малой странностью отбрасываются. Вторая выборка проводится по контрастной коллекции и образует список терминов, входящих в предметную область документа. Выделенные термины включаются в множество маркеров M, по ним ведется поиск значимых фрагментов.

Метод проверки полноты отчетной документации по ТЗ работает в 4 шага.

На *первом шаге* метода по тексту ТЗ ищутся ключевые фрагменты, к которым применяются приведенные выше правила. Каждый выделенный по правилам фрагмент из текста ТЗ $\mathbf{t_1}$ заносится в список F = {**f**}. Два значимых фрагмента могут быть объединены вместе, если их границы пересекаются или меж-

ду ними нет значимого текста: если $\mathbf{f}_m = <\mathbf{t}, s_1, e_1>$ и $\mathbf{f}_{m+1} = <\mathbf{t}, s_2, e_2>$: $e_1 >= s_2$, то $\mathbf{f}_m = <\mathbf{t}, s_1, e_2>$, а $\mathbf{f}_{m+1}$ удаляется.

На *втором шаге* проводится выделение признаков из значимых фрагментов из списка F, а также производится поиск терминов специализации текста. Ключевые фрагменты разбиваются на группы из *n* слов (*n*-граммы) и заносятся в список, куда попадают только *n*-граммы, которые включают находящиеся рядом друг с другом слова, не разделенные знаком. Параметр *n* показывает, длину извлекаемых *n*-грамм. Для значимых фрагментов рассчитывается вектор признаков: $\mathbf{a} = <(w, f)>$, где $w \in \mathbf{f}$ – *n*-грамма, а *f* – ее частота встречаемости.

Некоторые словосочетания, общие для всех отраслей, встречаются во всех текстах и не несут важной информации (к примеру, Минэнерго России). Из вектора признаков *n*-грамм отсеиваются термины, встречающиеся чаще p_1 и реже p_2 от встречаемости самой частой *n*-граммы (авторские особенности текста и шум). Оставшиеся *n*-граммы составляют множество векторов признаков ТЗ $VA_1 = \{\mathbf{a}\}$. Также для всех фрагментов создается множество их собственных списков *n*-грамм: $V = \{V_i\} = \{\mathbf{a}\}$. Схема разбора ТЗ представлена на Рис. 1.

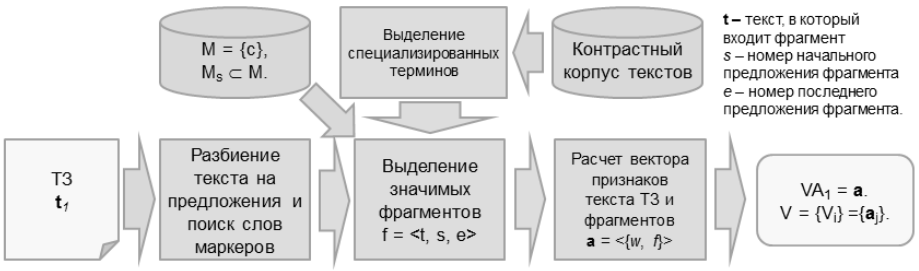

Рис. 1. Алгоритм выделения значимых частей ТЗ

На *третьем шаге* выбранные ключевые словосочетания ищутся в отчете. Находим те предложения $<\mathbf{s_j}>$ текста отчета $\mathbf{t_2}$, в которых встречаются *n*-граммы из $\mathbf{VA_1}$ (т.е. статистику нахождения упоминания требований в отчете). Для каждого $\mathbf{V_i}$ считаем меру mv_i, равную отношению количества найденных совпадений из $\mathbf{V_i}$ в тексте отчета $\mathbf{t_2}$ к количеству *n*-грамм в $\mathbf{V_i}$:

$$mv_i = m / n, \qquad (2)$$

где *m* – количество **f** таких что $\mathbf{f} \in \mathbf{V_i}$ и $\mathbf{f} \in \mathbf{t_2}$, а *n* – количество **f** в $\mathbf{V_i}$.

