OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 681.3

ОНТОЛОГИЧЕСКИЙ ПОДХОД К ФОРМИРОВАНИЮ КОНТЕКСТНЫХ ЗАПРОСОВ В ЭЛЕКТРОННОМ АРХИВЕ ТЕХНИЧЕСКИХ ДОКУМЕНТОВ

Наместников А.М.

Ульяновский государственный технический университет,
г. Ульяновск, Российская Федерация
nam@ulstu.ru

В работе приводятся модели и методы выполнения контекстно-ориентированных запросов к электронным архивам технических документов крупной проектной организации. Онтология рассматривается в контексте решения задачи уточнения пользовательских запросов с целью повышения точности и полноты информационного поиска. В статье содержатся результаты вычислительных экспериментов на базе современного предметно-ориентированного электронного архива.
Ключевые слова: онтология; контекстно-ориентированный запрос; электронный архив.

Введение

Основными задачами электронного архива крупной проектной организации является обеспечение коллективной работы проектно-конструкторских отделов над проектом, добавление, хранение и поиск технических документов (ТД). Поиск часто осуществляется по заранее определенным реквизитам документов и по ключевым словам [Маннинг и др., 2011]. Однако данные модели поиска не имеют представления об информационных потребностях пользователя и, тем самым, всегда присутствует вероятность того, что документы, которые были отобраны, не позволят сократить информационную неопределенность проектировщика. Современные системы информационной поддержки используют механизмы интеллектуального поиска. Интеллектуальный поиск – это ключевая тенденция в современном информационном поиске, которая предполагает способность поисковой системы к самоорганизации, осуществление независимого общения с пользователем, эффективный поиск текстовых документов, реагирующий на изменения информационной потребность пользователя. Знания могут быть представлены в виде онтологии предметной области [Добров Б.В. и др., 2006; Гаврилова Т.А. и др., 2000].

В данной статье представлена модель формирования контекстно-ориентированных поисковых запросов, основанная на использования знаний о предпочтениях проектировщиков в процессе поиска ТД. Фактически, речь идет об использовании индивидуального профиля проектировщика, который может применяться в задачах онтологически-ориентированного информационного поиска текстовых документов и позволит максимально полно удовлетворить информационную потребность пользователя.

1. Модель профиля проектировщика

Формализация *профиля проектировщика* осуществляется на основе предположения о том, что имеется возможность фиксировать результаты проектных запросов к электронному архиву в виде множества документов, удовлетворяющих информационной потребности, и множества документов, которые текущей информационной потребности не удовлетворяют.

Каждой информационной потребности In_j^i будем ставить в соответствие пару классов понятий онтологии предметной области $C^+ = \{c_1^+, c_2^+, ..., c_n^+\}$, $C^- = \{c_1^-, c_2^-, ..., c_m^-\}$, определяющие положительные и отрицательные подмножества понятий онтологии, соответственно (рисунок 1).

В процессе выполнения конкретным проектировщиком информационных запросов к электронному архиву определяется набор ТД, которые соответствуют его информационной потребности (D^+) и ТД, не соответствующие ей (D^-), с учетом текущей стадии (этапа) проектирования.

