OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

ОНТОЛОГИЧЕСКИ-ОРИЕНТИРОВАННАЯ МОДЕЛЬ КЛАССИФИКАЦИЙ ТЕКСТОВЫХ ДОКУМЕНТОВ

Наместников А.М., Субхангулов Р.А.[*]

[*] *Ульяновский государственный технический университет, г. Ульяновск, Россия*

nam@ulstu.ru,

subkhangulov-ruslan@yandex.ru

В статье рассматривается специфика информационных потребностей специалистов занятых в процессе проектирования технических средств. Представлено описание пользовательских профилей, их применение в процессе полнотекстового поиска текстовых документов. Рассматривается применение байесовского классификатора в терминах онтологических моделей информационного поиска. Приведены результаты экспериментальных исследований, доказывающие эффективность разработанной программной системы.
Ключевые слова: интеллектуальная система; онтология; полнотекстовый поиск, классификация.

Введение

Внедрение систем автоматизированного проектирования (САПР) CAD/CAM/CAE в современное производство - это пройденный этап автоматизации, который повысил качество и скорость разработки продукции. На современном этапе автоматизации проектирования стоят новые задачи к ним относятся использование средств PLM (Product life Cycle) поддержка жизненного цикла изделия, формирования систем электронного архива проектно-технической документации.

Основными задачами электронного архива является обеспечение коллективной работы проектно-конструкторских отделов над проектом, добавление, хранение и поиск данных в электронном архиве. Поиск знаний в таких архивах выполняется с использованием классических моделей поиска, к ним относятся поиск по реквизитам документов, поиск по ключевым словам. Для профессионального, в том числе научно-технического поиска информации требуется обеспечение поиска, основанного на знаниях, – использование синонимов, возможности автоматического расширения запроса, возможностей автоматического анализа результатов запроса и помощь в интерактивном поиске. Для решения подобного рода проблем применяются интеллектуальные модели поиска, функционирование которых основано на предметно-ориентированных знаниях. Эти знания могут быть представлены в виде онтологии предметной области [Добров и др., 2006]. В настоящее время наблюдается рост интереса к разработке новых методов интеллектуального анализа данных, основанных на применении онтологии [Гаврилова и др., 2000].

В данной статье предлагается использовать пользовательские профили, в основе которых лежит онтология предметной области. Профили содержат информационные предпочтения специалистов-проектировщиков. Они играют вспомогательную роль в задачах интерактивного поиска слабоструктурированных данных, что позволяет максимально удовлетворить информационную потребность пользователя электронного архива проектно-технических документов.

1.1. Информационная неопределенность проектировщика

Прежде чем рассмотреть пользовательские профили проектировщика и использование их в поддержке информационного поиска рассмотрим такой важный аспект, как информационную потребность специалиста-проектировщика, выделим специфику, которым обладает данная характеристика.

Дадим определение информационной потребности: «Информационная потребность – это информационная неопределенность, которую пользователь хочет уменьшить посредством получения информации из системы информационного поиска».

Для понимания специфики информационной потребности проектировщика необходимо детально рассмотреть процесс проектирования. Используем определение, которое дано Норенковым И.П. в работе [Норенков, 2009]: «Проектирование технического объекта – это создание, преобразование и представление в принятой форме образа этого еще не существующего объекта». В нашем случае проектирование необходимо рассматривать как информационный процесс, в котором выполняется преобразование входной информации о проектируемом объекте в выходную информацию в виде проектных документов.

Проектная деятельность имеет ряд специфических особенностей [Титов, 2009]:

1. Продуктом проектной деятельности является упорядоченная совокупность сведений, служащих знаковой моделью объекта, в момент проектирования реально еще не существующего.

2. Процедуры проектирования реального объекта соответствуют преобразованию его исходного описания в некотором конечном пространстве.

3. Способы преобразования информации при проектировании нельзя отразить в виде математических соотношений, т. е. невозможно построить строгую математическую модель такого процесса преобразования.

