Об использовании мер сходства при анализе документации♣

© А.Ю. Антонова, Э.С. Клышинский

Московский государственный институт электроники и математики
a.yu.antonova@gmail.com, klyshinsky@itas.miem.edu.ru

Аннотация

В ходе проектирования продукции подготавливается большое количество полнотекстовой документации, из которой можно выделить две части: постановку задачи (ТЗ, собственно постановка задачи, тех. условия и проч.) и итоговую документацию. Используя различные меры сходства и их комбинации, имеется возможность решать такие задачи, как поиск аналогичных документов по тексту технического задания, проверку полноты итоговой документации и так далее.

1. Введение

В настоящее время благодаря высокому развитию технологий большая часть современных крупных компаний уже осуществила переход на ведение безбумажного документооборота. Широкое внедрение получили системы, использующие концепцию и технологии ILM (системы, обеспечивающие хранение данных и оптимальное время доступа к ним) [3], а также PDM-системы [1, 9]. Система PDM управляет обменом данными об изделии, обеспечивает взаимодействие с любыми корпоративными приложениями и отслеживает внесения изменений в изделие и документацию о нем. Кроме того, существующие системы бизнес-аналитики осуществляют интеллектуальную обработку хранимых данных. Тем не менее, насколько нам известно, до сих пор не создано эффективных систем, обрабатывающих текстовую составляющую информации об изделии.

Однако очевидно, что вследствие постоянного увеличения объёма хранимых электронных данных, важнейшей задачей является организация эффективного поиска внутри документации компании. Так, например, начиная разработку нового продукта, разработчики должны провести анализ существующих решений. На крупном предприятии с большой историей производства в первую очередь изучаются собственные архивы. Поисковые машины, осуществляющие поиск по имеющейся документарной базе, как правило, предоставляют возможность формулирования запроса с помощью ключевых слов. Это ставит перед пользователем ряд проблем.

За годы использования документации введенная вначале терминология могла измениться и быть заменена новой. Тем не менее, при поиске релевантных документов пользователь должен получать все документы, в том числе и те, в которых, используется начальная терминология. Более того, существует и другая трудность. Мы не можем с уверенностью утверждать, что пользователь поисковой системы введет в качестве поискового запроса именно те термины, которые употребляются в нужном документе. А ввод синонимов значительно увеличит нагрузку на пользователя и снизит общую релевантность выдачи поискового механизма.

Наиболее уязвимы в данном случае однословные термины. В отличие от них, многословные конструкции обладают большей устойчивостью. В нашем подходе мы исходили из соображения, что, несмотря на определенную степень изменения терминологической лексики и даже системы понятий в более поздней документации, для всех релевантных документов в целом сохраняется большим множество общих слов.

В соответствии с этим, предлагается применить следующий подход: проводить поиск не по ключевым словам, а по характерной лексике. Для этого выбирается т.н. документ-образец, из которого с помощью изложенных ниже принципов выделяются неоднословные сочетания, которые в дальнейшем используются для прецедентного поиска внутри документации. Поскольку мы имеем дело с технической документацией, то в качестве документа-образца используется техническое задание (ТЗ). Техническое задание содержит в себе краткое, но полное описание постановки задачи, а также методов ее решения. Таким образом, в нем обязательно присутствует вся основная терминология.

2. Постановка задачи

Из множества предложенных ранее задач (см. [7]) мы выбрали для своего исследования две.

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

Первая из них – это поиск документов, которые могли бы быть полезны в текущем проекте, с использованием ТЗ как запроса. Для ее решения требуется разработать метод, который позволил бы вычислять меру сходства документов. На основе использования этой меры можно проводить ранжирование документов и подавать на выход наиболее релевантные. Данная мера должна учитывать и многословные конструкции, встречающиеся в тексте документации.

Вторая задача – это проверка полноты итоговой документации по поставленному ТЗ. Для этого необходимо автоматически выделять из текста ТЗ исходные данные и требуемые результаты для разрабатываемого продукта. На основе этой информации и проводится автоматическая проверка документации: насколько полно требования, указанные в ТЗ, фигурируют в итоговом отчете и в каких местах документации приведено их описание. Подобным образом мы можем оценить степень полноты документации в процентах от поставленных задач. При снижении полученного значения ниже порогового, документация признается неполной, и мы имеем возможность указать, какие именно вопросы из ТЗ отсутствуют в итоговой документации. В последнем случае исполнитель должен будет переделать документацию и, возможно, пересмотреть свою работу.

