УДК 81'322.2 : (047.31)

Э.С. Клышинский, Я.Б. Калачёв, В.В. Жаднов

Методика автоматизации проверки полноты технической отчетной документации*

Рассматривается новый метод автоматизации определения соответствия технического задания и итогового отчета в ходе его приемки. Предложенный метод позволяет экспертам получить предварительную оценку степени соответствия отчета техническому заданию. Используются выделение значимых фрагментов технического задания, поиск соответствующих им элементов отчета и проверка степени его покрытия. Разработанный метод, в отличие, например, от косинусной меры сходства, дает лучшее разделение отчетов по критерию хорошего и плохого изложения материала.

Ключевые слова: *информационные технологии, электронный документооборот, проверка документации, автоматическая обработка текстов.*

Оформление корректной документации при проектировании изделий является залогом успешного выполнения проекта. Качественно написанное техническое задание (ТЗ) на проект и проектная документация снижают шансы на срыв поставки изделия. Важную роль играет и отчет о выполненных работах, так как он позволяет повторить проделанные работы и (или) разобраться в них. Особую роль документация играет при работе в соответствии с принципами ИПИ (CALS)-технологии [1].

Существует несколько стандартов на оформление технической документации. В первую очередь – это ГОСТы, по которым оформляется документация в государственных учреждениях, на промышленных предприятиях и т.д. В качестве альтернативы можно привести стандарты International Standard Organization, которые дают рекомендации по составу документации. В области электротехники и телекоммуникаций – серия рекомендаций European Telecommunications Standards Institute, в области разработки программного обеспечения – рекомендации Rational Unified Process от IBM [2] и Microsoft Solutions Framework [3]. Хотя последние не утверждены в качестве государственных стандартов, договор может регламентировать работу в соответствии с этими рекомендациями или по стандартам предприятий, описывающим состав и содержание документации.

Текст отчета проверяется экспертами, что занимает много времени. Недобросовестный исполнитель может пытаться спрятать низкое качество отчета за его объемом, цитатами, уходом в смежную область, применением бюрократического стиля изложения текста и т.д. Для создания «первого эшелона обороны» в помощь экспертам необходимо создать автоматизированную систему, проверяющую степень соответствия отчетов или технической документации тексту ТЗ и помогающую принимать решения о необходимости проведения детальной экспертизы документации. Подобная система выявит наиболее очевидные несоответствия, которые должны быть проверены специалистом. На вход система должна получать тексты документов, содержащие постановку задачи и требования к проекту, а также отчетные документы. Выходом системы является оценка степени сходства и связности фрагментов текста, их смысловой нагруженности.

Для определения смысловой близости документов или их фрагментов применяется, например, метод шинглирования, который основывается на выделении последовательностей слов длины k с последующей оценкой вероятности совпадения текстов документов [5, 6]. Этот метод обладает высокой скоростью, однако, может быть применен лишь для выявления заимствованных фрагментов отчетов. Для сравнения текстов ТЗ и отчетов более подходят методы определения тематической близости документов. Пусть для двух документов вычислены векторы частот встречаемости слов **a** и **b**. Эти векторы определены на множестве всех слов, встречающихся в обоих документах. В этом случае косинусная мера сходства двух документов определяется следующим образом [7]:

$$\cos(\mathbf{a},\mathbf{b}) = \frac{\sum a_i * b_i}{\sqrt{\sum a_i * a_i} * \sqrt{\sum b_i * b_i}}$$

Развитием этого метода является использование векторов частот для словосочетаний различной длины. В этом случае точность определения степени сходства документов возрастает [8]. Косинусная мера сходства документов хорошо зарекомендовала себя при решении различных задач, но определяет лишь сходство тематики документов в целом.

К настоящему времени хорошо проработаны методы хранения технической документации с использованием PDM/PLM технологий [9]. Также существуют методики, учитывающие специфику

* Работа выполнена при поддержке РГНФ (грант № 12-04-00060) и РФФИ (грант № 11-01-00793).

обработки технических документов [10]. Современные достижения в области ИПИ-технологий дают основу для создания системы контроля качества документации, но не позволяют решить задачу определения ее полноты.