4. Ввиду сложности проектируемых объектов на каждом этапе разработки вовлекаются различные специалисты, что придает проектированию характер коллективной деятельности.

5. Проектируемый объект входит в упорядоченную иерархию объектов и, с одной стороны, выступает как часть системы более высокого уровня, а с другой — как система для объектов более низкого уровня. В соответствии с этим процесс проектирования можно разделить на два этапа: внешнего (объект — часть системы более высокого ранга) и внутреннего проектирования (объект — совокупность компонентов).

6. Проектирование, как правило, имеет итерационный многовариантный характер, для принятия проектных решений используются различные научно-технические знания.

1.2. Профили проектировщика

В традиционных поисковых системах поиск данных выполняется следующим образом: пользователь имеет информационную потребность, которая затем преобразуется в виде набора терминов в поисковый запрос к подсистеме информационного поиска. Подсистема информационного поиска, используя традиционные модели и алгоритмы, находит документы, которые должны уменьшить информационную потребность пользователя. Однако данная подсистема не имеет полного представления об информационных потребностях пользователя и тем самым всегда присутствует вероятность того, что документы, которые были отобраны, не уменьшат информационную неопределенность.

Для решения задачи максимального удовлетворения информационной потребности специалистов-проектировщиков в работе [Филиппов и др., 2013] предлагается использовать модели кластеризации и информационного поиска, в основе, которых лежит предметно-ориентированная онтология. В данной работе не будем рассматривать онтологические методы поиска и кластеризации. Отметим лишь, что формально предметная онтология имеет следующий вид [Наместников, 2009**,** Филиппов и др., 2013]:

$$\Theta = \langle r, S, C, W, R \rangle,$$

где r – корневая вершина онтологии, соответствующая классу проектных документов; S – множество структур документов, $C = \{c_1, c_2, \ldots, c_k\}$ – множество понятий предметной области электронного архива, $W = \{w_1, w_2, \ldots, w_l\}$ – множество терминов предметной области электронного архива, R – множество отношений онтологии.

В онтологии каждый термин $w_i \in W$ связан с понятием $c_j \in C$ отношением ассоциации $R_A^a : w_i R_A^a c_j$. Понятия C связаны друг с другом различными видами отношений (отношение обобщения, отношение включения и т.д.), образуя таксономию понятий предметной области. Понятийный уровень онтологии представим в виде ориентированного графа:

$$G = (C, E), \qquad (1)$$

где C – множество вершин графа, каждая вершина – это понятие онтологии; E – множество дуг вида

$$E = \{\langle c_i, c_k \rangle\}$$

для всех $c_i, c_k \in C$, для которых имеет место отношение $c_i R_G c_k$.

Онтологическое представление oV_j^d является вершинным подграфом графа G, определяемого выражением (1). Поскольку, в общем случае, в онтологическом представлении могут отсутствовать любые понятия из состава понятий предметной онтологии, результирующий граф онтологического представления oV_j^d может состоять из несвязанных деревьев и/или изолированных понятий.

В результате получаем, что проектно-технический документ в электронном архиве представляется не в лексическом пространстве терминов, которые удается выделить в документе, а в пространстве понятий предметной области, которые зафиксированы в онтологии электронного архива.

В данной работе предлагается дальнейшее развитие темы использования предметной онтологии в задачах информационного поиска. В частности использования пользовательских профилей, в основе которых лежит онтология предметной области электронного архива. Пользовательский профиль объединяет различных пользователей по следующим признакам:

- Пользователи одного проектно-конструкторского отдела;
- По должностям;
- Пользователи, которые работают над одним проектом.

Такие профили формируют своеобразный «портрет» группы пользователей, объединенные по определенному признаку. Структура профиля состоит из двух категорий, формально представляется в следующем виде:

$$Pr = \{Tr_+, Tr_-\}$$

где Tr_+ – множество, содержащее информацию о документах, которые принадлежат пользовательскому профилю, Tr_- – множество, содержащее информацию о документах, которые не принадлежат профилю.

