УДК 681.3

Р.А. Субхангулов

ОНТОЛОГИЧЕСКИ-ОРИЕНТИРОВАННЫЙ МЕТОД ПОИСКА ПРОЕКТНЫХ ДОКУМЕНТОВ

Субхангулов Руслан Айратович, *аспирант кафедры «Информационные системы» Ульяновского государственного технического университета, окончил факультет информационных систем и технологий УлГТУ. Имеет статьи в области интеллектуальных систем хранения и обработки информации. [e-mail: subkhangulov-ruslan@yandex.ru].*

Аннотация

В данной статье представлен новый метод информационного поиска текстовых проектных документов на основе онтологии предметной области. В работе рассмотрено формализованное представление метода информационного поиска, позволяющего улучшить показатель точности за счет использования внешней информации о состоянии предметной области. В отличие от известных методов анализ текстовой проектной документации производится не на уровне терминов документов, а на уровне концептов (понятий), представленных в онтологии проектной организации.

Ключевые слова: проектные документы, информационный поиск, онтология, термины, концепты.

Ruslan Ayratovich Subkhangulov, *post-graduate student at the Chair 'Information Systems'; graduated from the Faculty of Information Systems and Technology of Ulyanovsk State Technical University; author of articles in the field of intellectual systems for storage and processing of information. e-mail: subkhangulov-ruslan@yandex.ru.*

Abstract

The article presents a new method for information retrieval of text design documents, based on domain ontology. The article considers a formalized presentation of the information retrieval method enabling the improvement of accuracy factor due to the use of external information on domain status. Unlike known methods, the analysis of text design documents is not used at the level of document terms but at the level of concepts (notions) presented in the ontology of design organization.

Key words: design documents, information retrieval, ontology, terms, concepts.

ВВЕДЕНИЕ

Современная крупная проектная организация обладает значительным по объему электронным архивом проектно-конструкторской документации, большая часть которой представлена в текстовом неструктурированном виде. Фактически, такой проектный репозиторий текстовой документации содержит в себе опыт и знания большого количества высококвалифицированных специалистов, которые на протяжении многих лет занимались разработкой и проектированием сложных систем. При увеличении объемов проектного репозитория затрудняется поиск документов по заранее заданным реквизитам. В результате важный опыт предыдущих разработок остается невостребованным, и, как следствие, увеличивается производительный цикл выпускаемых предприятием изделий.

Решение указанной проблемы может основываться на применении интеллектуальных методов и алгоритмов полнотекстового информационного поиска проектных документов. Для эффективного применения методов интеллектуального анализа текстовой проектной документации недостаточно рассматривать отдельный проектный документ как набор терминов из ограниченной предметной области. Принимая во внимание слабоструктурированный характер проектных документов, актуальным является разработка нечетких методов полнотекстового информационного поиска, функционирование которых происходит не на уровне терминов предметной области, а на понятийном уровне. Следовательно, возникает необходимость в разработке методов и алгоритмов предметно-ориентированного поиска документов как информационных ресурсов электронного архива технической документации на основе прикладной онтологии.

1 ПРОБЛЕМЫ ИНФОРМАЦИОННОГО ПОИСКА ПРОЕКТНЫХ ДОКУМЕНТОВ

Понятие проектных документов основывается на том, что существует проектная информация и она соотносится с проектной документацией. Проектная информация - это более частный вариант информационного объекта, согласно работе [1], понимается как «описание некоторой сущности (реального объекта, явления, процесса, события) в виде совокупности логически связанных реквизитов (информационных элементов)». В данной работе в качестве проектных документов используются текстовые документы, в которых зафиксирована проектная информация, относящаяся к ограниченной предметной области проектной организации.

Информационный поиск (Information Retrieval) - это процесс поиска в большой коллекции (хранящейся, как правило, в памяти компьютера) некоего неструктурированного материала (обычно - документа), удовлетворяющего информационные потребности [2]. В современных информационных технологиях выделяют три вида моделей информационного поиска:

1) Теоретико-множественные (булевская, нечетких множеств, расширенная булевская).

Модель булева поиска - это модель информационного поиска, в ходе которого можно обрабатывать любой запрос, имеющий вид булева выражения, т. е. выражения, в котором термины используются в сочетании с операциями AND, OR и NOT [2].

