Перспективные методы обработки проектной документации♣

© Э.С. Клышинский

Московский государственный институт электроники и математики
klyshinsky@mail.ru

Аннотация

Ставятся задачи, решение которых могло бы вывести обработку проектной документации на качественно иной уровень. В качестве таких задач рассматриваются автоматизированное выделение требований к изделиям, поиск прецедентных документов в ходе анализа и проектирования изделия, проверка полноты документации при завершении работ над изделием, автоматическая генерация документации о составных частях изделия, автоматический подбор компонентов для изготовления изделия.

1 Введение

В ходе своего существования крупные предприятия формируют большой архив, содержащий в себе различного рода документацию, связанную с их функционированием. К подобным документам относятся не только результаты официального документооборота (приказы, распоряжения и пр.), но и техническая документация по выполняемым и выполненным проектам: технические отчеты, проектная документация, планы и т. д. В последнее время довольно широкое распространение получили системы ILM (Information Lifecycle Management) и PDM (Product Data Management). ILM охватывает все процессы управления размещением, хранением, распределением, миграцией, архивированием и удалением данных в инфраструктуре предприятия [1, 2]. Задачей ILM является хранение документов и обеспечение оптимального времени доступа к ним со стороны пользователя и его систем. Так, например, оперативная информация помещается «рядом» с пользователем на высокопроизводительных серверах. По мере устаревания информации и падения числа обращений к ней информация перемещается на удаленные серверы и далее в файловый архив, носители информации из которого могут заказываться как обычные книги. Для части информации может определяться время ее хранения и порядок уничтожения.

Труды 12й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

С помощью PDM-систем осуществляется отслеживание больших массивов данных и инженерно-технической информации, необходимых на этапах проектирования, производства или строительства, а также поддержки эксплуатации, сопровождения и утилизации технических изделий [3]. PDM-системы позволяют создавать отчеты о конфигурации выпускаемых систем, маршрутах прохождения в ходе технологического процесса изделий, их частей или деталей, составлять списки материалов и деталей, необходимых для производства изделия. Одной из задач, решаемых PDM-системами, является обеспечение возможности групповой работы над проектом.

Внедрение подобного рода систем позволило предприятиям перейти к безбумажному обороту проектной документации. Современные системы позволяют быстро и эффективно формировать рабочие группы, занимающиеся одним проектом, налаживать взаимодействие между сотрудниками таких групп, автоматизировать выпуск проектной документации, решить целый ряд организационных задач.

Однако развитие науки позволило перейти на качественно иной уровень работы с документацией. На данный момент ведется переход от электронного хранилища, являющегося заменой полок с документами, к интеллектуальной обработке документации. Часть проблем в данной области уже успешно решена. При помощи средств Business Intelligence проводится эффективный анализ собранной в ILM-системе данных. Специальные системы автоматически или автоматизированно формируют список деталей и механизмов, входящих в состав изделия, список и порядок работ и производственных процессов, необходимых для изготовления изделия. На основе этой информации рассчитывается себестоимость изделия. Однако большая часть документов содержит в себе текстовое описание проекта. В особенной степени это относится к начальным и конечным этапам создания изделия: анализу, проектированию и внедрению. В связи с этим ставится вопрос об автоматизации процессов обработки текстовой документации.

В данной работе рассматривается несколько перспективных задач, связанных с автоматической обработкой текстов при проектировании и производстве различных изделий. Решение этих задач

должно помочь перевести некоторые процессы, выполняемые до сих пор вручную, в новое русло. В основном для решения поставленных задач рекомендуется использование документации, подготавливаемой в ходе обычного цикла разработки нового продукта. При этом такой документ или набор документов используются в качестве поискового запроса к информации, хранимой, например, в ILM- или PDM-системе.

