OSTIS-2015
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

ОНТОЛОГИЧЕСКИЙ ПОДХОД К ФОРМИРОВАНИЮ ПРОЕКТНЫХ ЗАПРОСОВ ИНТЕЛЛЕКТУАЛЬНОГО АГЕНТА

Наместников А.М., Субхангулов Р.А.

Ульяновский государственный технический университет,
г. Ульяновск, Россия

nam@ulstu.ru,

subkhangulov-ruslan@yandex.ru

В статье рассматривается применение интеллектуальных агентов в процессе формирования информационных запросов к электронному архиву технических документов. Интеллектуальный агент содержит онтологическое описание профиля проектировщика, которое выражается в виде фрагмента онтологии предметной области. Задачами интеллектуального агента являются: пополнение информации о пользовательских потребностях и нахождение текстовых документов по запросу пользователя электронного архива.
Ключевые слова: интеллектуальная система; онтология; интеллектуальный агент; информационный поиск.

Введение

Основными задачами электронного архива 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. 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. В основе интеллектуального поиска есть возможность использовать интеллектуальных агентов, функционирование которых основано на предметно-ориентированных знаниях. Эти знания могут быть представлены в виде онтологии предметной области [Добров Б.В. и др., 2006, Гаврилова Т.А. и др., 2000]. Интеллектуальные агенты изучают историю пользовательских запросов, выполняют поиск документов, обмениваются метаинформацией между собой.

В данной статье представлена модель формирования поисковых запросов, основанная на использования интеллектуальных агентов, которые учитывает предпочтения проектировщика в процессе поиска ТД. Фактически, речь идет о формировании индивидуального профиля проектировщика, который активно взаимодействует с электронным архивом ТД проектной организации. Такой профиль может применяться в задачах онтологически-ориентированного информационного поиска текстовых документов, что позволит максимально удовлетворить информационную потребность пользователя.

1. Структура интеллектуального агента

Под интеллектуальным агентом [Рассел и др.] понимается сущность, которая через систему датчиков получает информацию о среде и воздействующая посредством исполнительных механизмов на эту среду. Под интеллектуальностью следует понимать наличие у агента модели пользовательских потребностей и механизма их удовлетворения. Таким образом, интеллектуальный агент должен обладать следующими свойствами:

- Автономность - агент должен выполнять большую часть своей работы автономно, не взаимодействуя с человеком или другими агентами.

- Коммуникабельность – агент должен уметь общаться с пользователем, получая от него задания и предоставляя результаты.
- Адаптируемость и адаптивность поведения – в ходе общения с пользователем агент должен уметь настраиваться (или, хотя бы быть настраиваемым) под привычки и методы работы конкретного пользователя.
- Восприимчивость – агент, находясь в окружающей его информационной среде, должен воспринимать некоторым образом изменения окружающей среды и реагировать на изменения.
- Проактивность - агент не только должен формально выполнять поставленную задачу поиска, но и должен собирать при этом полезную для пользователя информацию, относящуюся к запросу пользователя.

Задачами интеллектуального агента в данной работе являются: пополнение знаний о предметной области; анализ пользовательских потребностей; формирование ранжированного списка документа на основе анализа потребностей пользователя в поисковой системе.

Для данных задач будем применять многослойный интеллектуальный агент с иерархической базой знаний, который имеет следующую архитектуру (рисунок 1).

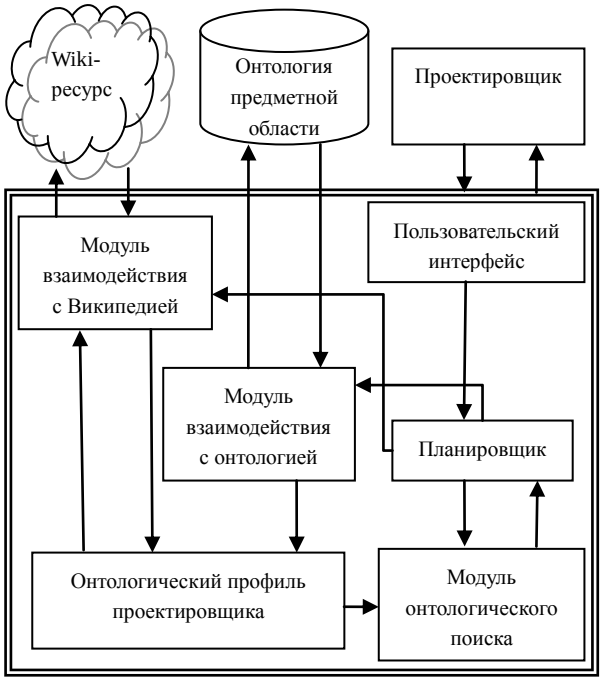

Рисунок 1. Архитектура интеллектуального агента

Интеллектуальный агент состоит из следующих слоев:

- Пользовательский интерфейс;
- Планировщик;
- Слой пользовательского профиля;
- Слой онтологического поиска;
- Слой взаимодействия с Википедией;
- Слой взаимодействия с онтологией проекта.