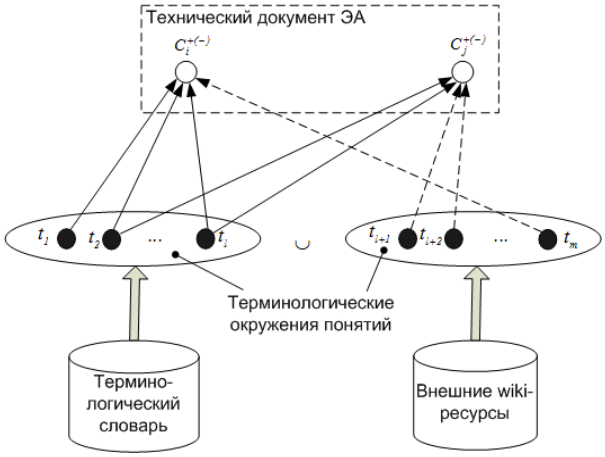


Рисунок 1 – Источники формирования ресурсов профиля проектировщика

Для каждого документа определяется его концептуальное представление. Другими словами, производится нечеткое онтологическое индексирование. Запишем нечеткое соответствие между множеством $C^{+(-)}$ и множеством $T^D = T^D_{in} \cup T^D_{ext}$, как $\tilde{\Gamma}_{CT} = (C^{+(-)}, T^D, \tilde{F}_{CT})$, где $\tilde{F}_{CT}$ – нечеткое множество в $C^{+(-)} \times T^D$. Определим нечеткое соответствие $\tilde{\Gamma}_{CT}$ в виде ориентированного двудольного графа с множеством вершин $C^{+(-)} \cup T^D$, каждой дуге $< c_i^+, t_j > (< c_i^-, t_j >)$ которого приписываем значение функции принадлежности $\mu_{F_{CT}} < c_i^+, t_j > (\mu_{F_{CT}} < c_i^-, t_j >)$. Указанное значение функции принадлежности вычисляется на основе нормализованной частоты встречаемости термина в терминологическом окружении понятия.

Образ множества T^D, при соответствии $\tilde{\Gamma}_{CT}$, фактически представляет собой нечеткое множество, элементами которого являются концепты с соответствующими степенями выраженности:

$$\tilde{\Gamma}_{CT}(T^D) = \{\mu_{\Gamma_{CT}}(c^{+(-)}) / c^{+(-)}\}, \quad (1)$$

где $\mu_{\Gamma_{CT}}(c^{+(-)}) = \underset{t \in T^D}{\vee} \mu_{F_{CT}} < c_i^{+(-)}, t_j >$.

В положительные и отрицательные подмножества понятий онтологии предметной области включаются такие понятия из онтологических представлений, степень выраженности которых наибольшая.

Формально профиль проектировщика будем представлять в виде кортежа [Наместников и др., 2014]:

$$Ex^i_j = \left\langle In^i_j, C^+, C^- \right\rangle, c^{+(-)} = \mu_{\Gamma_{CT}}(c^{+(-)}) =$$
$$= \max_{c \in D}(\underset{t \in T^D}{\vee} \mu_{F_{CT}} < c_i^{+(-)}, t_j >), c^{+(-)} \in C^{+(-)},$$

где i – индекс проектировщика, j – индекс стадии жизненного цикла проектирования автоматизированной системы (АС).

2. Формирование онтологического контекста

Следовательно, контекст проекта перепишем как граф вида [Наместников и др., 2012]:

$$G^{PT} = \left\langle C^{PT}, R^{PT} \right\rangle, \quad (2)$$

где C^{PT} – множество вершин-понятий проекта, R^{PT} – множество дуг, соединяющих вершины-понятия.

Множество понятий проекта определяется как результат функции онтологического индексирования технического задания (Tz) на реализуемый проект ($F_{oI}(Tz)$) и функции онтологического доопределения множества C^{Tz}, как результата $F_{oI}(Tz)$ с применением wiki-ресурсов сети Internet ($F_{cAdd}(C^{Tz})$). Алгоритм формирования онтологического контекста проекта АС (2) представим в виде следующих шагов [Наместников и др., 2012]:

Шаг 1. Загрузка файла технического задания (Tz).

Шаг 2. Онтологическое индексирование технического задания:

$$C^{Tz} = F_{oI}(Tz). \quad (3)$$

Шаг 3. Доопределение множества C^{Tz}.

На данном шаге выполняется анализ wiki-ресурса сети Internet и определяется множество дополнительных понятий, имеющих связи с понятиями множества C^{Tz}. Идентификация связей между понятиями определяется на основе существующих гиперссылок на соответствующие страницы сети, содержащие текстовые описания понятий.