Недостатки:

• на заданный запрос поисковая машина может вернуть очень много документов (или даже все документы коллекции). В этом случае пользователь вынужден последовательно добавлять условия в запрос, чтобы уменьшить результирующую выборку. Поиск производится методом проб и ошибок. В результате также часто возникает ситуация, когда условия булева запроса оказываются противоречивы, и пользователь не получает ни одного документа;

• как правило, полезную выборку обозримого размера можно получить, задав сложную логическую формулу. При этом от пользователя требуется не только знание правил построения формул, но и достаточно хорошее знакомство с «языком» предметной области;

• вследствие того, что существуют только два значения релевантности: «релевантен» (true) и «нерелевантен» (false), результирующая выборка не может быть упорядочена по релевантности. Все документы одинаково релевантны;

• все атомы формулы имеют одинаковую важность (вес), хотя некоторые из них могут быть «ключевыми», другие - вспомогательными.

2) Алгебраические (векторная, обобщенная векторная, латентно-семантическая, нейросетевая).

Векторная модель - это модель, основанная на векторном представлении коллекции документов. Данная модель в научной литературе называется «мешком слов» (bag of words model) [2]. В рамках этой модели точный порядок следования слов не рассматривается, так как предполагается, что документы с одинаковыми «мешками слов» сходны. Основное значение придается количеству вхождений терминов в документ. Вводимые пользователем запросы также представляются в виде векторов. Для вычисления подобия векторов используется косинусная метрика. Векторная модель была реализована в 1968 году Джерардом Солтоном в поисковой системе SMART (Salton's Magical Automatic Retriever of Text) [3].

Недостатки:

• эта модель не справляется с синонимией (когда разные слова имеют одно значение) и полисемией (когда одно слово имеет разные значения);

• векторы и массивы имеют высокую размерность, что приводит к сложности обработки.

3) Вероятностная модель поиска.

В качестве оценки соответствия документа запросу в вероятностной модели, разработанной в 1977 году Robertson и Sparck-Jones, используется вероятность того, что пользователь считает документ релевантным. В данной модели поиска вероятность, что документ релевантен запросу, основывается на предположении, что термы запроса по-разному распределены среди релевантных и нерелевантных документов. При этом используются формулы расчета вероятности, базирующиеся на теореме Байеса.

Недостатки:

• низкая вычислительная масштабируемость;

• необходимость постоянного обучения системы.

2 ФОРМАЛИЗАЦИЯ МЕТОДА ИНФОРМАЦИОННОГО ПОИСКА ПРОЕКТНЫХ ДОКУМЕНТОВ

В настоящее время наблюдается рост интереса к разработке новых методов интеллектуального анализа данных, основанных на применении онтологии [4]. Онтология в нашей работе имеет следующий вид:

$$O = (C,\ T,\ F,\ F\ R) \qquad (1)$$

где C - словарь понятий, формирующий верхний уровень онтологии $C = \{c_i\}$;

T - словарь термов, который формируется автоматически на основе пользовательских документов. Словарь представляет собой множество $T = \{t., w_i\}$, где $t_{.i}$ - термин и w_i - вес термина;

F_t - функция интерпретации терминов, сопоставляет термин с концептами. Функция F_t является функцией нечеткого соответствия;

F_c - функция интерпретации концептов, сопоставляет концепты с терминами из словаря терминов T. Функция F_c является инверсией функции F, в которой область отправления F_c совпадает с областью прибытия F_t, и наоборот, область прибытия F_t совпадает с областью отправления F_c;

R_c - конечное множество отношений между концептами (понятиями) заданной предметной области. Отношения между понятиями должны удовлетворять требованиям транзитивности и наследования.

На основе онтологии можно создать матрицу отношений вектора концептов и вектора терминов:

$$M_{CT} = \begin{array}{c} \\ c_1 \\ c_2 \\ \vdots \\ c_n \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_3 & \cdots & L \end{array} \\ \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nm} \end{bmatrix} \end{array},$$

где $\{c\}$ - концепты онтологии, n - количество концептов в онтологии, $\{t_j\}$ - термы, формирующие терминологическое окружение концепта, m - количество термов, w_{ij} - величина, характеризующая степень выраженности терма j в концепте i, где $0 \le w_{ij} \le 1$.

Величина w_{ij} вычисляется по следующей формуле [2]:

$$w_v = a + Q.-a)\frac{tfj*}{tf_{max}2u_n'} \quad (^2)$$

где $tfwx№) \simeq^{max} j \wedge d\, tfj,d$ - максимальная величина частоты встречаемости термина j в документе d (который описывает концепт '), a - сглаживающий коэффициент, принимающий значение между нулем и единицей (как правило устанавливают равным 0,4). Роль данного параметра заключается в уменьшении вклада второго члена.