Пользовательский интерфейс осуществляет взаимодействие пользователя с интеллектуальным агентом. Пользователь вводит поисковые запросы, выраженные в виде набора терминов, получает список релевантных документов и оценивает документы на степень удовлетворенности своих потребностей.

Планировщик осуществляет следующие функции: преобразует пользовательский запрос к онтологическому виду, выраженный в виде множества концептов предметной области; передает преобразованный запрос модулю онтологического поиска; получает список релевантных документов и выводит в модуле пользовательского интерфейса; пополняет пользовательский профиль знаниями о предметной области.

Слой пользовательского профиля содержит знания о предметной области. Данные знания выражены в виде концептов онтологии проектов и концептов извлеченных из Википедии. В дальнейшем, пользовательский профиль используется в процессе онтологического поиска.

Слой онтологического поиска получает от планировщика онтологический запрос, выполняет поиск релевантных документов и возвращает набор ранжированных документов планировщику.

В модуле взаимодействия с Википедией реализованы функции извлечения концептов. Далее подробно рассмотрено построение профиля проектировщика на основе онтологии предметной области.

2. Онтология предметной области

Онтология предметной области представим в виде кортежа:

$$O = \langle C, W, R^D, F \rangle,$$

где C - это множество понятий предметной области, которая образует основу понятийного аппарата процесса проектирования; $W = W^S \cup W^P$ - множество терминов предметной области (W^S - множество терминов на уровне стандартов, W^P - множество терминов на уровне проектов); R^D - множество отношений:

$$R^D = \{R_G^D, R_C^D, R_A^D\},$$

где R_G^D - антисимметричное, транзитивное, нерефлексивное бинарное отношение обобщения; R_C^D - бинарное транзитивное отношение композиции («часть-целое»); R_A^D - бинарное отношение однонаправленной ассоциации.

Множество понятий C записывается следующим образом:

$$C = (C^{S_1} \cup C^{S_2} \cup ... \cup C^{S_k}) \cup (C^{W_1} \cup C^{W_2} \cup ... \cup C^{W_k}) \cup C^P,$$

где $C^{S_i}, i = \overline{1,k}$ - множество понятий предметной области, рассматриваемых в рамках *i*-ой серий стандартов, используемых в проектной организации (например, ГОСТ 34.602-89, ГОСТ 19.201-78 и т.д.); C^{W_i} - множество понятий предметной области, извлекаемые из wiki-ресурса; C^P - множество понятий предметной области, извлекаемых из ТД по реализованным проектам.

Множество интерпретирующих функций представлено в виде:

$$F^D = \{F^D_{WC}, F^D_{CW}, F^D_{C^P C^S}, F^D_{C^S C^W}\},$$

где $F^D_{WC} : \{W\} \to \{C^P\}$ - функция, сопоставляющая набору терминов подмножество понятий предметной области, задаваемая алгоритмически; $F^D_{CW} : \{C^P\} \to \{W\}$ - функция интерпретации концептов, сопоставляющая каждому концепту набор терминов из словаря; $F^D_{C^P C^S} : \{C^P\} \to \{C^S\}$ - функция интерпретации подмножества понятий на проектном уровне онтологии, позволяющая осуществить переход на уровень понятий (концептов), определенных в стандартах; $F^D_{C^S C^W} : \{C^S\} \leftrightarrow \{C^W\}$ - функция интерпретации подмножества понятий на уровне стандартов, позволяющая осуществить переход на уровень понятий (концептов), определенных в Википедии и проектных документах.

3. Формирование онтологического профиля проектировщика

Рассмотрим процесс формирования онтологического профиля проектировщика на основе информации, извлекаемой из Википедии. Википедия — свободная общедоступная мультиязычная универсальная интернет-энциклопедия, реализованная на принципах Wiki. Концепты в данной библиотеке представлены в виде HTML-страницы, для связи между страницами используются гиперссылки, тем самым гиперссылки между страницами 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 данный алгоритм поэтапно:

1. На первоначальном этапе на вход модуля взаимодействия с Википедии поступают понятия из технического задания на проектирование.