С точки зрения потребителя, система должна проводить предварительное исследование итоговой документации на предмет явного отсутствия описания некоторых вопросов, описанных в ТЗ. Список требований может корректироваться экспертом вручную. Система не предусматривает явного противодействия ей, когда разработчики включают в итоговую документацию «шум», имитирующий фрагмент отчета. Таким образом, отрицательный ответ системы должен рассматриваться как повод для направления документации на согласование или доработку. Положительный ответ системы служит сигналом для ее проверки принимающей стороной, так как гарантирует лишь наличие информации в отчете, но не ее глубину.

3. Существующие методы

В настоящее время разработано немалое количество различных методов, используемых для решения сходных задач. Ниже мы рассмотрим некоторые из них, распределенные по трем группам.

Поиск по ключевым словам и словосочетаниям может производиться различными методами.

Традиционным является использование статистической меры tf*idf, показывающей «информативность термина» [10]. Tf (term frequency) представляет собой отношение числа вхождений некоторого термина ко множеству всех слов в документе. Idf (inverse document frequency) – инверсия частоты, с которой то же слово встречается во всех документах рассматриваемой коллекции. В целом эта мера учитывает степень важности термина внутри одного документа, снижая вес для общеупотребительной лексики (имеющей высокую частотность во всех документах рассматриваемой коллекции). Термин с высокой встречаемостью считается малоинформативным, так как, скорее всего, является элементом стилистики. В то же время редко встречающийся термин в рамках данного подхода считается элементом шума и получает низкое значение весовой характеристики.

Для словосочетаний наряду с мерой tf*idf используется метод выделения коллокаций (неслучайных сочетаний двух и более лексических единиц), например, меры MI, t-score или log-score [5]. С их помощью может быть получен набор неоднословных сочетаний, характерных для рассматриваемой предметной области. Для этого, как показали исследования, описанные в [11], лучше подходит мера MI. Мера t-score позволяет скорее выявить стилистические особенности текста и, с этой точки зрения, в рамках нашей задачи может служить скорее для выявления терминов, подлежащих отсеву.

Методы, основанные на модели векторного пространства, используют формируемый заранее вектор признаков документа, включающий в себя список наиболее значимых терминов. Дальнейшее сравнение документов проводится путем оценки сходства между полученными векторами с использованием различных мер сходства. В данной работе мы рассмотрим часто применяемые коэффициент Дайса и косинусную меру [10]:

$$\mathrm{Dice}(\mathbf{x}, \mathbf{y}) = 2*\frac{|\mathbf{xy}|}{|\mathbf{x}|+|\mathbf{y}|}, \quad (1)$$

$$\mathrm{Cos}(\mathbf{x}, \mathbf{y}) = \frac{|\mathbf{xy}|}{\sqrt{|\mathbf{x}|*|\mathbf{y}|}}, \quad (2)$$

где **x** и **y** – два сравниваемые документа;

|**x**| и |**y**| – мера встречаемости слов в каждом из документов, определяемая как скалярное произведение векторов частот встречаемости слов соответствующих документов (x или y);

|**xy**| – мера совместной встречаемости слов документов **x** и **y**, определяемая как скалярное произведение векторов, соответствующих этим документам.

Для формирования вектора признаков документа обычно используется некоторое подмножество слов. Помимо уже упоминавшейся меры tf*idf могут использоваться, например «опорные слова» [12]. Из множества слов всех документов коллекции выбирается N «опорных» и для каждого документа формируется N-мерный вектор признаков. В вектор заносится 1, если частота данного слова в документе превышает некоторое пороговое значение, и 0 в противном случае. Сходными считаются документы, у которых векторы признаков совпадают.

Кроме перечисленных, на данный момент существует целый ряд мер, использующих

небольшие **фрагменты документов для их сравнения** или их оценки. Так, например, группа алгоритмов, базирующихся на алгоритме случайных полиномов Карпа-Рабина [2], опирается на обнаружение сходных групп последовательно идущих слов длины *k* (шинглов) или дактилограмм – подстрок документа фиксированной длины. Для документа вычисляется фиксированное количество шинглов или дактилограмм, после чего проводится сравнение подобных последовательностей. Для более быстрых методов применяется вычисление хеш-функции для полученных последовательностей. Документы считаются сходными, если полученное значение хеш-функции совпадает. Более медленные, но более точные методы сравнивают полученные последовательности и на основании оценки меры их совпадения делается вывод о степени и вероятности сходства документов. Подробный обзор этих и других методов можно найти в [6].