Поскольку требования, поставленные в ТЗ, могут находиться в различных абзацах отчета, то мера mv_i показывает пользователю, какие конкретно предложения с требованиями не были найдены в отчете. Позиции вхождений *n*-грамм $\mathbf{VA_1}$ и $\mathbf{V_i}$ в текст отчета $\mathbf{t_2}$ заносятся в списки $\mathbf{PA_1}$ и $\mathbf{P_i}$ соответственно.

Из текста отчета формируется список *n*-грамм с частотами их встречаемости $\mathbf{b} = <(w, f)>$ и вектор признаков ПД $VA_2 = \{\mathbf{b}\}$. На основе совпадающих *n*-грамм в VA_1 и VA_2 косинусная мера сходства документов:

$$v = \cos(\mathbf{VA_1}, \mathbf{VA_2}) = \frac{\quad}{\quad}. \qquad (3)$$

Вместо косинусной меры может использоваться одна из метрик, вычисляемых при помощи онтологии, например, среднее по минимумам расстояний между словами, вычисленными как минимальный путь в графе.

Текст отчета разбивается на абзацы, каждому из которых присваивается вектор признаков **b**, который формирует списки VO_j ($\mathbf{VO} = \{\mathbf{VO_j}\} = \{\mathbf{b}\}$). Зна-

чимость абзаца с номером *j* вычисляется как максимум косинусной меры сходства вектора **b** с векторами **a** или равна нулю, если его значимость ниже порога:

$$vo_j = \max_i \cos(\mathbf{a_i}, \mathbf{b_j}), \quad (4)$$

где $\mathbf{a}_i \in \mathbf{V_i}$, а $\mathbf{b}_j \in \mathbf{VO_j}$.

Значение *i* (номер значимого фрагмента ТЗ), при котором мера vo_j (рассчитанная по формуле 4) принимает наибольшее значение, показывает, описанию какого требования текста ТЗ соответствует *j*-й абзац текста отчета. Схема разбора отчетного документа представлена на Рис. 2.

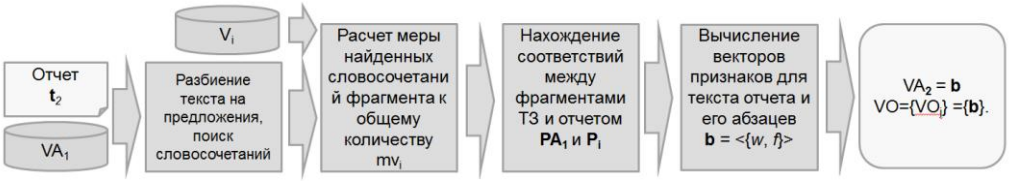

Рис. 2. Разбор текста ТЗ

На *четвертом шаге* метода ЛПР получает следующие данные:

- информацию о покрытии отчета фрагментами ТЗ в виде точечной диаграммы, построенной по полученным спискам $\mathbf{PA}_1$ и $\mathbf{P}_i$;
- отсортированный по возрастанию список мер mv_i и принадлежащим им фрагментам текста ТЗ;
- меру схожести документов v целиком и меру схожести по абзацам vo_j;
- позиции найденных в отчете упоминаний требований из ТЗ.