Документ в качестве запроса активно используется в таких предметных областях, как рубрикация текстов, составление списка заметок по той же тематике из новостной ленты, других задачах, связанных с тематическим анализом текстов [4, 5]. На конкурсе РОМИП документ-образец используется для уточнения запросов пользователя и фильтрации выдачи. Лобовое использование документа, как материала для поисковых запросов, наталкивается на огромную выдачу со стороны поисковой машины, ее существенное время работы. Как следствие, пользователю становится сложно отобрать из большого объема полученных документов те несколько, которые его интересуют. В связи с этим используются такие технологии, как сокращение пространства поиска путем отбора лишь наиболее значимых для данного документа слов [6]. Для сравнения документов используются специальные метрики, учитывающие совпадение максимального количества слов или даже их распределения. В результате скорость и точность работы подобных систем существенно возрастают.

На различных этапах жизненного цикла возникает необходимость в обработке различного вида документации. При этом различаются как задачи, так и методы работы с этой документацией. Рассмотрим некоторые до сих пор не реализованные задачи обработки документации в соответствии с жизненным циклом разработки изделия.

2 Выделение требований к изделию

На этапе анализа проводится создание списка требований к конечному продукту. Подобный список используется, например, для ранжирования требований, определения их связности и непротиворечивости. Для автоматизированного извлечения подобного списка подходит целый ряд документов. При наличии внешнего заказчика проводятся предварительные переговоры, призванные согласовать мнения сторон относительного видения проекта. Во многих случаях ведется стенограмма подобных переговоров, которая и может использоваться для выделения требований. Аналогичным образом возможно использование таких документов, как постановка задачи, техническое задание, спецификации, вербальные модели поведения системы, описания логики ее поведения и т. д.

Удобным инструментом для выделения требований являются синтаксические шаблоны [7]. Сами требования обычно формулируются с использованием типичных фраз. Выделение этих фраз и связанного с ними контекста позволяет выделить описание задаваемых ими требований. Список требований может оформляться как набор гиперссылок на найденные части документа, помогая тем самым не потерять важную информацию из окружения фразы.

Использование онтологии или тезауруса дает возможность сгруппировать выделенные требования по классам. Их применение позволяет перейти к задаче извлечения знаний из естественно-языковых текстов [8, 9]. В результате можно сформировать некоторую модель отношений между понятиями, связывающую отдельные слова в описание объекта окружающей среды. Вслед за формированием модели объекта можно перейти к выделению значений его параметров. Проделав подобные операции можно перейти к проверке непротиворечивости требований, проверяя, например, диапазоны приписываемых их параметрам значений. Становится возможной группировка требований по классам, за счет чего упрощается их анализ человеком, проверка связанности и зависимостей между требованиями.

Формально задачу выделения требований можно поставить следующим образом. Пусть дан документ, содержащий требования к продукту. Необходимо создать систему синтаксических шаблонов, выделяющих предложения, содержащие подобные требования. Кроме того, требуется разработать систему, проводящую частичный синтаксический анализ, позволяющий выделять требования из полученных предложений.

Для выделения параметров создаваемого продукта необходимо создать систему частичного синтаксического анализа, выделяющую группы существительного. Набор прилагательных позволит выделить набор кандидатов на свойства. Однако слишком большой набор прилагательных может существенно усложнить работу проектировщику. В связи с этим необходимо ограничивать количество рассматриваемых конструкций. Для этого могут использоваться статистические методы, когда просматриваются лишь наиболее часто встречающиеся конструкции, либо семантические, когда при помощи онтологии или тезауруса проводится группировка выделенных требований. В результате на рассмотрение специалиста передаются лишь наиболее значимые по выбранному критерию объекты и их свойства.

3 Поиск прецедентных документов

Крупная компания за время своего существования накапливает большое количество документации, которая должна использоваться в начале проектных исследований. Проектировщики должны проверить, есть ли в документарном хранилище информация о сходных проектах. В результате поиска может выясниться, что компания уже выпускала подобную продукцию, и нет необходимости в ее проектировании с нуля. Может потребоваться лишь некоторое перепроектирование системы, исходя из современного состояния дел: надо применить новую элементную базу, новые технологические решения,

добавить или заменить функциональность. В хранилище может найтись описание блоков, которые могут применяться при построении новой системы. Результат может оказаться и прямо противоположным: исследования показали, что существующие методы решения не позволяют эффективно получить результат. Иными словами информационное хранилище превращается в прецедентную библиотеку, хранилище опыта и знаний, применяющихся для работы во время начальных этапов создания изделия. При этом информационный поиск должен вестись по всей совокупности документов, так как результатом поиска должна стать информация не только о готовых изделиях, но и о невоплощенных идеях, проектных решениях, методиках и технологиях, имеющих отношение к новому проекту.