Еще одной группой методов, для которой проводятся исследования, является разработка формальных моделей текста документов [11], но для создания подобных систем необходимы большие онтологии предметных областей, не всегда доступные разработчикам. В настоящее время проводятся работы по автоматизированному составлению онтологий [12], однако и эти работы еще не доведены до программного обеспечения, удобного для применения. Наиболее разработанными в теоретическом и практическом плане являются работы в области формализации выделения спецификаций систем [13], однако и они далеки от широкого распространения.

Из сказанного становится очевидным, что существует потребность в создании теоретического метода автоматизации определения соответствия технического задания и итогового отчета, его практической реализации. Кратко изложим основные фрагменты предлагаемого нами метода.

Из текста ТЗ выделяются предложения, содержащие характеристики разрабатываемого изделия. Из выделенных фрагментов извлекаются все пары стоящих рядом слов (коллокации), незначимые коллокации удаляются.

Текст отчета разбивается на фрагменты, для которых также выделяется список коллокаций, после чего находится максимум меры сходства абзацев со значимыми фрагментами ТЗ. Это значение и будет мерой соответствия абзаца тексту технического задания. Полученный результат выдается лицу, принимающему решение (ЛПР), в визуальной форме. С его помощью ЛПР определяет меру соответствия отчета и технического задания, а также определяет необходимость дальнейшего анализа текста отчета или его фрагментов.

Теперь рассмотрим каждый из этапов более подробно.

Представим текст как упорядоченное множество предложений $\mathbf{t} = \langle \mathbf{s}_i \rangle$, предложение – как упорядоченное множество слов: $\mathbf{s}_i = \langle w_{ij} \rangle$. Под словосочетанием будем понимать упорядоченное множество слов: $\mathbf{c} = \langle w \rangle$. Будем считать, что словосочетание входит в i-е предложение ($\mathbf{c} \subset \mathbf{s}_i$), если в $\mathbf{s}_i$ существует контактное подмножество, эквивалентное $\mathbf{c}$.

Пусть $\mathbf{K} = \{\mathbf{c}\}$ – список ключевых коллокаций, вводящих требования к изделию, поставленные заказчиком (например, «должно обладать / состоять / …», «обеспечивает», «служит»). Тогда предложение $\mathbf{s}$, входящее в текст ТЗ, называется значимым, если $\exists \mathbf{c} \in \mathbf{K}: \mathbf{c} \subset \mathbf{s}$. Значимое предложение входит в значимый фрагмент, содержащий в себе одно или несколько предложений:

$$\mathbf{f} = \langle \mathbf{t}, s, e \rangle,$$

где $\mathbf{t}$ – текст, в который входит фрагмент, s – номер начального предложения фрагмента, e – номер последнего предложения фрагмента.

По результатам анализа текстов ТЗ были разработаны следующие правила, определяющие границы значимых фрагментов:

- если ключевая фраза встречается в предложении, после которого идет перечисление, то выделяется и весь текст до конца перечисления (например, «система должна состоять из следующих подсистем: …»);
- если ключевая фраза встречается в предложении, находящемся в связанном тексте, то выделяется предложение целиком;
- если фраза встречается отдельно (например, заголовок «необходимо»), то выделяется весь следующий абзац.

Эксперименты показали, что качество результата, полученного с помощью метода, возрастает, если помимо значимого предложения в фрагмент включается одно предложение до и два предложения после значимого, так как они чаще всего связаны по смыслу. Предыдущее предложение часто вводит некоторые определения или определяет общее направление, последующие – расшифровывают требования.

На первом шаге по тексту ТЗ ищутся ключевые фразы, к которым применяются приведенные правила. Если условие выполняется, выделяется очередной значимый фрагмент, который заносится в список $\mathbf{F} = \{\mathbf{f}\}$. Два значимых фрагмента могут быть объединены вместе, если их границы пересекаются или между ними нет значимого текста: если $\mathbf{f}_m = \langle \mathbf{t}, s_1, e_1 \rangle$ и $\mathbf{f}_{m+1} = \langle \mathbf{t}, s_2, e_2 \rangle$: $e_1 >= s_2$, то $\mathbf{f}_m = \langle \mathbf{t}, s_1, e_2 \rangle$, а $\mathbf{f}_{m+1}$ удаляется.