Руководствуясь результатами, пользователь принимает решение о пригодности отчета. Общий алгоритм работы пользователя состоит из следующих шагов. На вход алгоритма подается текст ТЗ и отчетный документ. Пользователь определяет параметры метода, пригодные для конкретного случая, добавляет специфичные словосочетания маркеров и выбирает термины текста ТЗ. Значение меры сходства в результате разбора документа должно находиться в пределах от **0,3** до **0,9** у близких по смыслу текстов и в пределах от **0,4** до **0,8** между документами ТЗ и принадлежащим им отчетам. Показателями пригодности текста по точечной диаграмме являются отношение выделенных фрагментов к количеству фрагментов и кучность выделения. Хорошим процентным показателем выделения предложений со словосочетаниями-маркерами из документов, написанных по ГОСТ, является 8-17% (70-80% на коротких технических записках). При большом проценте выделения значимых частей ТЗ (15% и более), но маленьком значении процентного отношения найденных ключевых словосочетаний в тексте отчета (15-20%), пользователь может просмотреть список мер *mv*, показывающий, какие требования ТЗ не представлены в отчете. Значение меры *mv* варьируется в пределах от 0 до 1, со средним показателем 0,6 для описанных в отчете требований, 0,2 и меньше, для плохо описанных. На основе результатов пользователь принимает решение о насыщенности документа.

Для принятия решения о необходимости проверки отчетного документа и поиска недостающих требований был разработан метод визуализации результатов. Принимая во внимание среднее количество символов и слов в предложении, был выбран максимальный размер фрагмента, пригодного для поставленной задачи. Размер ТЗ может варьироваться от 10 до 90 страничных текстов, на-

писанных по всем правилам ГОСТ. Можно предполагать, что верхняя граница символов в тексте ТЗ равна 162 000 знаков, а нижняя – 18 000 (С учетом, что страница содержит 1800 символов). Принимая средний размер предложения в 100 символов за фрагмент, можно подсчитать их минимальное и максимальное количество: 180-1 620.

Представим цвет каждой точки в виде аддитивной цветовой модели с переменными *R, G, B,* варьирующимися в диапазоне от 0 до 255. Введем начальный цвет (R_1,G_1,B_1), конечный цвет (R_f,G_f,B_f), число фрагментов ТЗ *f*, шаг d:

——— . (5)

Цвет фрагмента меняется в заданном диапазоне, где переменные R, G, B каждого следующего фрагмента повышаются на значение d. Поскольку структура взаимного расположения требований в ТЗ дублируется в отчетном документе, можно предположить, что чем ближе цвет на диаграмме, тем ближе фрагменты текста по значению. Резкая смена цвета может наблюдаться при переходе от одной части отчета к другой, если сначала описываются общие требования к системе, а в последующем тексте они последовательно расшифровываются. Во всех остальных случаях резкая и частая смена цвета означает непоследовательное изложение требований, либо чрезмерное их перемешивание.

В третьей главе рассматриваются практическая реализация и возможности использования систем документооборота.

Для оценки эффективности метода было разработано программное обеспечение (ПО) на языке C++ (используемая среда – Embarcadero C++ Builder XE2) для ОС Windows. Параметрами его работы являются размер значимого фрагмента, список словосочетаний-маркеров, термины предметной области и проценты отсечки слишком частотных или слишком редких значимых *n*-грамм.

Для уточнения данных параметров был проведен ряд экспериментов с документами ТЭК. В результате их анализа были выбраны параметры метода (приведены в таблице 1) и список словосочетаний-маркеров, представленный на Рис. 3. В качестве контрастной коллекции текстов была выбрана свободно распространяемая библиотека Мошкова (http://www.lib.ru/, 680 млн словоупотреблений), которая содержит в себе тексты, написанные литературным языком различными авторами. Это позволяет выделить разные авторские стили и исключить их из рассмотрения при анализе ПД.

Таблица 1. Параметры работы метода

Проценты отсечки (p_1 и p_2)		Количество слов (n)	Размер фрагмента (r_1 и r_2)	
85%	15%	2	1	1

Возможность	Обеспечивает	Предназначен для	Назначение
Должен	Состоит из	Реализовать	Необходимо
Имеет следующую структуру	Обусловливает необходимость	Обеспечит следующие возможности	Требуется
Осуществляемой	Основной целью	Требования	

Рис. 3. Словосочетания-маркеры

Система получает на вход ТЗ и отчет, проводит их разбор, получая списки значимых *n*-грамм, и частей ТЗ, визуализирует результат для ЛПР. На выход пользователю выдаются графики распределения найденного текста и стати-

стическая информация о найденных *n*-граммах. Основными задачами разработанной системы являются: обработка документов, хранение информации и вывод результатов. Структура разработанной программы представлена на Рис. 4.