4. Подход к решению

4.1 Оценка сходства

Вместо традиционно используемых мер оценки сходства документов, таких как косинусная мера, коэффициент Дайса и др., учитывающих относительную встречаемость слова в документе, в нашем подходе оценка сходства ведется с помощью альтернативных мер, предложенных в статье [8].

Формулы (1) и (2) могут быть модифицированы к виду (3) и (4) соответственно. Введем две новые меры: упрощенная косинусная мера сходства (simplified cosine) (3), которая рассматривает сочетания слов длины *n* без учета частоты их встречаемости, а также упрощенная мера Дайса (4):

$$s\_cos(\mathbf{x},\mathbf{y},n) = \frac{\|\mathbf{xy}\|_n}{\sqrt{\|\mathbf{x}\|_n * \|\mathbf{y}\|_n}} \text{ и} \qquad (3)$$

$$s\_Dice(\mathbf{x},\mathbf{y},n) = 2 * \frac{\|\mathbf{xy}\|_n}{\|\mathbf{x}\|_n + \|\mathbf{y}\|_n}, \qquad (4)$$

где $\|\mathbf{x}\|_n$ – количество выделенных словосочетаний длины *n* в документе **x**,

$\|\mathbf{xy}\|_n$ – количество словосочетаний длины $n>0$, имеющихся как в документе **x**, так и в документе **y**.

Однако, как показали эксперименты, не менее убедительные результаты показали и две другие меры.

Первую из них будем называть несимметричной упрощенной мерой сходства:

$$NSL(\mathbf{x},\mathbf{y},n) = \frac{\|\mathbf{xy}\|_n}{\|\mathbf{x}\|_n}. \qquad (5)$$

Эта мера показывает относительный объем совпадающей лексики документов **x** и **y** в документе **x**. Очевидно, что подобная мера будет несимметричной, то есть NSL(**x**, **y**, *n*) ≠ NSL(**y**, **x**, *n*). Для того чтобы избавиться от несимметричности, просуммируем меру (5), нормированную на первый и второй документы, что даст нам симметричную упрощенную меру сходства:

$$SSL(\mathbf{x},\mathbf{y},n) = \frac{\|\mathbf{xy}\|_n}{\|\mathbf{x}\|_n} + \frac{\|\mathbf{xy}\|_n}{\|\mathbf{y}\|_n}. \qquad (6)$$

Простые арифметические преобразования приводят (6) к следующему виду:

$$SSL(\mathbf{x},\mathbf{y},n) = \|\mathbf{xy}\|_n * \frac{\|\mathbf{x}\|_n + \|\mathbf{y}\|_n}{\|\mathbf{x}\|_n * \|\mathbf{y}\|_n}. \qquad (7)$$

Так как правая часть произведения в (7) показывает вероятность пересечения списков, то мера в целом отражает удвоенную вероятность нахождения $\|\mathbf{xy}\|_n$ словосочетаний в обоих списках.

Также использовались модификации косинусной меры и меры Дайса, использующие не только отдельные слова, но и их комбинации.

$$Dice(\mathbf{x},\mathbf{y}) = 2 * \frac{|\mathbf{xy}|_n}{|\mathbf{x}|_n + |\mathbf{y}|_n}, \qquad (8)$$

$$Cos(\mathbf{x},\mathbf{y}) = 2 * \frac{|\mathbf{xy}|_n}{\sqrt{|\mathbf{x}|_n * |\mathbf{y}|_n}}. \qquad (9)$$

Здесь $|xy|_n$ – скалярное произведение векторов, содержащих в себе частоты встречаемости сочетаний из n слов в документах x и y. Для вычисления значений брались все группы из расположенных рядом n слов, встретившиеся более одного раза.

Каждая из этих мер хорошо показала себя при оценке сходства документов. Однако, для различных задач следует использовать разные меры определения сходства и их комбинации. Так, например, поскольку перед нами стоит задача не просто оценить степень сходства, но и определить, в какой степени выделенная лексика из одного документа (ТЗ) встречается в других документах, то наиболее целесообразным является использование несимметричной меры (5). В задачах поиска документов, наиболее релевантных ТЗ в целом, лучшим образом себя показали сочетания мер на основе (3).