Множества представляются следующим образом:

$$Tr_i = \{c_{ij}, d_{il}\}$$

где $i \in \{+,-\}$, c_{ij} – j-ый доминирующий концепт из предметной онтологии в i- ой категории, d_{il} - технический документ из электронного архива.

Рассмотрим детально, каким образом формируются пользовательские предпочтения. В работе [Филиппов и др., 2013] выдвигается гипотеза о том, что любой текстовый документ можно разделить на множество непересекающихся фрагментов, в каждом из которых будет доминировать тот или иной концепт предметной области. Для нахождения значения доминирования концептов применяется метод сравнения текстового входа каждого понятия в онтологии предметной области с анализируемым текстом.

Алгоритм вычисления степени доминирования понятия в текстовом фрагменте состоит из следующих шагов [Филиппов и др., 2013]:

Шаг 1. Определение максимальной степени выраженности концептов в текстовом фрагменте:

$$\hat{\mu}_{S_p^d}(c) = \max_c(\mu_{S_p^d}(c)).$$

Шаг 2. Определение среднего значения степени выраженности концептов онтологии, исключая концепт с максимальной степенью выраженности (определенный на предыдущем шаге):

$$\tilde{\mu}_{S_p^d}(c) = \frac{1}{n-1}\sum_{i=1}^{n-1}\mu_{S_p^d}(c_i),$$

где $c_i \in c - c_k$, $c_k = \arg\max_c(\mu_{S_p^d}(c))$, n – количество концептов с ненулевой степенью выраженности для текстового фрагмента S_p^d.

Шаг 3. Определение степени детерминированности понятия в текстовом фрагменте S_p^d:

$$\Delta_{S_p^d}(c) = \hat{\mu}_{S_p^d}(c) - \tilde{\mu}_{S_p^d}(c), \quad (2)$$

Выражение (2) фактически определяет качество выделения текстового фрагмента в проектно-технических документах с целью ограничения в тексте определенного понятия предметной области, которое зафиксировано в онтологии электронного архива.

Опираясь на эту гипотезу, а также воспользуемся алгоритмом «Вычисления степени доминирования концепта в текстовом фрагменте», разработаем метод, который добавляет доминирующие концепты в категории пользовательского профиля. Рассмотрим данный алгоритм более подробно по этапам:

Шаг 1. При входе пользователя в поисковую систему активируется профиль, к которому принадлежит пользователь.

Шаг 2. Поисковая система, в ответ на пользовательский запрос, выдает список релевантных документов. В основе информационного поиска лежит онтологически-ориентированная модель [Филиппов и др., 2013].

Шаг 3. Специалист-проектировщик, просматривая отобранный поисковой системой документ, отмечает, удовлетворяет или не удовлетворяет документ его информационным потребностям.

Шаг 4. Подсистема находит доминирующие концепты в отмеченном документе. Алгоритм вычисления степени доминирования понятия в документе рассмотрен выше.

Шаг 5. Если документ отмечен положительно, то есть удовлетворяет потребностям проектировщика, тогда этот документ рассматривается как документ, который принадлежит пользовательскому профилю и, следовательно, доминирующие концепты и наименование документа добавляются подсистемой информационного поиска в категорию Tr_+. Если документ отмечен отрицательно, то есть не удовлетворяет потребностям пользователя, тогда данный документ рассматривается как документ, который не принадлежит пользовательскому профилю и, следовательно, доминирующие концепты и наименование документа добавляются

подсистемой информационного поиска в категорию $Tr\_$ пользовательского профиля.

Таким образом, происходит формирование обучающей выборки пользовательского профиля, который в дальнейшем будет использоваться в задачах классификации и информационного поиска.