Рис. 4. Общая структура программы

Реализация метода обеспечивается работой основных модулей:

- Модуль выделения терминов предметной области – выделяет термины на основе документа ТЗ и контрастной коллекции текстов.
- Модуль поиска ключевых фрагментов – выделяет фрагменты ТЗ, содержащие термины и словосочетания-маркеры.
- Модуль разбора отчетного документа – ищет *n*-граммы в тексте отчета.
- Модуль визуализации результатов – выводит результаты работы. строит графики, предоставляет оба текста и результаты выделения требований.

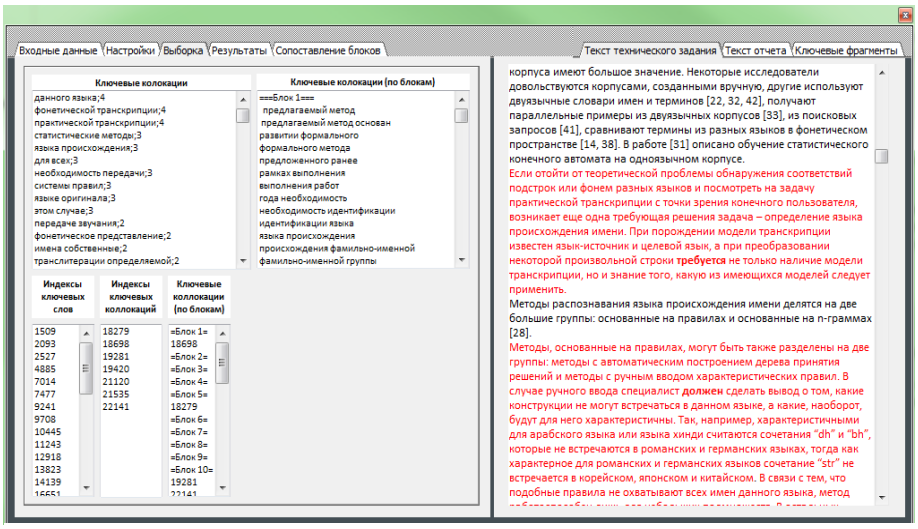

Рис. 5. Пример работы программы

Пользователь может сохранять и загружать настройки системы и создавать профили обработки определенных типов документов (Рис. 5). Всего создано 4 хранилища документов: база данных документов (хранит ТЗ и обработанные отчетные документы организации), хранилище коллекции документов (содержит контрастные коллекции), база данных терминов (хранит выделенные программой термины предметной области), база настроек метода (хранит пользовательские наборы настроек).

Анализ ПД, разрабатываемой в ходе проектирования, отличается от анализа текущего документооборота. В ходе проектирования на предприятиях ТЭК в ТЗ может ставиться задача проектирования комплексной системы, относящейся к нескольким предметным областям. При этом для производства пишется несколько ТЗ, описывающих свою предметную область. В связи с

этим актуальность разрабатываемого метода в области САПР, применяемых на предприятиях ТЭК, значительно выше, чем в области СЭД или анализа технической документации на производстве. В целом, СЭД позволяют организовать работу с документами в организации, но не участвуют в процессах проектирования, рассматриваемых в данной работе. При этом применение СЭД возможно для автоматизации процесса поиска и перемещения документации в организации. Предложенная система может быть встроена в качестве одного из модулей СЭД предприятия, но сложностью здесь является привлечение специалиста, имеющего опыт разработки подобных модулей, т.к. проектирование модулей для СЭД требует знания их архитектуры.