Традиционно подобный поиск ведется по ключевым словам с использованием стандартных технологий. Однако зачастую выбор ключевых слов определяется здесь самим проектировщиком и, как следствие, может быть неполон или ограничен. В результате имеющиеся в базе необходимые документы не будут найдены, так как оказались не релевантными введенному запросу.

Однако к моменту начала проектирования системы должно быть сформулировано техническое задание, описывающее требования к системе. Обычно это связанный многостраничный документ, содержащий максимально имеющееся на данный момент формальное описание разрабатываемой системы. Данный документ может быть использован в качестве запроса к информационному хранилищу. По документу-запросу строится профиль документа, например, список наиболее часто встречающихся слов. Аналогичные профили хранятся в базе и для документов. Сравнение профилей позволяет выделить наиболее релевантные документы. При этом вероятность пропустить необходимые ключевые слова существенно снижается. Выделение профилей позволит решить задачу хранения информации «близко» к пользователю. Подобная проблема встает в связи с тем, что большинство документов по уже завершенным проектом обычно переносится из области актуальных документов в область «медленного» хранения, доступ к которой более ресурсоемок.

Вопрос построения профиля уже разбирался, например, при кластеризации документов и достаточно хорошо проработан [10]. Кроме того, вместо использования отдельных слов можно оценивать их сочетания. При таком подходе можно применять такие методики, как латентно-семантический анализ [11], Bag-of-Words [12, 13] или коллокации [14]. Применение этих подходов позволяет повысить релевантность выдаваемых документов. При использовании в качестве запросов отдельных слов подобные методы применимы значительно меньше, так как запрос может состоять из несвязанных фраз или попросту содержать в себе единственное слово.

Формально задача может быть поставлена следующим образом. Пусть дано множество документов, а также документ-запрос, при этом каждый из документов характеризуется своим распределением слов. Необходимо найти функцию от двух параметров: распределение слов для документа из множества и распределение слов документа-запроса, такую, что она позволяет ранжировать документы из множества по релевантности документа запросу. Вид функции будет зависеть от используемых средств. Так, в простейшем случае в качестве распределения слов может использоваться относительная частота встречаемости слов в тексте, а функция ранжирования будет представлять собой квадрат разности между ними. Как показывает практика, даже такие простые средства могут давать неплохие результаты, помещая на первое место в выдаче наиболее релевантный документ. Однако остальные документы в выдаче будут иметь произвольную релевантность, так как, например, документ большого объема может дать лучший результат, чем маленький документ, имеющий фрагмент на заданную тему, однако имеющий небольшое совпадение по лексике.

Прямое использование коллокаций оказывается затруднительным. Имеющиеся меры расчета коллокаций дают величины, позволяющие ранжировать словосочетания по неслучайности их появления в рамках одного документа. Однако для документов различного размера вычисляемые значения оказываются не сопоставимыми между собой. Таким образом, такая мера, как квадрат разности, использоваться не может. Прямой подсчет количества коллокаций, имеющихся в обоих документах, дает относительно неплохие результаты, но также не позволяют отсеять нерелевантные документы, ранжируя всё множество.

Возможным вариантом решения задачи может служить отсев словосочетаний, равномерно распределенных по всем или почти всем документам [14]. Подобные словосочетания имеются в той или иной степени в любом входном документе и, скорее всего, относятся к общей лексике. В этом случае в верхнюю часть выдачи не попадут документы, имеющие большое количество стандартных словосочетаний. Более общим случаем является введение дискриминирующей силы словосочетаний. Здесь для документа определяется набор словосочетаний, максимально отличающих данный документ от всех остальных. При этом дискриминирующая сила определяется не только для словосочетаний, представленных в документе, но и для отсутствующих в нем, так как отсутствие словосочетаний может служить отличительным стилевым или тематическим признаком.

