САПР и ГРАФИКА

Метод проверки содержательной полноты отчетной документации

Эдуард Клышинский, Ярослав Калачев

Введение

Выполнение проектных работ сопровождается оформлением множества различных документов. Задание на разработку оформляется в форме технического задания (ТЗ), частного технического задания, плана-проспекта, технических условий или других документов (здесь мы будем использовать термин ТЗ для всех видов документации начального уровня). При завершении работ зачастую оформляется отчетная документация различного вида: технический отчет, руководства для пользователей и обслуживающего персонала и т.д.

При работе в соответствии с принципами CALS-технологии [1] проект должен заканчиваться оформлением полной и понятной документации, с помощью которой можно понять суть и принципы проведенных работ. Заказчик или исполнитель должен иметь возможность без привлечения дополнительных исследований повторить или улучшить полученный результат.

Согласно стандартам оформления документации, ТЗ должно включать строго определенные пункты. Существует несколько стандартов на оформление технической документации. В первую очередь это серия ГОСТов, по которым оформляется документация в государственных учреждениях. В качестве альтернативы можно назвать стандарты International Standard Organization, которые дают рекомендации по составу документации. В области электротехники и телекоммуникаций существует серия рекомендаций European Telecommunications Standards Institute (http://www.etsi.org); в области разработки программного обеспечения — рекомендации Rational Unified Process от IBM[2] и Microsoft Solutions Framework [3]. Названные рекомендации не утверждены государствами в качестве стандартов, но договор может предусматривать проведение проекта в соответствии с одним из них. Помимо этого имеются собственные разработки компаний, определяющие состав и содержание документации. Например, работающая в области систем автоматизированного проектирования компания Integra (http://www.integra.jp/) использует собственные бизнес-процессы, гарантирующие качество разрабатываемого программного обеспечения. Знание структуры ТЗ помогает проводить его анализ, в том числе и в автоматизированном режиме.

Оформляемые по данному ТЗ отчеты должны раскрывать те или иные аспекты задания. В некоторых случаях вручную полностью проверить соответствие отчета техническому заданию бывает затруднительно. Например, недобросовестный исполнитель может пытаться спрятать низкое качество отчета за его объемом, цитатами из открытых источников, уходом в смежную предметную область, за бюрократическим стилем изложения текста. Также он может вставить в текст отчета фрагменты предыдущих отчетов, изменив порядок следования фрагментов текста.

Для помощи в приемке отчетов необходимо создать автоматизированную систему, проверяющую степень соответствия отчетов или технической документации тексту ТЗ. Подобная система должна выявлять блоки итоговой документации, тематически сходные с ТЗ. При этом она не должна применяться автономно для принятия решения о приемке или отклонении отчета. Она является лишь первым этапом проверки, выявляющим наиболее явные несоответствия, которые должны быть подтверждены специалистом при анализе документации. На вход система должна получать тексты документов, содержащих постановку задачи и требования к проекту, а также отчетные документы. На выходе система обнаруживает и оценивает степень сходства и связности фрагментов текста, их смысловой нагруженности в рамках данного проекта. Подобные оценки могут использоваться ответственным лицом для принятия решения о прохождении документации.

В предыдущих работах мы уже показывали работоспособность предложенного метода [4]. В данной статье будет продемонстрировано его развитие и результаты вновь проведенных экспериментов.

Существующие методы сравнения документов

Задача определения идентичности документов или их фрагментов была решена достаточно давно. В этой области можно выделить несколько наиболее распространенных задач: определение заимствований в документах, определение нечетких дублей документов, определение тематической близости документов.

Задача определения нечетких дублей документов сейчас сводится к быстрому вероятностному определению сходства. Одним из методов, работающих в этой области, является метод шинглов [5]. В статье [6] проведен обзор модификаций нескольких методов определения неполного сходства документов. Суть метода состоит в том, что из текста выделяются шинглы — последовательности слов длины *k*. Для каждого документа выделяется фиксированное количество шинглов. Выбор шинглов может проводиться по разным алгоритмам: частоте встречаемости слов, мере IDF, частоте встречающихся комбинаций букв. Считается, что если шинглы двух документов существенно совпадают, то документы являются сходными. Значение меры сходства документов резко падает при расхождении наборов их шинглов. Метод шинглов может использоваться для сравнения небольших фрагментов текста (около 1000 знаков и более). При сравнении документов большого объема точность метода снижается. Здесь также необходимо тщательно подходить к выбору сравниваемых фрагментов.

Задача систем антиплагиата состоит в нахождении сходных фрагментов в разных документах. В отличие от метода шинглов, не сравнивается весь документ, а производится поиск отдельных сходных фрагментов. Сходство также может быть неполным, то есть фрагменты могут не совпадать полностью.