Шаг 4. Загрузка терминологического словаря Dic, который формируется на основе технической документации электронного архива, в том числе из набора основных терминов и понятий из ГОСТ, принятых к исполнению в проектной организации.

Шаг 5. Сравнение терминологических окружений $T_{sur}(\hat{C}^{Tz})$ понятий $\hat{C}^{Tz} = F_{cAdd}(C^{Tz})$ с терминами из Dic. Если $\forall \hat{w} \in T_{sur}(\hat{C}^{Tz})$ выполняется условие $\hat{w} \notin Dic$, то необходимо удалить понятие $\hat{c} \in \hat{C}^{Tz}$.

Шаг 6. Проверка очередного $\hat{c} \in \hat{C}^{Tz}$. Если сравнение терминологических окружений со словарем выполнено не для всех элементов множества $\hat{C}^{Tz}$, тогда выполняется переход к шагу 5.

Шаг 7. Определение множества дуг R^{PT} на основе анализа гиперссылок страниц wiki-ресурса.

Шаг 8. Сохранение графа G^{PT}.

Рассмотрим детально процесс формирования концептуальной сети, извлекаемой из «Википедии» – свободной общедоступной мультиязычной универсальной интернет-энциклопедии, реализованной на принципах wiki. Концепты в данной библиотеке представлены в виде HTML-страниц. Для связи между страницами используются гиперссылки, которые символизируют семантическую связь между понятиями. Опираясь на систему гиперссылок, существует возможность в автоматическом режиме переходить от одной страницы к другой, извлекая знания о понятиях предметной области.

Рассмотрим определение множества дуг R^{PT} (шаг 7) на основе анализа гиперссылок wiki-ресурсов:

1. Извлекаются понятия из онтологии проекта.
2. Выполняется извлечение понятий из wiki. В основе данного процесса лежит модифицированный алгоритм волновой трассировки (применяемый при трассировке печатных плат радиоэлектронных устройств). Данный процесс состоит из ряда последовательных шагов:

2.1. На вход данного алгоритма поступают множества понятий полученных на этапе 1.

2.2. Выполняется поиск страниц в wiki, в которых концепты являются заголовками.

2.3. Страницы, полученные на этапе (2.2) анализируются с целью нахождение тех концептов, для которых одновременно выполняются условия:

- существует страница, которая описывает концепт;
- анализируемая страница содержит гиперссылки на страницу найденного концепта;
- страница концепта содержит обратную гиперссылку на анализируемую страницу.

2.4. Обнаруженные концепты добавляются в предварительное ядро онтологии.

2.5. Проверяется условие существования маршрута между всеми первичными концептами, которые получены на этапе (1).

2.6. Если условие пункта 2.5 выполняется, то это означает окончание модифицированного алгоритма волновой трассировки, если не выполняется, то пункты 2.2-2.5 выполняются снова для концептов, извлеченных на этапе 2.3.

Таким образом, на выходе второго этапа получаем множество понятий, между которыми существуют неидентифицированные семантические отношения. Однако может оказаться так, что это множество содержит понятия, которые выходят за границы исследуемой предметной области.

3. Полученные на втором этапе концепты приводятся в нормальную форму, т.е. с помощью алгоритма стемминга выделяются словарные основы концептов.

4. Для каждого извлеченного концепта формируется терминологическое окружение. Терминологическое окружение создается путем извлечения терминов из страницы концепта wiki-ресурса и вычисления частоты встречаемости термина на данной странице по следующей формуле:

$$ntf_{t,d} = a + (1-a)\frac{tf_{t,d}}{tf_{\max}(d)}, \qquad (4)$$

где $tf_{\max}(d) = \max_{\tau \in d} tf_{\tau,d}$ – максимальная величина tf в документе d, a – сглаживающий коэффициент, принимающий значение между нулем и единицей (экспериментально устанавливают равной 0,4). Роль данного параметра состоит в уменьшении вклада второго члена. Нормировка частоты термина по максимуму предназначена для того, чтобы избежать следующей аномалии [Маннинг и др., 2011]: в более длинных документах наблюдаются более высокие частоты терминов, так как в более длинных документах чаще содержатся повторяющиеся слова.