4 Проверка полноты документации

Работы в данной области ведутся уже довольно продолжительное время. Так, например, была создана система «ЛоТА», предназначенная для анализа документации с использованием системы онтологий [15, 16]. Данная система предназначена для применения при анализе документов «Логика работы системы …» в авиационной промышленности [10]. В

ней предварительно проводится морфологический анализ текста документа, с использованием онтологии выделяются термины предметной области, проводится частичный синтаксический анализ. Далее система позволяет использовать документ как источник для ответов на запросы пользователя. Подобным образом система может извлекать такую информацию, как название алгоритма, содержащегося в документе, его задача, входные данные и т. д. При отсутствии информации в документе формируется соответствующее сообщение. Таким образом, может быть сформирован список вопросов, которые должны быть освещены в документах того или иного рода. Подготовленный документ контролируется при помощи системы на предмет его полноты.

Однако проверка полноты может осуществляться и с других позиций. Итоговая документация на систему должна отражать заданный список вопросов. Список этих вопросов изначально формулируется в техническом задании при проектировании системы. Таким образом, техническое задание может рассматриваться как документ, генерирующий запросы к комплекту итоговой документации. Для этого необходимо провести разбиение самого технического задания на монотематические фрагменты, представляющие собой описание постановки задачи на основные положения разработки. Далее, при помощи одного из разработанных методов строится профиль каждого из положений, состоящий из ключевых слов или словосочетаний. Полученные слова и словосочетания используются для информационного поиска по массиву итоговой документации. Если по какому-либо из запросов выдача отсутствует, то это служит индикатором отсутствия информации о разделе.

Аналогичным образом могут сравниваться профили фрагментов документа-запроса с профилями итоговой документации. Слабая степень соответствия профилей означает неполноту представленной информации.

К сожалению, метод не защищен от фрагментарного упоминания решения, когда после постановки вопроса следуют весьма урезанное его описание или пространные рассуждения из другой области. Для распознавания такой ситуации можно оценить дисперсию частоты упоминаний отдельных терминов, причем подобная оценка может проводиться по всем выделенным фрагментам.

Другим недостатком метода является возможная смена терминологии при переходе от технического задания к итоговой документации. Например, техническое задание было написано сотрудниками заказчика, привыкшими к собственной терминологии, возможно даже сленгу, тогда как проект разрабатывался профессионалами, тонко чувствующими разницу между различными определениями. В этом случае лексический состав документов будет существенно различным. Здесь на помощь могут прийти семантические методы анализа документов, например, методы, опирающиеся на онтологии. Мера соответствия профилей документов будет определяться не количеством совпадений, а степенью близости понятий в графе онтологии [17, 18].

Формально задача может быть поставлена следующим образом. Пусть имеется документ, описывающий требования к проектируемому продукту. Требуется разделить документ на несколько фрагментов, относящихся к различным аспектам описания продукта. В качестве метода разделения используется кластеризация документа по абзацам.

Далее, пусть дан набор документов, подготовленных на завершающей стадии разработки продукта. На этом этапе требуется найти связанные фрагменты документов, релевантные выделенным на предыдущем этапе кластерам. Отсутствие соответствия тому или иному кластеру означает неполноту документации.

Формальная структура документа, оговоренная тем или иным стандартом, может помочь в двух направлениях. Во-первых, при кластеризации документа могут быть отброшены такие его части, как преамбула, введение, заключение и подобные им элементы, обычно касающиеся обобщений, а не описаний. Во-вторых, стандарт оговаривает состав документации и поднимаемые в каждом из документов вопросы. Таким образом, может быть составлен список документов, которые должны быть подготовлены при разработке продукта. В этом случае проверка полноты может осуществляться тривиальным способом – проверкой наличия соответствующего документа в документарной системе.