2. Далее выполняется извлечение понятий из Википедии. В основе этого процесса лежит модифицированный алгоритм волновой трассировки. Данный процесс состоит из ряда последовательных этапов:

2.1. На вход системы поступают множество понятий полученных на этапе (1).

2.2. Выполняется поиск страниц в Википедии, соответствующих полученным концептам.

2.3. Выполняется анализ страницы, полученный на этапе (2.2), результатом является нахождение тех концептов, для которых одновременно выполняются следующие условия:

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

2.4. Обнаруженные концепты добавляются в черновик профиля проектировщика.

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

2.6. Если условие (2.5) выполняется, то это означает окончания модифицированного алгоритма волновой трассировки, если не выполняется, то пункты (2.2)-(2.5) выполняются снова для концептов, полученных на этапе (2.3).

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

3. Полученные на втором этапе концепты приводятся в нормальную форму (с помощью алгоритма стемминга выделяются словарные основы концептов).

4. Удаление тех концептов, которые отсутствуют в словаре проектной организации, сформированного из терминов ТД электронного архива.

5. Идентификация типов отношений сведем к использованию логических правил, выявляющие отношения между понятиями. Экспертом подготавливается корпус текстов предметной области, которые содержат множество предложений. Каждое такое предложение представляет собой некоторую ситуацию, в котором оказались концепты, и которые свойственны тому или иному отношению и имеют предикатное представление. В основе отношений лежат трехместные предикаты [Найханова, 2008]. Для отношения «обобщение» используется предикат *PHier(a,x,y)*, которое описывает отношение $род \leftrightarrow вид$. Отношению «целое-часть» соответствует предикат *Pwp(a,x,y)*. Определим отношение «ассоциация» между концептами, если между ними не существуют отношения «целое-часть» и «обобщение». Для распознавания вида отношения между концептами в текстах предметной области, необходимо дополнительная информация об отношениях в виде термов-спутников, которые составляют устойчивые

словосочетания с глаголом семантического отношения. Для отношения «обобщение» термами-спутниками являются: «к видам», «родом», «имя», «значения» и другие. Термы-спутники: «целое», «часть», «состав», «элементом» соответствуют отношению «целое-часть». В работе [Маркарова, 1996] подробно рассмотрено различные виды конструкции с предикатами, выражающие отношения «Целое-Часть». Используя знания об отношениях, построим логические правила по их идентификации. Формально логическое правило выглядит следующим образом:

$$R=(S=\{P,\{C^W\},F^C\},\{G\}),$$

где S – ситуация, которая описывается в предложении; $P=\{p\}$ - подмножество предложений из корпусов текста предметной области; C^W - концепты, извлеченные из wiki-ресурса; F^C - множество конструкций предикатов и терм-спутников; G - подмножество, описывающее результат исполнения логических правил.

4. Формальное представление концептуального индекса электронного архива

Пусть $C=\{c_i\}, i\in I, I=\{1,2,3,...,n\}$ – конечное множество понятий предметной области, зафиксированных в онтологии; $D=\{\tilde{d}_j\}, j\in J=\{1,2,3,...,m\}$ – семейство нечетких подмножеств в C [Берштейн Л.С. и др., 2005]. Пара $\tilde{C}I=(C,D)$ называется нечетким неориентированным гиперграфом, если $\tilde{d}_j\neq\varnothing$, $j\in J$ и $\bigcup_{j\in J}\tilde{d}_j=C$. При этом $c_1,c_2,...,c_n\in C$ являются вершинами гиперграфа, а множество D, состоящее из $\tilde{d}_1,\tilde{d}_2,...,\tilde{d}_m$ – множество нечетких ребер гиперграфа.

Принимая во внимание, что отдельно взятый ТД имеет онтологическое представление как результат концептуального индексирования, множество $D=\{\tilde{d}_j\}$ будем понимать как множество ТД в концептуальном индексе, а $\tilde{d}_j$ – отдельно взятое онтологическое представление j-го ТД. Получаем, что нечеткий неориентированный граф $\tilde{C}I=(C,D)$ – концептуальный индекс электронного архива.