3. Классификация запросов на основе байесовской модели

Рассмотрим алгоритм формирования проектных запросов для улучшения показателей точности и полноты запросов к электронным архивам технических документов проектной организации.

Пусть множество $\hat{w} = \{\hat{w}_1, \hat{w}_2, ..., \hat{w}_n\}$ – есть множество ключевых слов проектного запроса к электронному архиву. Используя функцию $F_{DC} : \{T^D\} \to C$ для построения терминологических окружений понятий онтологии предметной области, данное множество отображается в нечеткое множество:

$$\tilde{I}_q = \{\mu_1 / c_1, \mu_2 / c_2, ..., \mu_m / c_m\}, \qquad (5)$$

как результат онтологического преобразования терминов в набор степеней выраженности понятий онтологии.

Для представления модели формирования проектных запросов на концептуальном уровне будем использовать наивный байесовский классификатор. Вероятность того, что понятие запроса c принадлежит классу $k \in \{C^+, C^-\}$, будем определять по формуле Байеса:

$$P(k \mid c) = \frac{P(c \mid k) \cdot P(k)}{P(c)}, \qquad (6)$$

где $P(c \mid k)$ – вероятность встретить понятие c среди всех понятий класса k; $P(k)$ – безусловная

вероятность понятия класса k в онтологии предметной области; $P(c)$ – безусловная вероятность понятия c в онтологии предметной области.

Наиболее вероятный класс для понятия запроса определяется, используя оценку апостериорного максимума:

$$k_{map} = \arg\max_{k \in K} \frac{P(c \mid k) \cdot P(k)}{P(c)}. \quad (7)$$

Поскольку $P(c) = const$ в рамках одной онтологии и учитывая, что

$$P(c \mid k) \approx P(w_1 \mid k) \cdot P(w_2 \mid k) \cdot ... \cdot P(w_n \mid k) = \prod_{i=1}^{n} P(w_i \mid k), \quad (8)$$

получаем:

$$k_{map} = \arg\max_{k \in K} [P(k) \cdot \prod_{i=1}^{n} P(w_i \mid k)]. \quad (9)$$

Для больших документов количество множителей $P(w_i \mid k)$ в выражении (9) может быть большим, а, следовательно, возникает проблема исчезновения порядка вследствие перемножения большого количества малых чисел. Перепишем выражение (9) с учетом свойств логарифма:

$$k_{map} = \arg\max_{k \in K} [\log P(k) \cdot \sum_{i=1}^{n} \log P(w_i \mid k)] \quad (10)$$

Оценка вероятностей $P(k)$ и $P(w_i \mid k)$ выполняется на основе обучающей выборки, сформированной для каждой информационной потребности In_j^i. Вероятность класса будем записывать как:

$$P(k) = \frac{D_k}{D}, \quad (11)$$

где D_k – количество документов, принадлежащих классу k и определяемое на основе результатов выполнения запросов проектировщика к электронному архиву; D – общее количество документов в обучающей выборке.

Величина $P(w_i \mid k)$ определяет вероятность встретить термин w_i среди терминов документов, принадлежащих классу k. Значение данной величины будем определять с учетом того, что термин из окружения понятия, включенного в проектный запрос, может отсутствовать в документах анализируемого класса. Применяя метод аддитивного сглаживания (сглаживания Лапласа), получаем:

$$P(w_i \mid k) = \frac{f_{ik} + 1}{\sum_{i' \in V} (f_{ik} + 1)}, \quad (12)$$

где f_{ik} – частота встречаемости i - го термина в документах класса k; V – терминологический словарь проектной организации (список всех уникальных терминов).