Нормировка частоты термина по максимуму предназначена для того, чтобы избежать следующей аномалии: в более длинных документах наблюдаются более высокие частоты терминов, так как в более «длинных» документах чаще содержатся повторяющиеся слова.

Центральное место в информационном поиске занимает понятие запроса. Запрос - это формализованный способ выражения информационных потребностей пользователем системы. Запрос будем представлять в виде множества $Q = \{t_1, w\}\}$, где t_i - термины из словаря предметной области T, w_i - величина, характеризующая степень выраженности терма в концепте, где $0 \le w_i \le 1$, вычисленная по формуле (2).

Индексы запроса представляются в виде одномерного вектора M_{QC}:

$$M_{QC} = \begin{matrix} C\backslash & C_2 & \dots & C_n \\ Hi & Ш & \dots & \mu_n \end{matrix}$$

где c_i - концепты онтологии, n - количество концептов в онтологии, д. - величина, характеризующая степень влияния концепта в запросе, где $0 \le д. \le 1$, вычисляемая по формуле (3).

Индексы документа представляются в виде матрицы M_{DC}:

$$M_{DC} = \begin{matrix} & \text{<1} & c_2 & \dots & c_n \\ 4 & Hи & Ml2 & \dots & M\text{-}1, \\ d_2 & м™ & И22 & \dots & Д_{2л} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ dk & .Htl & \mu_{k2} & \dots & Ufa. \end{matrix},$$

где d_i' - документ в хранилище репозитория $D = \{d_1, d_2, \dots, d_k\}$, k - количество документов в хранилище репозитория, c_i - концепт онтологии, n - количество концептов в онтологии, $д_{ij}$ - величина, характеризующая степень влияния концепта в документе, где $0 \le w_{ij} \le 1$, вычисляемая по формуле [5]:

$$\mu_{ij} = 1 - \frac{l}{l_U} \sum_{s=\backslash}^{l_k} \left| f_s^k - f_s^j \right|, \quad (3)$$

где f - частоты встречаемости термина s в j-м документе и в описании k-го понятия онтологии соответственно; l_k- мощность текстового входа понятия c_k. В том случае, если термин s отсутствует в j-м документе, тогда fS принимается равным нулю.

В основе метода онтологически-ориентированного поиска лежит нахождение меры включения вектора концептов запроса в вектор концептов документа. Для решения данной задачи предлагается воспользоваться формулой нечетких включений множеств [6]:

$$\gamma\left(\bar{I}_q, \bar{I}_d\right) = \quad - (^C))' \quad (4)$$

где $\bar{I}_q, \bar{I}_d$ - индексы запроса и документа, c - концепты онтологии, где $0 \le Y\,(\bar{j}_q, \bar{Id}) \le 1$.

Представляем четыре варианта алгоритмов поиска информации:

1) «Простой» алгоритм поиска:

Данный алгоритм, блок-схема которого представлена на рисунке 1, является базовым для других алгоритмов.

Этапы выполнения алгоритма:

1. Формирование индекса запроса M_{QC} путем вычисления $д_{ij}$ для каждого концепта по формуле (3);

2. Вычисление меры включения запроса Q в документ D по формуле (4);

3. Ранжирование документов по убыванию значения $\gamma\left(\bar{I}_q, \bar{I}_d\right)$;

4. Вывод результатов поиска.

В ходе проведенных экспериментов были получены результаты, представленные в таблице 1.

2) Иерархический онтологически-ориентированный алгоритм поиска:

Рис. 1. Блок-схема первого алгоритма

Таблица 1

Результаты онтологически-ориентированной модели поиска и векторной модели

Онтологически-ориентированная модель поиска	Векторная модель поиска
Системные объекты базы данных	Технология создания распределенных информационных систем
Непроцедурный доступ к данным	Системные объекты базы данных
Процедурное расширение языка SQL	tz_20
Язык манипуляции данными	Этапы разработки проекта_ заключительные стадии проектирования, схема базы данных
Этапы разработки проекта_ заключительные стадии проектирования, схема базы данных	tz_05
Этапы разработки проекта_ реализация, тестирование, эксплуатация и сопровождение	…
Методы композиции и декомпозиции исполняемых UML-моделей	Непроцедурный доступ к данным
Условия целостности базы данных	Процедурное расширение языка SQL

$$\gamma\left(\bar{I}_q,\bar{I}_d\right)=\&\left(F\left(\mu_{I_q}(c)\right)\rightarrow\mu_{I_d}(c)\right), \quad (5)$$

где $\bar{I}_q$, $\bar{I}_d$ - индексы запроса и документа,

c - концепты онтологии,

F (jI^ (c)^—> (c) - функция перехода от общего к частному, выполняемая следующим образом:

Рис. 2. Блок-схема иерархического онтологически-ориентированного алгоритма поиска

• система определяет, имеет ли концепт потомков и ненулевой вес;

• если концепт удовлетворяет обоим условиям на первом шаге, система присваивает нулевой вес концепту, для того чтобы убрать влияние концепта на результат поиска;

• действие предыдущего шага выполняется в цикле, двигаясь по иерархии концептов онтологии вниз до тех пор, пока не дойдет до значимых концептов, которые не имеют потомков.