На данный момент для русского языка наиболее крупной является система «Антиплагиат.РФ» [7]. Однако существуют решения, которые применимы для анализа локальных документов (наибольшее распространение они получили в высших учебных заведениях). Подобные методы поиска заимствований используются для определения фрагментов отчетной документации, применяемой повторно. Однако они не могут использоваться для решения задачи определения соответствия отчетной документации техническому заданию, так как цитирование фрагментов ТЗ в отчетных документах является разрешенным, но не распространенным приемом.

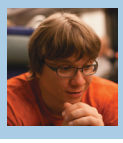
Для определения тематического сходства чаще всего применяются меры, определен-

Эдуард Клышинский

Канд. техн. наук, доцент МИЭМ НИУ ВШЭ, закончил МИЭМ в 1997 году, защитил кандидатскую диссертацию в 2000-м.

Ярослав Калачев

Аспирант МИЭМ НИУ ВШЭ.



ные на векторах, содержащих в себе частоты встречаемости слов рассматриваемых документов. Пусть для двух документов вычислены векторы частот встречаемости слов *a* и *b*. Векторы определены на множестве всех слов, встречающихся в обоих документах. В этом случае, например, косинусная мера сходства определяется следующим образом:

$$\cos(a,b) = \frac{\Sigma a_i * b_i}{\sqrt{\Sigma a_i * a_i} * \sqrt{\Sigma b_i * b_i}}$$ [8].

Развитием метода является использование векторов частот не отдельных слов, а их сочетаний различной длины. В этом случае точность определения степени сходства документов возрастает [9].

Однако косинусная мера сходства — это лишь средство определения меры похожести двух документов. В связи с этим задача выбора сходных фрагментов ТЗ и документации может быть решена методами поиска заимствований.

Метод анализа документации

Кратко суть предложенного метода можно описать следующим образом. Из текста ТЗ выделяются предложения, содержащие ключевые характеристики разрабатываемой системы. Стоящие рядом или пересекающиеся фрагменты ТЗ объединяются в один. Далее из выделенных фрагментов извлекаются все пары стоящих рядом слов (коллокации), незначимые коллокации удаляются. На следующем шаге текст рассматриваемого отчета разбивается на фрагменты, для которых также выделяется список коллокаций. Далее находится максимум меры сходства текущего абзаца со значимыми фрагментами ТЗ. Это значение и будет мерой соответствия абзаца тексту технического задания. Полученный результат выдается лицу, принимающему решение (ЛПР), в визуальной форме. На основании этих данных ЛПР выявляет меру соответствия отчета и ТЗ, а также определяет необходимость дальнейшего анализа текста отчета или его фрагментов. Опишем теперь каждый из этапов более подробно.

Для нахождения значимых фрагментов ТЗ используется список ключевых слов или словосочетаний, обозначающих требования к системе, которые сформулированы заказчиком (должна обладать/состоять, обеспечивает). Список ключевых слов и фраз может редактироваться пользователями системы в зависимости от их собственного опыта или предметной области, в которой они работают.

По результатам анализа текстов ТЗ были разработаны следующие правила, определяющие границы значимых фрагментов ТЗ. Если слово встречается в предложении, после которого идет перечисление, то выделяем предложение и весь текст до конца перечисления (например, «система должна состоять из следующих подсистем:…»). Если слово встречается в предложении, находящемся в связном тексте, то предложение выделяется целиком. Если слово встречается отдельно (например, заголовок «необходимо»), то выделяется весь следующий абзац.

На первом шаге метода по тексту ТЗ ищутся ключевые фразы, к которым применяются приведенные нами правила. Если условие выполняется, выделяется очередной значимый фрагмент, который заносится в список. Эксперименты показали, что эффективность метода возрастает, если помимо собственно значимого предложения во фрагмент включается одно предложение до и два предложения после значимого. Это происходит потому, что соседние предложения чаще всего связаны по смыслу. Предыдущее предложение часто вводит некоторые уточнения или определяет общее направление, тогда как последующие предложения расшифровывают предложение, содержащее требования.

После выделения значимых фрагментов проводится их объединение. Для этого, помимо текста выделенного фрагмента, хранится еще его начало и длина. Два значимых фрагмента могут быть объединены вместе, если их границы пересекаются или между ними нет значимого текста (например, между ними расположена подрисуночная подпись).

На втором шаге проводится выделение коллокаций из значимых фрагментов. Из полученного списка удаляются коллокации, частота встречаемости которых выше 0,75 и ниже 0,25 от максимальной. Это позволяет избавиться от служебных слов и авторских особенностей текста, отсеять редко встречающиеся сочетания. В результате для каждого значимого фрагмента будет получен вектор признаков. В итоге можно сформировать множество векторов признаков технического задания $S=\{a_i\}$, здесь *a* содержит коллокации и частоты их употребления.