1.3. Модель классификации документов

Рассмотренный профиль пользователя содержит профессиональную информацию о пользовательских предпочтениях. Таким образом, подсистема информационного поиска имеет представление об информационных потребностях пользователя. Тем самым предполагается, что снизится неопределенность о том, какую информацию хочет получить пользователь. Это позволит улучшить качество информационного поиска в условиях неопределенности.

Будем применять пользовательские профили в рамках решения задач классификации в условиях неопределенности с использованием математического аппарата теории вероятностей.

Применительно к данной работе постановка задачи классификации документов выглядит следующим образом. Пусть даны: множество проектно-технических документов электронного архива $d \in \{d_1, d_2 ... d_n\}$ и множество пользовательских профилей $pr \in \{pr_1, pr_2 ... pr_m\}$. Множество документов имеет большую размерность, чем множество профилей. Количество профилей задается экспертом, по каким критериям могут формироваться пользовательские профили, рассмотрены выше в данной статье. Кроме того, имеется обучающее множество:

$$< d, pr > \in D \times Pr$$

Информация в профиле является динамической, то есть изменяется в процессе работы пользователей с подсистемой информационного поиска. Следовательно, результаты классификации проектных документов также будут изменяться в процессе работы всех пользователей с подсистемой информационного поиска, которые состоят в одном профиле. Используя метод обучения, получим функцию классификации, которая отображает документы в пользовательские профили:

$$\gamma : D \to Pr$$

В качестве метода обучения будет использоваться наивный метод Байеса, который имеет следующий вид:

$$P(pr \mid d) = P(pr) \prod_{1<i<n} P(c_i \mid pr)$$

где $P(c_i \mid pr)$ – условная вероятность того, что концепт c_i из профиля будет доминирующим в документе, $P(pr)$ – априорная вероятность того, что документ принадлежит профилю.

Так как необходимо рассчитать максимальную апостериорную вероятность величины $P(pr \mid d)$, то получим следующее выражение:

$$c_{map} = \arg\max_{pr \in \Pr} (P(pr) \prod_{1<i<n} P(c_i \mid pr))$$

Величины $P(d)$ и $P(d \mid pr)$ вычисляются следующими выражениями:

$$P(d) = \frac{N_{Tr_i}}{N}$$

где N_{Tr_i} – количество документов в категории профиля, N – общее количество документов в пользовательском профиле *pr*.

Как было отмечено раннее, в данной работе документы имеют концептуальное представление, в виде множеств концептов и степени выраженности концептов в документе. Следовательно, условную вероятность будем вычислять как относительную частоту концепта в документах обучающей выборки, принадлежащих профилю *pr*.

$$P(d \mid pr) = \frac{k_i}{\sum_{i' \in V} k_{i'}}$$

где k_i – количество появлений концепта *c* в в категории профиля, V список всех уникальных концептов в профиле.

Может оказаться так, что в обучающей выборке отсутствует концепт, который является доминирующим в анализируемом документе. Для того чтобы, решить данную проблему воспользуемся сглаживанием Лапласа. К каждой частоте добавляется единица. Таким образом, окончательно формула для вычисления условной вероятности будет выглядеть следующим образом:

$$P(d \mid pr) = \frac{k_i + 1}{\sum_{i' \in V} k_{i'} + \mathrm{B}}, \qquad (2)$$

где k_i – количество появлений концепта *c* в категории профиля, $\mathrm{B} = |\mathrm{V}|$ – список всех уникальных концептов в профиле.

Рассмотрим алгоритм поисковой системы, основанный на пользовательских профилях по этапам:

Шаг 1. При входе пользователя в поисковую систему активируется профиль, к которому принадлежит пользователь.

Шаг 2. Классификационная модель, обладая информацией о предпочтениях пользователя, в реальном времени формирует дополнительный список релевантных документов, которые возможно

удовлетворят информационную потребность пользователя, следующим образом:

Шаг 2.1. На вход классификационной модели поступает концептуальные представления документов электронного архива.

Шаг 2.2. Для каждого документа с помощью выражения (2) и категории Tr_+ активного профиля вычисляется вероятность принадлежности документа к профилю, затем с помощью выражения (2) и Tr_- вычисляется вероятность того, что документ не принадлежит профилю.