В результате классификации понятий запроса (5) понятия, принадлежащие к положительному классу понятий информационной потребности проектировщика, остаются в составе запроса. Понятия, которые классифицированы как отрицательные, исключаются из исходного запроса.

4. Способ редукции понятий контекстно-ориентированного запроса

В случае достаточно большой онтологии и неявной принадлежности запроса конкретному фрагменту предметной области проектирования мощность множества C^q (количество понятий, включенных в нечеткое представление проектного запроса) может быть большой. Поэтому возникает необходимость в процедуре редукции запроса (5).

Способ редуцирования понятий проектного запроса основывается на разделении исходного графа $G^q = \langle C^q, R^q \rangle$ на несколько подграфов, каждый из которых содержит только вершины, соединенные дугами одной семантической категории. В данной работе используются две семантические категории: «обобщение» («isA») и «часть-целое» («part_of»). На рисунке 2 представлен иллюстративный пример графа проектного запроса, в котором присутствуют вершины $c_1, ..., c_{14}$ – понятия онтологии предметной области и два типа дуг: дуги «isA» (штриховая стрелка) и дуги «part_of» (сплошная стрелка).

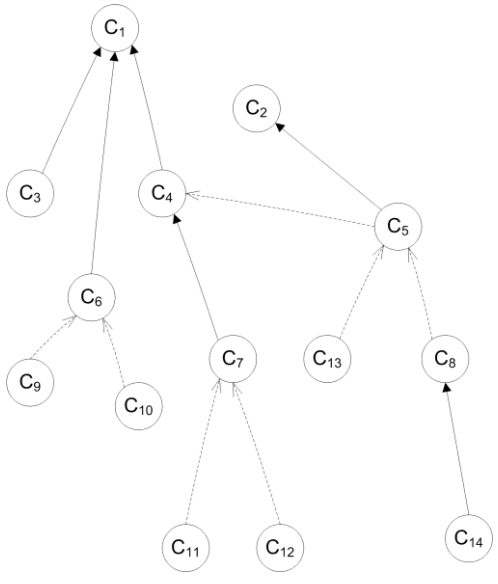


Рисунок 2 – Иллюстративное представление проектного запроса в виде графа

Редуцирование понятий графа проектного запроса предполагает выполнения ряда следующих шагов.

Шаг 1. Разбиение графа проектного запроса на несколько подграфов с учетом семантических категорий («isA» или «part_of»). На рисунке 3 показан подграф проектного запроса $G_{isA}^{q} = \left\langle C_{isA}^{q}, R_{isA}^{q} \right\rangle$, понятия которого связаны между собой отношениями типа «isA». Соответственно, на рисунке 4 представлен подграф $G_{p}^{q} = \left\langle C_{p}^{q}, R_{p}^{q} \right\rangle$ с отношениями «part_of».

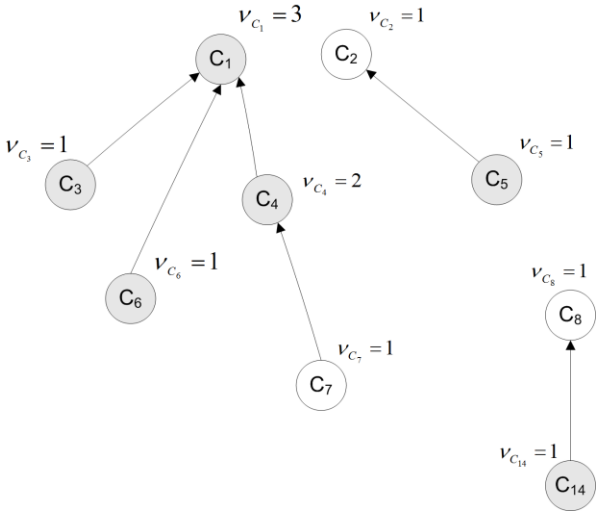
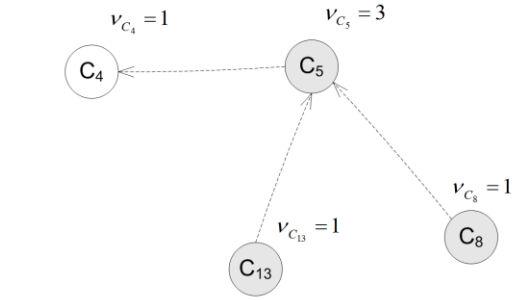