На втором шаге выделяются коллокации из значимых фрагментов, для значимых фрагментов рассчитывается вектор признаков:

$$\mathbf{a} = \langle \{w, f\} \rangle,$$

где $w \in \mathbf{f}$ – коллокация, а f – ее частота встречаемости.

Из вектора признаков отсеиваются коллокации с частотами выше 0,75 и ниже 0,25 от максимальной частоты. Эти меры позволяют избавиться от служебных слов и авторских особенностей текста, отсеять редко встречающиеся сочетания. В итоге будет сформировано множество векторов признаков ТЗ $\mathbf{S}_1 = \{\mathbf{a}\}$.

На третьем шаге текст отчета разбивается на абзацы, для которых формируется список коллокаций с частотами их встречаемости (вектор признаков $\mathbf{b}$: $\mathbf{S}_2 = \{\mathbf{b}\}$). Значимость абзаца с номером j вычисляется как максимум косинусной меры сходства вектора $\mathbf{b}$ с векторами $\mathbf{a}$ ТЗ или равна нулю, если найденная значимость ниже заданного порога:

$$v_j = \max_i \cos(\mathbf{a}_i, \mathbf{b}_j),$$

где $\mathbf{a}_i \in \mathbf{S}_1$, а $\mathbf{b}_j \in \mathbf{S}_2$.

На четвертом шаге лицо, принимающее решение, получает информацию о покрытии отчета фрагментами ТЗ в виде точечной диаграммы. Так как в работе метода возможны ошибки при выделении свойства или значимого фрагмента, эксперт может получить более подробную информацию о фрагментах отчета и технического задания: соответствие значимых фрагментов, список коллокаций и т.д.

По текстам ТЗ и отчетов формируются точечные диаграммы. На них точка соответствует 100, а строка – 10 000 символов текста. Темные точки показывают части текста, содержащие ключевые слова.

Проверка метода проводилась в два этапа. На первом этапе экспертам давали ознакомиться с содержанием ТЗ и отчета, и высказать свое мнение относительно их содержания, затем документы проверялись в соответствии с разработанным методом. На втором этапе проводилась кросс-проверка документации. Все ТЗ проверялись со всеми отчетами, чтобы проверить гипотезу о том, что максимум совпадения для качественно написанных отчетов должен находиться на соответствующих им ТЗ.

На рис. 1а представлена диаграмма разбора ТЗ, содержащего ненужную информацию. В нем большая часть текста говорит о составе и планах организации, проводимой ею научной работе. Техническое задание, соответствующее рис. 1б, написано в строгом стиле и по требованиям ГОСТа. Ключевые предложения найдены в середине текста в разделе, описывающем требования к изделию. В заключительной части ТЗ формируются сроки разработки, требования к рабочим местам и интерфейсу. Хотя число ключевых фрагментов в первом и втором случае почти одинаково, второй текст выигрывает из-за сжатости и точно поставленных требований.

На рис. 2а показана диаграмма для отчета полного «воды». Блоки из компактно расположенных 5-10 темных точек описывают заявленные в ТЗ требования. Отдельно стоящие темные точки соответствуют единичным коллокациям (например, в заголовке). Здесь на более чем 130 000 знаков отчета было найдено лишь 470 коллокаций, относящихся к ТЗ (считая единичные вхождения в заголовках). Максимальная длина связного текста, имеющего отношение к одному из значимых фрагментов ТЗ, – 700 символов.

На рис. 2б представлен качественно написанный отчет, в котором ключевые коллокации встречаются везде, за исключением начала (содержание, авторы, введение) и конца отчета (юридический и экономический разделы). При длине отчета свыше 130 000 знаков в нем найдено более 3500 коллокаций. Максимальная длина текста, имеющего значимые фрагменты ТЗ – 1500 знаков.

Для кросс-проверки были использованы тексты шести ТЗ и девяти отчетов. Отчетам присвоен номер соответствующих ТЗ. ТЗ с номерами 1-3 написаны по одной тематике, отчеты 5 и 6 написаны по близким тематикам. Отчет с номером 0 не связан ни с одним из ТЗ. Отчеты 3 и 6 представлены в двух версиях. Вторая версия отчетов, отмеченная знаком «+», содержит исправления найденных заказчиком недостатков.