Шаг 2.3. Полученные на этапе (2.2) вероятности сравниваются между собой, если значение вероятности принадлежности к профилю оказывается выше, то анализируемый документ отбирается в дополнительный список.

Шаг 3. Сформированный на втором этапе список дополнительных документов выводятся пользователю.

2. Результаты экспериментов

В ходе работы разработан прототип подсистемы информационного поиска, в котором были реализованы пользовательские профили, классификационная модель текстовых документов, в основе которой лежат онтология предметной области и профили пользователей.

В процессе работы были проведены эксперименты с разработанной подсистемой. Для оценки качества информационного поиска использовались следующие характеристики [Маннинг и др., 2011]:

Полнота (P) – доля релевантных документов в выборке, по отношению ко всем релевантным документам коллекции.

$$\text{P} = \frac{\text{кол-во релевантных найденных документов}}{\text{кол-во релевантных документов}}$$

Точность (T) – доля релевантных документов выборки, по отношению ко всем документам выборки.

$$T = \frac{кол-во\ релевантных\ найденных\ документов}{кол-во\ найденных\ документов}$$

Для удовлетворения баланса между двумя этими параметрами использовался параметр F-мера (F measure) [Маннинг и др., 2011]. Данный параметр представляет собой среднее гармоническое взвешенное:

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

где $\beta^2 = \frac{1-\alpha}{\alpha}, \alpha \in [0,1]$, т.е. $\beta^2 \in [0,\infty]$. По умолчанию сбалансированная F-мера присваивает точности и полноте одинаковые веса, т.е. $\alpha = 1/2$, или $\beta = 1$. Если $\beta < 1$ предпочтение отдают точности поиска, при $\beta > 1$ полноте поиска. При $\beta = 1$ формула принимает вид:

$$F_\beta = \frac{2PR}{P+R}$$

В данной работе использовалась сбалансированная F-мера.

Результаты работы разработанной подсистемы информационного поиска сравнивались со следующими поисковыми системами:

1) Яндекс. Персональный поиск (ЯПП).

2) Архивариус 3000 (A300).

3) AOL Desktop Search (AOL).

4) Copernic Desktop Search (CDS).

Для анализа результатов работы подсистемы информационного поиска использовалась выборка, состоящая из 100 текстовых документов, 20 из которых являются техническими заданиями, а 80 – статьями из области построения информационных систем. Результаты экспериментов представлены в таблице 1.

Таблица 1 – Результаты экспериментов с поисковыми системами (для первых 10 документов).

Оценка	Новая модель	ЯПП	A3000	AOL	CDS
Запрос «Microsoft Solution Framework»					
T	**0.4**	**0.4**	**0.4**	**0.4**	**0.4**
P	**1**	**1**	**1**	**1**	**1**
F-мера	**0.6**	**0.6**	**0.6**	**0.6**	**0.6**
Запрос «Rational Unified Process»					
T	**0.4**	**0.4**	**0.4**	**0.4**	**0.4**
P	**1**	**1**	**1**	**1**	**1**
F-мера	**0.6**	**0.6**	**0.6**	**0.6**	**0.6**
Запрос «хранимые процедуры»					
T	0.5	0.4	0.4	**0.6**	**0.6**
P	0.5	0.3	0.3	**0.4**	**0.4**
F-мера	0.5	0.3	0.3	**0.5**	**0.5**
Запрос «вариантов использования»					
T	**0.5**	0.3	0.2	0.2	0.2
P	**0.7**	0.4	0.3	0.3	0.3
F-мера	**0.6**	0.3	0.2	0.2	0.2

Результаты эксперимента проранжированы в порядке семантической неопределенности запроса. Семантическое значение запросов, состоящие из терминов «Microsoft Solution Framework» и «Rational Unified Process», явно указывают на конкретную предметную область, и как видно из результатов эксперимента все поисковые системы хорошо справились с поставленной задачей. Запрос, состоящий из терминов «вариантов использования», несет более слабое семантическое значение, для того чтобы понять, что он указывает на предметную область «Унифицированный язык моделирования (UML)» и, как видно из результата эксперимента, наиболее качественный результат показывает разработанная поисковая модель. Следовательно, можно сделать вывод, что поисковые и классификационные модели, обладающие информацией о пользовательских потребностях показывают более качественный результат в условиях неопределенности, чем традиционные модели.