Алгоритм представлен в виде блок-схемы на рисунке 2.

В ходе проведенных экспериментов были получены результаты, представленные в таблице 2.

3) Использование терминологического окружения концепта. Вводимый пользователем запрос $Q = \{t_i, w_i\}$ предварительно обрабатывается системой, для уточнения запроса. Алгоритм представлен в виде блок-схемы на рисунке 3.

Этапы выполнения алгоритма:

1. С помощью функции F_t - функции интерпретации терминов, определяются концепты, вокруг которых вводимые пользователем термины формируют терминологическое окружение;

2. С помощью функции F_c - функции интерпретации концептов, определяются уникальные термины, формирующие терминологическое окружение концептов, отобранных на первом этапе;

3. Пользователь выбирает термины из списка, полученного на втором этапе, тем самым уточняя запрос;

4. Формирование вектора запроса M_{QC} путем вычисления p_{ij} для каждого концепта по формуле (3);

5. Выполнение действия по первому алгоритму или по второму в зависимости от выбора пользователя.

4) Алгоритм основан на использовании уточнения запросов посредством отбора концептов. В данном алгоритме уточняются концепты, по которым следуют искать документы. Вводимый пользователем запрос $Q = \{t_i, w_i\}$

Таблица 2

Результаты иерархической онтологически-ориентированной модель поиска и векторной модели

Иерархическая онтологически-ориентированная модель поиска	Векторная модель поиска
Непроцедурный доступ к данным	tz_20
Процедурное расширение языка SQL	Непроцедурный доступ к данным
Этапы разработки проекта_ заключительные стадии проектирования, схема базы данных	Процедурное расширение языка SQL
Системные объекты базы данных	tz_10
Встроенные операторы SQL	Встроенные операторы SQL
Язык манипуляции данными	Технология создания распределенных информационных систем
Технология создания распределенных информационных систем	tz_19
Этапы разработки проекта реализация, тестирование, эксплуатация и сопровождение	Модели представления данных в СУБД

предварительно обрабатывается системой для его уточнения (рис. 4).

Этапы выполнения алгоритма:

1. Формирование вектора запроса Mq_C путем вычисления μ_{ij} для каждого концепта по формуле (3);
2. Пользователь отбирает те концепты, по которым следует проводить поиск документов;
3. Вычисление меры включения запроса Q в документ D по формуле (4);
4. Ранжирование документов по убыванию значения $\gamma\left(\bar{I}_q, \bar{I}_d\right)$;
5. Вывод результатов поиска.

Рис. 3. Блок-схема использования терминологического окружения концепта

Рис. 4. Блок-схема поиска, основанного на уточнении запроса посредством отбора концептов

ЗАКЛЮЧЕНИЕ

В данной работе представлен новый метод информационного поиска на основе онтологии предметной области. Учет понятийной структуры предметной области проектной организации позволяет повысить качество ранжированного информационного поиска проектных документов по сравнению с традиционными методами. В разработанном методе применение стратегии онтологического уточнения запросов способствует получению удовлетворительных результатов даже в тех случаях, когда пользователь, не имея должной квалификации и знаний предметной области, формулирует запрос в виде неполного набора ключевых слов. Приведенные в статье методы и алгоритмы легли в основу разработанной программной системы онтологически-ориентированного информационного поиска проектных документов, с которой проведены вычислительные эксперименты.

СПИСОК ЛИТЕРАТУРЫ

1. Наместников А.М. Интеллектуальные проектные репозитории. - Ульяновск : УлГТУ, 2009.

2. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. - М. : Вильямс, 2011.

3. Сэлтон Г. Автоматическая обработка, хранение и поиск информации. - М. : Советское радио, 1973.

4. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. -СПб. : Питер, 2000. - 384 с.

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

6. Берштейн Л.С., Боженюк А.В. Нечеткие графы и гиперграфы. - М. : Научный мир, 2005.