Результаты проверки показаны в табл. 1. Соответствия ТЗ и отчетов выделены рамкой. Результаты удачных проверок выделены темным фоном. Успешные проверки с другими отчетами показаны светло-серым фоном.

Как видно из табл. 1, разработанный метод и программное обеспечение определило высокое качество отчетов, написанных для технических заданий 1-3 и 6. При этом результат работы системы для отчета 3 и 6 совпал с мнением заказчика. Отчет 0 не показал совпадений ни для одного из ТЗ.

Технические задания 4 и 5 не предполагали подробного описания результатов работы и требований к ним. Также в ТЗ 5 требовалось дать рекомендаций по улучшению изделия, это усложнило поиск соответствия. Отчет 4 содержал информацию по предметной области ТЗ 5, в связи с чем их сходство выше.

Для проверки метода были вычислены значения косинусной меры сходства между ТЗ и отчетами. Для этого использовались частоты встречаемости отдельных слов, отсеивание слов не проводилось. Результаты кросс-проверки для косинусной меры сходства сведены в табл. 2.

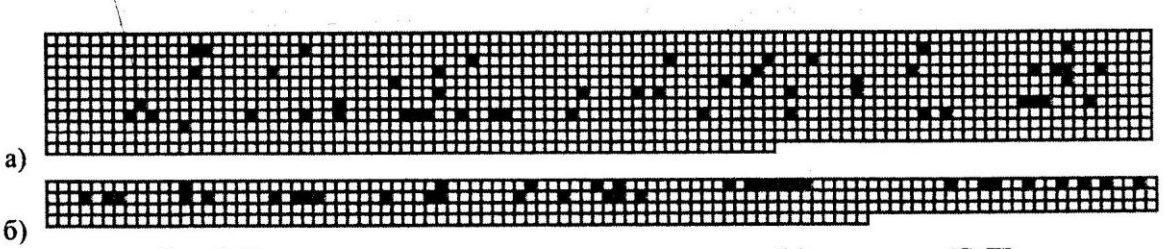

Рис. 1. Визуализация результатов анализа неудачного (а) и удачного (б) ТЗ

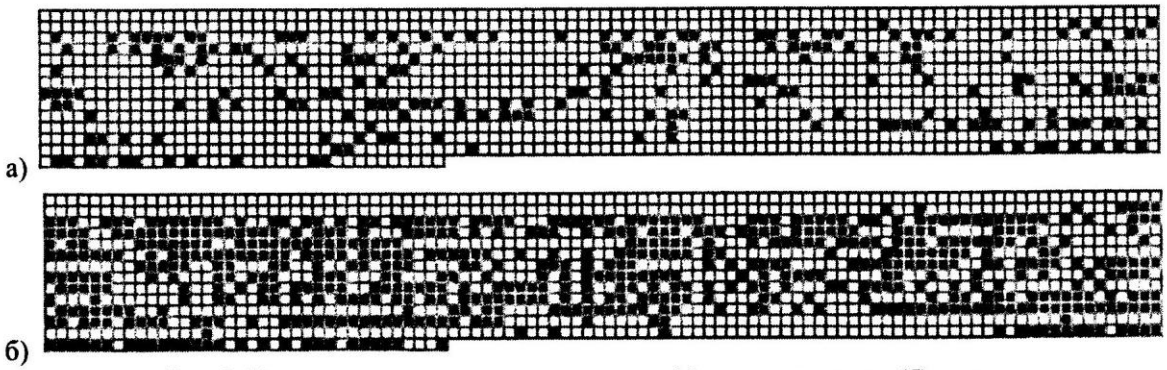

Рис. 2. Точечная диаграмма для неудачного (а) и качественного (б) отчета

Таблица 1

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

Таблица 2

Результаты кросс-проверки для косинусной меры сходства

		Технические задания					
		1	2	3	4	5	6
Отчеты	1	0,533	0,546	0,546	0,198	0,263	0,292
	2	0,532	0,554	0,880	0,151	0,169	0,162
	3	0,554	0,779	0,579	0,183	0,205	0,191
	3+	0,534	0,763	0,612	0,191	0,204	0,192
	4	0,331	0,238	0,326	0,189	0,212	0,167
	5	0,418	0,302	0,331	0,182	0,317	0,243
	6	0,161	0,091	0,116	0,089	0,091	0,443
	6+	0,163	0,091	0,117	0,089	0,091	0,443
	0	0,257	0,174	0,207	0,139	0,175	0,185

Как видно из табл. 2, косинусная мера успешно определяет тексты с общей тематикой, но не показывает качество отчетной документации: максимальные значения достигаются при сравнении ТЗ, не соответствующих данному отчету; разделимость хороших и плохих отчетов отсутствует.