5 Автогенерация документации

В ходе проектирования изделия разрабатывается большое количество документации в специальных форматах. Это могут быть чертежи, UML-диаграммы функционирования, специализированные описания систем, подсистем и их взаимодействия. Производимые в них изменения требуют вносить изменения и в текстовое описание проделанной работы. Чтобы избежать постоянной переделки текстовой документации, можно попытаться автоматизировать процесс ее составления. Так, например, к отдельным блокам и подсистемам могут привязываться текстовые фрагменты. Таким образом, документация может быть превращена из текстовой в интерактивную. Разработчик выбирает заданный блок и имеет возможность ознакомиться с его описанием. Далее, для проектной документации задается логика изложения обычного линейного документа, собираемого из отдельных текстовых блоков. Последовательность изложения может задаваться и последовательностью действий, если речь идет об описании логики функционирования системы.

Кроме того, к различного вида связям могут привязываться различные шаблонные языковые конструкции. Так, например, соединение с валом позволяет говорить, что «вал вращает» и добавлять название детали. Для задания подобного рода описаний и генерации связанного текста по ним можно использовать уже упоминавшуюся объектно-ориентированную модель представления знаний [8].

Задав сценарий описания создаваемого изделия, мы получаем возможность генерировать текстовую документацию при внесении тех или иных изменений.

Аналогичная методика может использоваться для порождения текстовых описаний проводимых экспериментов. Подобные описания могут проводиться по текстовому шаблону, в который включены спецификаторы, на место которых подставляются определенные параметры модели. Подобная схема уже много лет используется в Канаде при составлении прогноза погоды на английском и французском языках [19].

6 Автоматический подбор компонентов

Представим, что у нас имеется база данных, содержащая в себе описания различных компонент, используемых в производстве. При этом помимо описания характеристик компоненты имеется и текстовое описание данного компонента. Используя подобную базу, можно провести подбор компонент, оптимально подходящих для выпускаемого изделия. На данный момент уже существуют системы B2B и B2C, а также сервисы, подобные Яндекс.Маркет.

Результатом проектирования системы является ее детальное описание, которое содержит в себе, среди прочего, разбиение системы на отдельные блоки. Используя поисковые алгоритмы, можно провести поиск необходимых компонент по хранимым в базе описаниям. По результатам информационного поиска обычно сравниваются лишь основные характеристики, имеющие, как правило, числовое значение, или значение из заданного множества (цвет, тип, …). Однако ряд дополнительных характеристик может помещаться в текстовом виде, и поиск по ним будет вестись с использованием ключевых слов. Как это уже замечалось выше, подобный поиск может оказаться неполным. Однако компонентам проектируемой системы в документации приписывается как словесное описание, содержащее в себе как диапазоны необходимых значений основных параметров, так и описание условий эксплуатации, некоторых особенностей реализации и другие параметры, описание которых не может быть кратким. Описание требующихся нам компонент может использоваться для формирования запроса к базе, хранящей подобные компоненты. Для этого необходимо извлечь из текстового описания документа-запроса список интересующих нас параметров и значения этих параметров. Кроме того, необходимо сформировать профиль остальной части описания и использовать его для сравнения с профилями хранимых описаний.

7 Выводы

Создание систем нарастающей сложности остро ставит вопросы дальнейшей автоматизации всех процессов, начиная с анализа требований и заканчивая технической поддержкой пользователя и уничтожения отработанных изделий. Наибольший успех в этой области был достигнут для задач создания и хранения документации. Одним из следующих шагов должна являться автоматизация обработки массивов накопленной и поступающей информации в интересах дальнейшего сокращения времени и стоимости разработки, повышения уровня управляемости протекающих процессов. Поставленные в данной работе проблемы призваны зафиксировать направления для дальнейших исследований.

Литература

[1] Головченко А. ILM – концепция и инструментарий // PCWeek Review. – 2008. – № 1.

[2] Орлов С. Жизненный цикл ILM // LAN. – 2007. – № 7.

[3] Беспалов В., Клишин В., Краюшкин В. Развитие систем PDM: вчера, сегодня, завтра ... // САПР и графика. – 2001. – № 12.