На третьем шаге текст отчета разбивается на абзацы, так как они содержат некоторую законченную мысль и должны обладать тематическим сходством. Для каждого из абзацев формируется список коллокаций с частотами их встречаемости (вектор признаков *b*), после чего просчитывается максимум косинусной меры сходства вектора *b* с векторами *a* ТЗ. Тогда значимость абзаца с номером *j* вычисляется как $v_j = \max_i \cos(\mathbf{a}_i, \mathbf{b}_j)$, номер *i*-абзаца, для которого был получен максимум сходства, также сохраняется. Если значимость абзаца ниже заданного порога, то считается, что значимый фрагмент не найден и абзац не содержит в себе информации, описывающей какую-либо из характеристик разрабатываемой системы.

На четвертом шаге ЛПР получает информацию о списке значимых фрагментов ТЗ, для которых не найдено соответствия в отчете. На практике может оказаться, что система совершила ошибку при выделении свойства или значимого фрагмента. Кроме того, описание свойства может проводиться с использованием синонимичной лексики. Но чаще всего такая ситуация говорит о том, что описание свойства выполнено неполно. Частота совпадений выбранного значимого фрагмента с абзацами отчета может показывать степень подробности описания свойства.

На последнем шаге проводится визуализация результатов. Эксперту показываются значимые фрагменты, выделенные из ТЗ, и степень соответствия абзацев отчета каким-либо значимым фрагментам ТЗ. Для каждого значимого фрагмента ТЗ выводится количество совпавших абзацев, их расположение в тексте отчета, а при необходимости — и сам текст абзаца.

Результаты экспериментов

В эксперименте использовались имеющиеся у нас тексты ТЗ. Результаты выделения значимых фрагментов показаны на рис. 1 и 2. Значимые фрагменты представлены в виде точечной диаграммы, где каждая точка является представлением 100 символов текста. В каждой строке диаграммы 100 точек, то есть 10 000 символов. Диаграмму следует читать слева направо построчно. Зеленые точки показывают части текста, содержащие ключевые слова.

На Рис.1 представлена диаграмма разбора неудачно (по мнению эксперта, прочитавшего ТЗ) написанного ТЗ, которое содержит большое количество ненужной информации. Значимые фрагменты разбросаны по всему тексту ТЗ. Большинство абзацев подобного текста повествует о планах организации, составе организации и исследованиях, а также о проводимой научной работе.

По итогам разбора было получено более 400 коллокаций, число которых после отсеивания было сокращено примерно до 200. Из-за частого упоминания в тексте названия организации заказчика, в ключевые коллокации попали сочетания, не относящиеся к теме проекта: Российской Федерации, Правительства Российской Федерации, Министерство энергетики.

ТЗ, соответствующее рис. 2, написано в строгом стиле и по требованиям ГОСТа. Как видно по распределению зеленых блоков, все ключевые предложения были найдены в середине

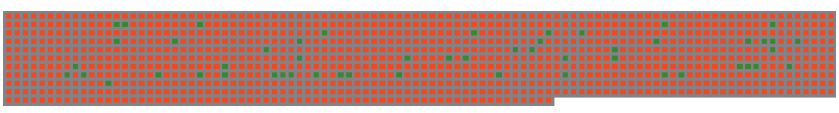


Рис. 1. Пример визуализации результатов анализа неудачного ТЗ

Рис. 2. Пример визуализации результатов анализа удачного ТЗ

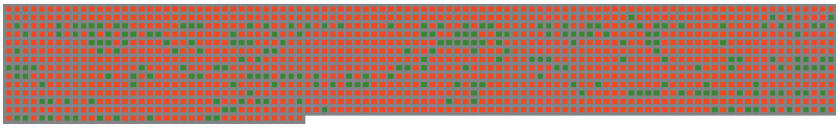


Рис. 3. Точечная диаграмма для неудачного отчета

Рис. 4. Точечная диаграмма для качественно написанного отчета

текста и соответствуют разделу, описывающему требования к системе. Заключительная часть ТЗ относится к срокам разработки, требованиям к рабочим местам и к интерфейсу. Хотя число ключевых фрагментов в первом и втором случаях практически одинаково, второй вариант выигрывает в точности выделенных коллокаций из-за сжатости текста, точно поставленных требований и отсутствия ненормативных предложений.

Приведенные ТЗ сопровождались текстами отчетов. Как и следовало ожидать, исполнитель, пренебрегающий точностью при оформлении ТЗ, также не следит за качеством текста и при написании отчета. На рис. 3 показана диаграмма для отчета, полного «воды», то есть включающего мало информативного текста. Зелеными точками показаны найденные ключевые коллокации. Блоки из компактно расположенных 5-10 зеленых точек описывают заявленные в ТЗ требования. Отдельно стоящие зеленые квадраты — единичная встреча коллокации в тексте (например, заголовок).