В четвертой главе описываются результаты экспериментов, подтверждающих качество работы метода. На 1-м этапе эксперты читали ТЗ и отчеты и высказывались о их содержании. Далее документы проверялись автоматически. На 2-м этапе проводилась перекрестная проверка текстов 6 ТЗ с текстами 9 отчетов. ТЗ 1-3 имеют близкую тематику, ТЗ 5 и 6 имеют близкую, но не связаны с 1-3, отчет 0 не имеет ничего общего с ТЗ 1-6, отчеты 3+ и 6+ были переписаны по требованию заказчика.

Результаты проверки приведены в Табл. 2. Результаты удачных проверок в корректной паре выделены темно-серым, а успешные проверки с другими отчетами – светло-серым фоном. Разработанный метод и ПО определили высокое качество отчетов, написанных для ТЗ 1-3 и 6. Результат работы системы для отчета 3 и 6 совпал с мнением заказчика. Отчет 0 не показал совпадений ни для одного из ТЗ. ТЗ 4 и 5 не предполагали подробного описания результатов работы и требований к ним. Также в ТЗ 5 требовалось дать рекомендации по улучшению изделия, что усложнило поиск соответствий. Отчет 4 содержал информацию по предметной области ТЗ 5, поэтому их сходство выше.

Таблица 2. Результаты кросс-проверки для предложенного метода

		Технические задания					
		1	2	3	4	5	6
Отчеты	1	0,521	0,157	0,192	0,032	0,025	0,072
	2	0,394	0,592	0,543	0,056	0,054	0,062
	3	0,37	0,39	0,158	0,05	0,049	0,05
	3+	0,494	0,45	0,535	0,045	0,051	0,054
	4	0,032	0,032	0,066	0,032	0,002	0,031
	5	0,032	0,009	0,02	0,307	0,057	0,095
	6	0,006	0,011	0,007	0,002	0,031	0,638
	6+	0,006	0,009	0,006	0,002	0,016	0,725
	0	0,011	0,043	0,035	0,006	0,006	0,017

Для повышения точности выделения ключевых фрагментов текста ТЗ была произведена еще одна серия экспериментов, связанная с выделением терминов предметной области. Исходя из результатов экспериментов, можно сказать, что предложенный метод выделяет корректную информацию, помогающую ЛПР принимать правильные решения о тематике ТЗ, изложенных в нем требованиях и полноте изложений этих требований в отчетных документах.

В следующей серии экспериментов проводилась проверка точности визуализации. В результате работы метода пользователь получает на выходе алго-

ритма 3 диаграммы: маркеры в ТЗ, ключевые *n*-граммы в отчете по всему тексту, ключевые *n*-граммы в отчете по фрагментам. Точечные диаграммы отчетных документов представлены на Рис. 6 и 7. Каждому фрагменту текста ТЗ сопоставлен цвет, отображавшийся и в диаграмме отчета для соответствующих фрагментов. Цвет меняется от синего к зеленому в зависимости от номера фрагмента. Блоки из компактно расположенных 5-10 цветных точек описывают заявленные в ТЗ требования. Отдельно стоящие цветные квадраты – единичная *n*-грамма в тексте. Основным критерием оценки полученных результатов является процентное отношение выделенных фрагментов (точек диаграммы) к общему количеству фрагментов.

На Рис. 6 видно, что цветные точки разбросаны по отчету, почти нет блоков больше 5 точек. На 130 000 знаков отчета было найдено лишь 470 групп, относящихся к ТЗ. Максимальная связная длина текста, имеющего отношение к одному из значимых фрагментов ТЗ – 700 символов. Т.е. отчет должен быть переработан разработчиком. Проверка эксперта подтвердила оценку программы.

На рис. 7 представлен качественно написанный отчет, в котором ключевые *n*-граммы встречаются везде, за исключением начала (содержание, авторы, введение). При длине отчета более 130 000 знаков найдено более 3500 групп.