Заключение

В данной работе рассматривается специфика информационной потребности специалистов занятых в процессе проектирования технических средств. Предлагается использовать информационную потребность пользователей в процессе информационного поиска в условиях неопределенности. Рассматривается новая модель информационного поиска, основанная на использовании пользовательских профилей, которые содержат информационную потребность и предметно-ориентированной онтологии. Приведенная в статье модель легла в основу разработанной программной системы информационного поиска проектных документов, с которой проведены вычислительные эксперименты.

Библиографический список

[Добров и др., 2006] Добров Б.В., Лукашевич Н.В., Лингвистическая онтология по естественным наукам и технологиям: основные принципы разработки и текущее состояние // Десятая национальная конференция по искусственному интеллекту с международным участием (Обнинск, 25-28 сентября 2006 г.) – М.: Физматлит, 2006.

[Гаврилова и др., 2000] Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. –СПб. : Питер, 2000. – 384 с.

[Маннинг и др., 2011] Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. М: Вильямс, 2011

[Наместников 2009] Интеллектуальные проектные репозитории. – Ульяновск: УлГТУ, 2009. Наместников А.М. Интеллектуальные проектные репозитории. – Ульяновск: УлГТУ, 2009.

[Наместников и др., 2010] Наместников А.М., Филиппов А.А. Концептуальная индексация проектных документов // Автоматизация процессов управления. – 2010. – №2(20). – С. 34-39.

[Норенков, 2009] Норенков И.П. Основы автоматизированного проектирования. М: МГТУ имени Баумана, 2009.

[Титов, 2009] Титов Ю. А. САПР технологических процессов. Ульяновск, 2009.

[Филиппов др, 2013] Филиппов А.А., Наместников А.М., Субхангулов Р.А. Применение нечетких моделей в задачах кластеризации и информационного поиска текстовых проектных документов // Интегрированные модели и мягкие вычисления в искусственном интеллекте. Сборник научных трудов VII-й Международной научно-практической конференции (Коломна, 22-22 мая 2013 г.) В 3-х томах. Т.3. – М.:Физматлит, 2013. С. 1278-1289.

ONTOLOGICALLY-ORIENTED MODEL OF CLASSIFICATIONS OF TEXT DOCUMENTS

Namestnikov A.M. *, Subkhangulov R.A. *

** Ulyanovsk State Technical University, Ulyanovsk, Russia*

nam@ulstu.ru

subkhangulov-ruslan@yandex.ru

In article the description of specifics of information needs of experts occupied in design process the technical means. We considered the user profiles, their application in the course of full-text query search in text documents. We considered to use of the bayesian qualifier in terms of ontological models of information search. The results of the experiments proving efficiency of developed program system are given.

Introduction

In CAD archive the designer has no possibility to solve semi-structured problems of search. We suggest to use the user profiles and to apply them in the course of information search. Such profiles use ontology of data domain and contain information on the user preferences.

Main Part

In article it is considered features information need of a designer. During operation we developed ontology of data domain. The domain ontology includes two levels: conceptual and terminological. Created a row of the user profiles for designers, using ontology of data domain. Developed the information retrieval system. This system uses the user profiles and classification model on a basis Bayes's rule in the course of information search. Were carried out a row of experiments.

Conclusion

Experiments showed that in the conditions of uncertainty our system shows good results. Results of researches showed that method of Ontologically-oriented model of classifications of text documents is more effective than a tradition method.