В данном отчете, состоящем из более чем 130 тыс. знаков, 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. Проверка текста отчета экспертом подтвердила оценку программы.

На рис. 4 представлена диаграмма качественно написанного отчета, в котором ключевые коллокации встречаются везде, за исключением начала (содержание, авторы, введение) и конца отчета (юридическая и экономическая информация). При длине отчета более 130 тыс. знаков найдено более 3500 коллокаций. Максимальная длина текста, имеющего значимые фрагменты ТЗ, — 1500 знаков.

Помимо точечной диаграммы, показывающей наличие или отсутствие коллокаций в тексте, мы применили другой вид диаграммы, отображающей связность фрагментов отчета и ТЗ. Каждому фрагменту текста ТЗ сопутствует цвет, отображавшийся и в диаграмме отчета для соответствующих фрагметов. На рис. 5 и 6 показаны новые диаграммы для рассмотренных отчетов. Цвет меняется от синего к зеленому в зависимости от номера фрагмента. Поскольку требования пишутся последовательно и обычно связаны по смыслу, можно предположить, что чем ближе цвет на диаграмме, тем ближе фрагменты текста по значению.

На примерах хорошо видно, что первый отчет содержит фрагменты, которые пытаются охватить сразу несколько тем. Фрагмент максимальной длины относится именно к такому смешанному типу описания. В противоположность первому примеру на рис. 6 видны большие фрагменты, в которых цвет перетекает плавно, то есть тема раскрывается последовательно. Очевидно, что во втором ТЗ блоки текста тематически более связаны, чем в первом.

Выводы и анализ результатов

Проверка экспертом результатов работы показала пригодность метода для применения в практических задачах. В отчетном документе одноцветные блоки, длина которых превышает 5-6 фрагментов (квадратов на диаграмме), действительно содержат информацию, соответствующую по смыслу оригинальному ТЗ. Одноцветные блоки большей длины (10-20 точек на диаграмме) полноценно описывают работу разрабатываемой системы. Если несколько блоков идут подряд с плавным цветовым переходом, есть вероятность, что такой фрагмент текста скопирован из ТЗ. Отдельные блоки в 1-3 квадрата представляют случайно встретившиеся или отдельно стоящие ключевые слова.

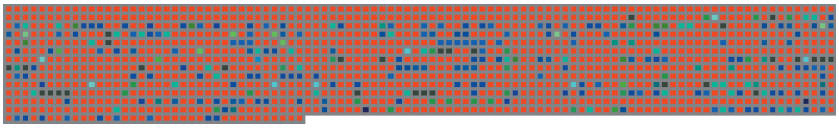


Рис. 5. Близость фрагментов для неудачного отчета

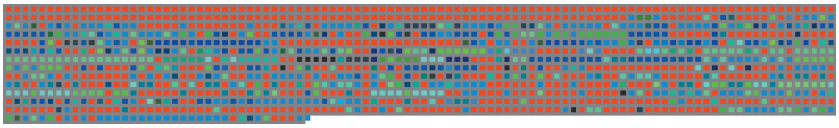


Рис. 6. Близость фрагментов для качественно написанного отчета

Таким образом, для определения процента полноты отчета необходимо предварительно отбросить все отдельно стоящие блоки, имеющие совпадения с ТЗ. Также рекомендуется не рассматривать начало документа, содержащее формальные его части. Проверка информативности отчета, а также степени освещения требований ТЗ служит хорошим показателем качества итоговых документов. В ходе экспериментов было выявлено, что отчет, более чем на 70% состоящий из блоков, связанных с текстом ТЗ, может считаться выполненным на высоком уровне, качественно.

Предложенный метод может применяться как часть системы ведения и хранения документации по проекту и использоваться экспертом для принятия решения о необходимости доработки отчета или о его приемке. Созданная на его основе система проверки отчетов не заменяет человека, а лишь помогает ему отсеять заведомо некачественно написанные тексты.

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. 246-250 с.
5. A. 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.
7. Романов М.Ю., Житлухин Д.А. Внедрение системы «Антиплагиат» в Российской государственной библиотеке // Сб. трудов 11-й Всероссийской научной конференции RCDL'2009.
8. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск. — М.: Вильямс, 2011. 528 с.
9. Клышинский Э.С. Анализ комплексных мер тематического сходства документов // Научно-техническая информация. Сер. 2: Информационные системы и процессы. Серия 2. 2011. № 9. С. 6-11.

Данная работа поддержана грантами РГНФ № 12-04-00060 и РФФИ № 11-01-00793.