[4] Чугреев В.Л., Яковлев С.А. Выделение критериев поиска текста на основе подобия значимых документов // ВУЗОВСКАЯ НАУКА – РЕГИОНУ: Материалы 1-й Общероссийской нучн.-техн. конф. – Вологда: ВоГТУ, 2003. – С. 200-202.

[5] Некрестьянов И., Некрестьянова М. РОМИП' 2006: отчет организаторов // Российский семинар по Оценке Методов Информационного Поиска. Труды четвертого российского семинара РОМИП'2006, Суздаль, 19 октября 2006 г. – Санкт-Петербург: НУ ЦСИ, 2006. – С. 7-29.

[6] Пескова О.В. Автоматическое формирование рубрикатора полнотекстовых документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Десятой Всерос. науч. конф. RCDL'2008, Дубна, 7 – 11 октября 2008 г. – Дубна: ОИЯИ, 2008. – С. 139-148.

[7] Большакова Е.И., Баева Н.В., Бордаченкова Е.А., Васильева Н.Э. Морозов С.С. Лексико-синтаксические шаблоны в задачах автоматической обработки текста // Компьютерная лингвистика и интеллектуальные технологии: Труды межд. конф. «Диалог 2007». – М.: Изд-во РГГУ, 2007. – С. 70-75.

[8] Лебедев А.С. Естественно-языковое программирование как средство извлечения знаний // Труды ИВМиМГ, Информатика, Вып. 9. – Новосибирск, 2009. – С. 64-71.

[9] Андреев А.М., Березкин Д.В., Симаков К.В. Модель извлечения знаний из естественно-языковых текстов // Информационные технологии. – 2007. – №12. – С. 57–63.

[10] Абрамов А.П., Выдрук Д.Г., Федунов Б.Е. Компьютерная система оценки реализуемости алгоритмов деятельности экипажа // Изв. РАН. Теория и системы управления. – 2006. – № 4. – С. 122-134.

[11] Katz G., Giesbrecht E. Automatic identification of non-compositional multi-word expressions using latent semantic analysis // Proc. of Multiword Ex-

pressions: Identifying and Exploiting Underlying Properties, Sydney, 2006. – P. 12-19.
[12] Mladenic D. Text-learning and related intelligent agents: a survey// IEEE Intelligent Systems. – 1999. – V. 14, No 4. – P. 44-54.
[13] Caropreso F., Matwin S. Beyond the bag of words: a text representation for sentence selection// Proc. of AI2006, Québec, QC. – P. 324-335.
[14] Ягунова Е.В., Пивоварова Л.М. Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов// Сб. НТИ, Сер. 2. – М., 2010. – № 5.
[15] Невзорова О.А. Подход к разработке методов автоматизированного контроля информационной целостности технических текстов //Труды десятой национальной конф. по искусственному интеллекту КИИ-2006. Т. 2. – М.: Физматлит, 2006. – С. 564-571.
[16] Невзорова О.А., Федунов Б.Е. Система анализа технических текстов «ЛоТА»: основные концепции и проектные решения // Изв. РАН. Теория и системы управления. – 2001. – № 3. – С. 138-149.
[17] Заболотняя Т.Н., Михайлюк А.Ю., Михайлюк Е.С. Инверсионный контекстно-ассоциативный метод автоматической орфокоррекции // Штучний інтелект. – Киев, 2008. – № 3. – С. 78-88.
[18] Budanitsky A., Hirst G. Evaluating WordNet-based measures of lexical semantic relatedness // Computational Linguistics. – 2006. – V. 32. – No 1. – P. 13-47.
[19] Sripada S., Reiter E., Davy I., Nilssen K. Lessons from deploying NLG technology for marine weather forecast text generation//Proc. of PAIS-2004, 2004. – P. 760-764.

Some perspective methods of project documentation processing

E.S. Klyshinsky

This paper describes some of tasks that can dramatically change the project documentation creation process. There discussed tasks like automated requirements extraction, precedent documents searching for the analysis and development stage support purposes, documentation fullness check on final stages, automatic documentation generation about some part of designed product, automated components selection for designed product.

♣ Работа выполнена при финансовой поддержке Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России на 2009-2013 годы»