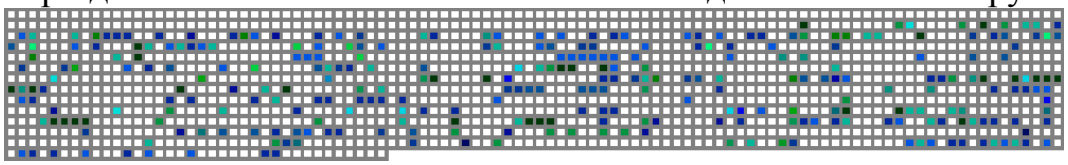

Рис. 6. Точечная диаграмма для неудачного отчета (17%)

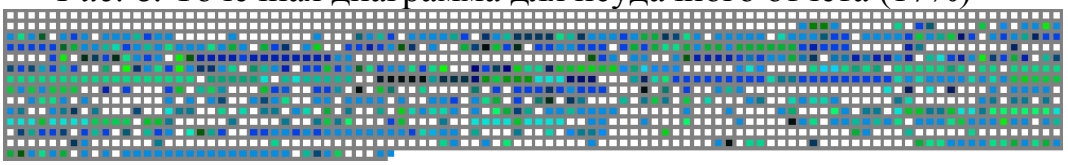

Рис. 7. Точечная диаграмма для качественно написанного отчета (66%)

Эксперименты подтвердили корректность разработанного метода визуализации и его применимость на практике. Для оценки точности выделения требований ТЗ методом было произведено сравнение с ручным выделением информации экспертами. На рис. 8 представлены требования ТЗ, выделенные экспертами (верхняя часть) и предложенным методом (нижняя часть). Текст ТЗ был разделен на несколько равных фрагментов, по 3 фрагмента на страницу документа, представленных полосами. Интенсивность цвета верхнего рисунка растет с количеством экспертов, отметивших данных фрагмент как содержащий требования (4 оттенка заливки: меньше 33% экспертов, от 33% до 66% экспертов, от 66% до 99% экспертов и 100% экспертов).

*Рис. 8.*Сравнение выделения фрагментов

Как видно из рисунка, метод чаще выделяет фрагменты, выделенные несколькими экспертами. В эксперименте были рассчитаны значения полноты и точности. Вычисление мер производилось на основе количества предложений:

- выделенных методом, но не выделенных большинством экспертов;

- выделенных как методом, так и большинством экспертов;
- не выделенных методом, но выделенных большинством экспертов.

Общее качество оценивалось при помощи F_1-меры, которая определяется как взвешенное гармоническое среднее точности P и полноты R:

——. (6)

Полученное в ходе экспериментов среднее значение F_1-меры для разработанного метода равняется 0,685. F_1-мера для результатов, показанных экспертами (относительно их консолидированного мнения) не превысила 0,6 ни в одном из экспериментов. То есть точность и полнота работы предложенного метода как минимум не ухудшает результатов оценки, а скорее улучшает их. При этом затраты времени на выделение терминов предметной области и выделение значимых фрагментов разработанным методом не превышает 8 минут, а ручное выделение экспертами в среднем занимало 30 минут.

Эффективность метода вытекает из времени работы пользователя. При учете норм РФ по чтению и обработке документации в 25 страниц в день, можно предположить, что для полной проверки отчета размером в 100 страниц потребуется 32 часа рабочего времени пользователя, а максимальное время работы предложенным методом составляет 1 час. То есть, при анализе неудачного отчета экономия времени ЛПР составит до 31 часа.

В заключении описываются основные результаты работы.