4.2 Метод поиска документов

Меры (1)-(9) могут использоваться для определения меры сходства документов между собой. В качестве запроса подается документ, для которого мы осуществляем поиск по сходству. Вычисленные значения меры позволяют ранжировать документы из хранилища по релевантности. Очевидно, что меры будут давать различные результаты. При учете специфики задачи мы исходили из следующих простых соображений. Различные задачи требуют разного использования информации, содержащейся в документе. Так, при поиске плагиата или неявных дубликатов следует обращать внимание на частоты употребления терминов, так как одни и те же термины могут использоваться в различном порядке и разных фрагментах. Однако тематическое сходство документов основывается скорее на полноте совпадения лексики, чем сходстве предложений внутри текста. В связи с этим для данной задачи скорее должны подходить меры (3)-(7), не учитывающие относительную частоту встречаемости терминов, чем (1), (2), (8) и (9). Также следует использовать информацию о многословных конструкциях, лучше описывающих информацию о предметной области. Кроме того,

комбинация различных мер может давать лучшие результаты, чем использование каждой меры по отдельности.

Как следствие, требовалось провести серию экспериментов, определяющих пригодность той или иной меры для задачи поиска тематически сходных технических документов. Результаты этих экспериментов приведены ниже.

4.3 Поиск по шаблонам

На основе эмпирического анализа текстов технической документации, содержащей ТЗ или постановку задачи, было замечено, что довольно распространенными в тексте являются глагольные конструкции определенного вида. Предполагается, что внутри предложений с такими конструкциями содержится информация, касающаяся требований, указанных в ТЗ, и/или описание их реализации. Примеры таких глагольных конструкций: «*предусматривает следующие действия»,* «*осуществляется в соответствии с*», «*должен контролировать*» и др. В ряде случаев можно искать лишь глаголы, определяющие требования к системе: *«требуется», «следует», «должен»*. Вся обработка документов, в том числе и поиск шаблонов, проводилась с использованием подсистемы морфологического анализа «Кросслейтор» [4].

В соответствии с данным предположением сравнение документации проводится следующим образом. Заранее определяется список глагольных конструкций, характерных для технических текстов и текстов выбранной предметной области и описывающих требования к продукту. Из имеющихся конструкций формируется список шаблонов. Далее осуществляется поиск отобранных конструкций в тексте документа-образца (в нашем случае это ТЗ). Поскольку нас интересует лексика, характеризующая данный документ, то для дальнейшего анализа из текста выбирается отрезок, больший, чем предложение, непосредственно содержащее найденную конструкцию. Определение размера т.н. «окна поиска» представляет собой отдельную задачу. В первом приближении было решено использовать для дальнейшей обработки интервал, включающий, помимо найденного предложения, еще два соседних с каждой стороны. Внутри этого интервала с наибольшей вероятностью будет встречаться тематически значимая лексика, поскольку трудно предположить, что описание задачи / требований будет выражено в одном предложении.

Нами была поставлена задача получить максимальную полноту при умеренном проценте шума, т.е. чтобы по каждому пункту требований из ТЗ найти по возможности *все* соответствия в тексте итоговой документации, избегая при этом ложных совпадений. Таким образом, приоритет отдается скорее полноте выдачи, чем точности.

5. Промежуточные результаты

Для проведения экспериментов по поиску сходных документов была собрана коллекция из 450 документов различной тематики. В качестве основы коллекции были взяты научные статьи, диссертации и авторефераты диссертаций из различных отраслей науки, произведения художественной литературы. Документы были атрибутированы экспертом по 26 кластерам. В качестве показателя бралась f-мера как удвоенное среднее гармоническое точности и полноты. На вход подавался документ из коллекции, на выходе получались 15 документов, ранжированных по релевантности. Полнота рассчитывалась как отношение документов, принадлежащих тому же кластеру, к количеству документов кластера. Точность рассчитывалась как количество документов, принадлежащих тому же кластеру к количеству документов на выходе (то есть, 15). В итоге были получены следующие средние значения f-меры: s_cos(1) = 0.449; s_cos(2) = 0.491;
s_cos(1) + s_cos(2) + s_cos(3) = 0.498; cos(1) = 0.555; cos(2) = 0.481; cos(1) + cos(2) + cos(3) = 0.576; cos(1)+ cos(2)+ cos(3)+ s_cos(1)+ s_cos(2)+ s_cos(3) = 0.581.