Проверка результатов показала повышение точности работы метода по сравнению с методами, разработанными ранее. Проблему представляют ТЗ, описывающие лишь основные цели работы. Кроме того, даже при соответствии 70% и выше, детальная экспертиза необходима, так как метод не гарантирует полностью достоверных результатов. Для решения этой задачи необходимо применять специализированные методы (например, для экспертизы отчёта по надёжности электронных средств – методику, описанную в [14]).

Тем не менее, метод может применяться как часть автоматизированной системы ведения и хранения документации по проекту и помогать в принятии решений о доработке отчета или о его детальной экспертизе.

СПИСОК ЛИТЕРАТУРЫ

1. Яблочников Е.И., Молочник В.И., Миронов А.А. ИПИ-технологии в приборостроении: учебное пособие. – СПб: СПбГУ ИТМО, 2008. – 128 с.
2. Кролл П., Крачтен Ф. Rational Unified Process - это легко. Руководство по RUP для практиков:пер. с англ. – М.: КУДИЦ-ОБРАЗ, 2004. – 432 с.

3. Тернер М. Основы Microsoft Solutions Framework. – СПб.: Питер, 2008. – 336 с.
4. Клышинский Э.С., Антонова А.Ю. Об использовании мер сходства при анализе документов // Сб. трудов 13-й Всероссийской научной конференции RCDL'2011, с. 246-250.
5. Broder S., Glassman M. Manasse and G. Zweig. Syntactic clustering of the Web // Proc. of the 6th International World Wide Web Conference, April 1997.
6. Зеленков Ю.Г., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для WEB-документов // Сб. трудов 9-й Всероссийской научной конференции RCDL'2007, с. 166-174.
7. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск. – М.: Вильямс, 2011. – 528 с.
8. Клышинский Э.С. Анализ комплексных мер тематического сходства документов // Научно-техническая информация. Сер. 2. 2011. – № 9. – С. 6-11.
9. Колчин А. Что такое PDM? // PC Week. – 2001. – № 38.
10. Черников Б. В. Технологии подготовки документов на основе кибернетических методов. – М.: Финансы и статистика, 2009. – 206 с.
11. Тарасенко А.В. Разработка и исследование методов и моделей автоматической проверки текстов на соответствие требованиям технической документации: автореф. дис. ... д-ра техн. наук. – Таганрог, 2009.
12. Волкова Г.А. Создание «онтологии всего». Проблемы классификации и решения // Сб. трудов научно-практического семинара «Новые информационные технологии в автоматизированных системах». – М., 2013. – С. 293–300.
13. Заболеева-Зотова А.В., Орлова Ю.А. Автоматизация процедур семантического анализа текста технического задания // Известия Волгоградского гос. технического университета. – 2007. – Т. 9, № 3. – С. 52-55.
14. Жаднов В.В. Методические указания по проведению экспертизы конструкторского документа РР01 «Расчет надежности» для электронных модулей первого уровня с использованием технологии прогнозирования надежности АСОНИКА® – М.: МИЭМ НИУ ВШЭ, 2012. – 16 с.

Материал поступил в редакцию 21.02.14.

Сведения об авторах

КЛЫШИНСКИЙ Эдуард Станиславович – кандидат технических наук, доцент Московского института электроники и математики Национального исследовательского университета – Высшая школа экономики (МИЭМ НИУ ВШЭ)
e-mail: eklyshinsky@hse.ru

КАЛАЧЁВ Ярослав Борисович – аспирант МИЭМ НИУ ВШЭ
e-mail: Kalachyov-YB@sac.minenergo.gov.ru

ЖАДНОВ Валерий Владимирович – кандидат технических наук, доцент, профессор МИЭМ НИУ ВШЭ
e-mail: vzhadnov@hse.ru