Рисунок 3 – Подграф графа проектного запроса с типом отношений «isA»



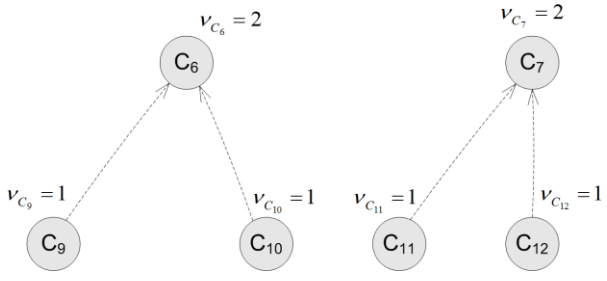


Рисунок 4 – Подграф графа проектного запроса с типом отношений «part_of»

Шаг 2. Для каждой вершины подграфов определяется ее степень v_{C_j} (количество входящих и исходящих дуг). Как видно из рисунков наибольшие степени имеют понятие C_1 в подграфе «isA» и понятия C_5, C_6 и C_7 в подграфе «part_of».

Шаг 3. В редуцированное множество понятий отдельно взятого подграфа проектного запроса включаются понятия согласно следующим правилам.

Правило 1: Если в подграфе существует вершина с максимальной степенью, то в результирующее множество соответствующего подграфа включается данная вершина и связанные с ней вершины, дуги от которых направлены к вершине с максимальной степенью.

Правило 2: Если подграф содержит две вершины, соединенные дугой, то в результирующее множество включается вершина с исходящей дугой.

Правило 3: Если подграф состоит из одной изолированной вершины, то данная вершина включается в результирующее множество понятий.

На рисунках 3 и 4 темным фоном отмечены те вершины подграфов, которые включены в результирующие множества понятий C_{isA}^{q+} и C_{p}^{q+} с использованием вышеприведенных правил.

Шаг 4. В редуцированное множество понятий проектного запроса включаются понятия, которые включены как во множество C_{isA}^{q+}, так и во множество C_{p}^{q+}: $C^{q+} = C_{isA}^{q+} \cap C_{p}^{q+}$.

5. Результаты вычислительных экспериментов

Вычислительные эксперименты проводились в центре обработки данных под управлением операционной системы Red Hat Enterprise Linux с виртуализацией Red Hat Enterprise Virtualization со следующими характеристиками: 8-ми ядерный процессор Intel Xeon MP E7-2830 Westmere-EX, размер оперативной памяти − 256 Гб.

В ходе эксперимента была построена прикладная онтология, которая включает в себя около 300 концептов и 30000 уникальных терминов.

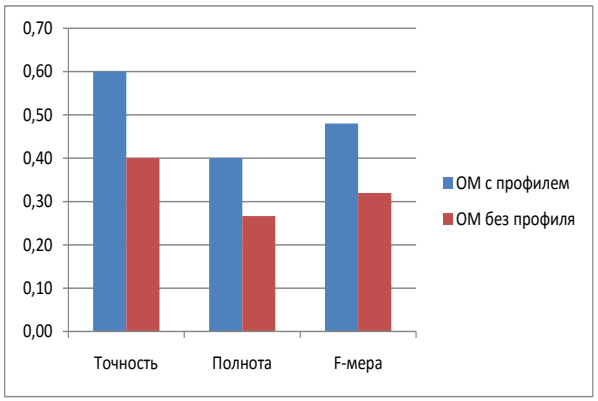


Рисунок 5 -- Сравнение результатов экспериментов

Разработанная интеллектуальная система

контекстно-ориентированного поиска документов позволила увеличить точность поиска технических документов в электронном архиве до 30%..