Результаты экспериментов показали, что комбинация косинусной и модифицированной косинусной меры позволяют получить более релевантные результаты, чем при использовании исходных формул. Однако более оптимальным с точки зрения вычислительных затрат представляется использование в качестве критерия суммы косинусных мер по сочетаниям в 1-3 слово.

В качестве материала для проведения эксперимента по определению полноты документации использовалась коллекция технической документации, имеющаяся в нашем распоряжении. В состав коллекции входят несколько десятков документов, содержащих в себе ТЗ и итоговые отчеты по выполнению работ. При анализе документов осуществляется морфологическая обработка текста с помощью анализатора [4], дальнейшая работа проводится только над нормальными формами.

Исходя из перечисленных выше соображений, была реализована программа, позволяющая говорить о следующих промежуточных результатах. На основании лексики, извлекаемой из текста технического задания с использованием методики шаблонов, по всем текстам коллекции итоговой документации ищутся соответствия фрагментам из ТЗ. Найденные соответствия сравниваются с исходными с помощью мер сходства (упомянутые NSL, SSL, s_Cos, s_Dice). Числовые характеристики, которые мы получили по этим мерам, позволяют сделать однозначный вывод о тематической связанности исходного ТЗ и соответствующей ему коллекции документации.

Кроме того, был проведен опыт ранжирования документов по каждой из используемых мер (NSL,

SSL, а также модифицированным косинусной мере и коэффициенту Дайса). Это делалось для того, чтобы убедиться в том, что рассматриваемый метод не дает высокой оценки нерелевантным документам. Системе было предложено сравнить исходный документ (ТЗ) с коллекцией других. В коллекции, помимо документов, соответствующих данному ТЗ, были представлены заведомо нерелевантные документы (как из близких, так и из несвязанных предметных областей). Эксперименты показали, что в каждом случае наивысшую степень сходства получали документы, соответствующие ТЗ, причем числовые характеристики, получаемые по мерам, отличались, в случае релевантных документов, на порядок.

Эти результаты подтверждают целесообразность дальнейшего развития выбранного подхода. Как следствие, требуется разрешение целого ряда проблем, связанных с обеспечением качества работы метода.

Литература

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

[2] Гасфилд Д. Строки, деревья и последовательности в алгоритмах. СПб.: Невский диалект, 2003. 654 с.

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

[4] Елкин С.В., Клышинский Э.С., Стекляников С.Е., Проблемы создания универсального морфосемантического словаря // Сб. трудов Международных конференций IEEE AIS'03 и CAD-2003, том 1, Дивноморское. 2003. стр. 159-163.

[5] Захаров В.П., Хохлова М.В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (2010). Вып. 9 (16), сс. 137-143

[6] Зеленков Ю.Г, Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для Web-документов // сб. трудов Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции - 2007», Переславль-Залесский, 2007. – Том 1, С. 166-174.

[7] Клышинский Э.С. Перспективные методы обработки проектной документации// Труды 12-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010.

[8] Клышинский Э.С. Методика поиска документов в документарных системах. Новые информационные технологии в автоматизированных системах: материалы четырнадцатого научно-практического семинара. - Моск. гос. ин-т электроники и математики. М., 2011, стр. 58-67

[9] Колчин А. Что такое PDM? // PC Week, №38, 2001.

[10] Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск // Вильямс. М.: 2011. 528 с.

[11] Пивоварова Л., Ягунова Е. Извлечение и классификация терминологических коллокаций на материале лингвистических научных текстов // Терминология и знание: Материалы II Международного Симпозиума (Москва, 21-22 мая 2010г.) – М., 2010.

[12] S. Ilyinsky, M. Kuzmin, A. Melkov, I. Segalovich. An efficient method to detect duplicates of Web documents with the use of inverted index // In Proc. 11th Int. Conf. on World Wide Web, 2002.

On Using of Similarity Measures During Documentation Analysis

© A.Yu. Antonova, E.S. Klyshinsky

A large amount of documentation is made while a product is designed. This documentation contains two big classes: task statement (requirements specification, technical requirement and so on) and final documentation. Using different similarity measures one can solve tasks like precedent documentation searching using task statement, final documentation completeness detecting and so on.

♣ Работа выполнена при финансовой поддержке гранта РФИИ № 11-01-00793.