Разработанный метод определения полноты отчетной документации позволяет выделять значимые фрагменты ТЗ, содержащие условия разработки и находить их упоминание в отчете. Метод визуализации результатов сравнения текстов отчетной документации и ТЗ представляет результаты в виде точечной диаграммы, благодаря которой пользователь может оперативно принять решение о дальнейших действиях по проверке ПД. Предоставляемая пользователю информация позволяет найти описание выполнения каждого из выделенных требований в отчете, найти требования, которые не были упомянуты в отчете и показать взаимосвязь между абзацами отчета и фрагментами ТЗ. Разработанный метод уменьшает время работы пользователя, заменяя чтение всего документа на изучение значимых фрагментов. Выделение фрагментов не привязано к правилам русского языка, то метод может работать на европейских языках при правильном подборе словосочетаний-маркеров.

При развитии системы сохраненные позиции вхождений требований ТЗ и выделенные термины предметной области могут применяться для выделения стандартизированной конструкции, показывающей распределение требований ТЗ по документу. Конструкцию можно использовать для создания шаблонов автоматизации документирования в САПР, а выделенные термины – как базис для создания онтологии предметной области.

Метод реализован в виде ПО для ОС Windows и позволяет производить проверку полноты документов в основных текстовых форматах. Система не заменяет человека, а лишь помогает ему отбросить заведомо пустые тексты или специально увеличенные в размере тексты, не несущие смысловой и информационной нагрузки. На основании результатов экспериментов можно судить о

полноте и точности работы метода, о повышении эффективности работы ЛПР и уменьшении времени, затрачиваемого на проверку отчетного документа.

Развитием ПО является интеграция системы в системы документооборота организации для автоматической обработки документов при получении новой информации и последующей передачи ее пользователю.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

- разработан новый алгоритм выделения требований из текста технического задания с использованием лексико-синтаксических шаблонов;
- разработан метод поиска описаний выделенных требований в тексте проектного документа, учитывающий особенности выбранной предметной области;
- разработан метод определения степени покрытия списка требований технического задания, упрощающий проверку проектной документации, относящейся к нескольким предметным областям;
- разработан метод визуализации полученных результатов;
- разработаны рекомендации по применению указанных методов в ходе приемки проектной документации;
- спроектирован и разработан программный комплекс для анализа отчетных документов на предмет определения полноты описания требований технических заданий;
- путем проведения вычислительных экспериментов показана корректность разработанного метода, эффективность его применения на практике, определены границы применимости разработанного метода.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в изданиях, рекомендованных ВАК:

1. Калачёв, Я.Б. К вопросу об автоматизации проверки полноты отчетной документации / Э.С. Клышинский, Я.Б. Калачев, В.В. Жаднов // Информационные технологии в проектировании и производстве. – 2014. – №2.С. 68-72.
2. Калачёв, Я.Б. Методика автоматизации проверки полноты технической отчетной документации / Э.С. Клышинский, Я.Б. Калачев, В.В. Жаднов // Научно-техническая информация. – 2014. - Сер. 2: Информационные процессы и системы. № 5. – С.11-15

Другие статьи:

3. Kalachev, Ya.B. Method for visual representation of technical documentation’s completeness / Ya.B. Kalachev, E.S. Klyshinsky // Журнал «НаучнаяВизуализация». – 2014. - № 3. – С. 96-104.
4. Калачёв, Я.Б. Метод проверки содержательной полноты отчетной документации / Э.С. Клышинский, Я.Б. Калачев // Журнал «САПР и графика». -2013-№11. - С. 94-96.
5. Калачёв, Я.Б. Оптимизация временных затрат проверки полноты проектной документации на предприятиях топливно-энергетического комплекса / Я.Б. Калачёв // Журнал «Системный администратор». – 2015. - №1-2. С.126-131.

Подписано в печать 23.04.15. Формат бумаги 60x84/16.
Бумага офсетная. Печать офсетная. Печ. л. 1,0.
Гарнитура «Times New Roman».
Тираж 100 экз. Заказ № .

Отпечатано с готового оригинал-макета
в типографии Издательства СПбГЭТУ «ЛЭТИ»

Издательство СПбГЭТУ «ЛЭТИ»
197376, Санкт-Петербург, ул. Профессора Попова, д. 5.