Заключение

В данной статье рассмотрен новый подход к анализу контекстно-ориентированных проектных запросов к электронному архиву технических документов проектной организации.

Результаты вычислительных экспериментов на реальном множестве документов из реализованных проектов по созданию автоматизированных систем демонстрируют, что формализация профилей проектировщиков и описание предметной области в виде онтологии позволяет улучшить качество человеко-машинного взаимодействия с предметно-ориентированными электронными архивами.

Исследование выполнено в рамках государственного задания №2014/232 на выполнение государственных работ в сфере научной деятельности Минобрнауки России по проекту «Разработка нового подхода к интеллектуальному анализу слабоструктурированных информационных ресурсов».

Библиографический список

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

[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.

[Наместников и др., 2014] Наместников А.М., Субхангулов Р.А. Формирование информационных запросов к электронному архиву на основе концептуального индекса // Радиотехника №7 – 2014. – С. 126-129.

[Наместников и др., 2012] Наместников А.М., Субхангулов Р.А. Разработка инструмента инженерии онтологии в интеллектуальном проектном репозитории // Автоматизация процессов управления №2 (28) – 2012. – С.38 – 43.

[Наместников и др., 2013] Наместников А.М., Субхангулов Р.А., Филиппов А.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, 20-22 мая 2013 г.). В 3-х томах. Т3. – М.: Физматлит, 2013. – С. 1278–1289.

ONTOLOGICAL APPROACH TO FORMATION OF CONTEXTUAL QUERIES IN ELECTRONIC ARCHIVE OF TECHNICAL DOCUMENTATION

Namestnikov A.M.

Ulyanovsk State Technical University, Ulyanovsk, Russian Federation

nam@ulstu.ru

In work models and methods of performance of the contextual-oriented queries are given to electronic archives of technical documentation in a large design organization. The ontology is considered in the context of the solution of a problem of specification for user queries for the purpose of increase of accuracy and completeness of information retrieval. The article contains results of computing experiments on the basis of modern subject-oriented electronic archive.

Introduction

The main objectives of electronic archive of a large design organization is ensuring collective work of design departments on the project, addition, storage and search of technical documentation. The model of formation of the contextual focused search queries based on uses of knowledge of designer preferences in the course of search of the document is presented in this article. Actually, it is about use of an individual cross-section of the designer which can be applied in problems of the ontology focused search of text documents and will allow to satisfy of user information need the most fully.

Main Part

Formalization of a designer cross-section is carried out on the basis of the assumption that there is an opportunity to fix results of design queries to electronic archive in the form of a set of documents, the satisfying information requirement, and a set of documents which don't satisfy the current information requirement.

For each document its conceptual representation is defined. In other words, indistinct ontological indexing is made.

For representation of design queries formation model at the conceptual level the naive Bayesian qualifier is used.

In case of a big ontology and implicit accessory of inquiry to a concrete fragment of domain area the set power (amount of the concepts included in indistinct representation of design query) can be big. Therefore there is a need for procedure of a reduction of query.

The mode of reduction of concepts of design inquiry is based on division of the initial count into some subgraphs, each of which contains only the tops connected by arches of one semantic category. In this work two semantic categories are used: "generalization" and "part - whole".

Conclusion

Results of computing experiments on a real set of documents from the realized projects on the automated systems creation show that formalization of designer cross-sections and the description of subject domain in the form of ontology allows to improve quality of human-machine interaction with subject-oriented electronic archives.