Два понятия концептуального индекса c_α и c_β (вершины гиперграфа) называются нечетко смежными, если существует документ (нечеткое ребро гиперграфа), которое включает оба понятия, причем величина

$$\mu(c_\alpha,c_\beta)=\bigvee_{d_i\in D}\mu_j(c_\alpha,c_\beta),\quad \text{где}$$

$$\mu_j(c_\alpha,c_\beta)=\mu_{d_j}(c_\alpha)\,\&\,\mu_{d_j}(c_\beta)$$

называется степенью смежности понятий c_α и c_β. Величина $1-\mu(c_\alpha,c_\beta)$ представляет собой расстояние между понятиями c_α и c_β, основываясь фактически на содержании архива.

Данный показатель может найти применение при уточнении запроса пользователя к архиву ТД в том случае, когда известно доминирующее понятие в запросе, но результат оставляет желать лучшего. Для уточнения запроса привлекается терминологическое окружение того понятия онтологии, которое находится на меньшем расстоянии от исходного.

Два документа $\tilde{d}_\gamma$ и $\tilde{d}_\delta$ называются нечетко смежными, если $\tilde{d}_\gamma\bigcap\tilde{d}_\delta\neq\varnothing$, причем величина

$$\mu(\tilde{d}_\gamma,\tilde{d}_\delta)=\bigvee_{c\in(\tilde{d}_\gamma\cap\tilde{d}_\delta)}\mu_{\tilde{d}_\gamma\cap\tilde{d}_\delta}(c)$$

называется степенью смежности документов $\tilde{d}_\gamma$ и $\tilde{d}_\delta$. Величина $1-\mu(\tilde{d}_\gamma,\tilde{d}_\delta)$ описывает расстояние между документами в архиве, основываясь на содержании документов и онтологии предметной области. Данный показатель может применяться в задаче нечеткой кластеризации содержимого электронного архива, т. е. там, где важную роль в целевой функции играет расстояние между центром кластера (в качестве которого может выступать гипотетический документ) и анализируемыми документами.

5. Онтологический поиск документов

Поисковый запрос, сформулированный в терминах на естественном языке $Q=\{t_i\}$, преобразуется планировщиком к концептуальному виду, но прежде выполняется ряд предобработок:

- удаление стоп-слов,
- стемминг (выделение основы слова, получение термов).

С полученным запросом $Q=\{t_i\}$ возможны два варианта обработки:

1) $Q=C$ - поисковый запрос совпадает с названием концепта онтологии предметной области.

2) $Q=W$ - поисковый запрос совпадает с терминами онтологии предметной области.

В первом случае расширение поискового запроса достигается с использованием функции интерпретации концептов, т.е. запрос дополняется терминами, формирующими терминологическое окружение концепта:

$$Q=Q\cup F_{CW}^D(c_i).$$

Во втором случае применяется функция, сопоставляющая набор терминов подмножеству понятий предметной области. Таким образом, получим множество концептов, в терминологическом окружении которых присутствуют термины запроса. К полученному множеству концептов применим функции интерпретации концептов, дополняя запрос терминами, семантически связанных с концептами:

$$Q = Q \cup F_{CW}^{D}(F_{WC}^{D}(w_i)).$$

Таким образом, расширение запроса сводится к определению прямой и обратной функции интерпретации. Полученный после обработки расширенный запрос преобразуется к концептуальному виду с помощью следующего выражения:

$$\mu_{ij} = 1 - \frac{l}{l_k}\sum_{s=1}^{l_k}\left| f_s^k - f_s \right|,$$

где f_s, f_s^k — частоты встречаемости s-го термина в запросе и в описании k-го понятия онтологии проекта, соответственно; l_k — мощность текстового входа понятия c_k. В том случае, если термин s отсутствует в запросе, f_s принимается равным нулю; μ_{ij} - величина, характеризующая степень выраженности концепта [Филиппов и др., 2013] в запросе, где $0 \le \mu_{ij} \le 1$.

В традиционных моделях информационного поиска нахождение степени релевантности документа обозначается величиной, которая называется мерой сходства запроса к документу. В данной работе информационный запрос и ТД имеют концептуальные представления, т.е. рассматриваются в виде нечеткого множества. Для вычисления меры сходства между запросом и документом воспользуемся термином «степень включения», которое соответствует операцией нечеткого включения множества [Берштейн и др., 2005]. Мера включение запроса и ТД вычисляется по следующей формуле:

$$\gamma(\overline{I_q}, \overline{I_d}) = \&(\mu_{I_q}(c) \to \mu_{I_d}(c)),$$

где $\overline{I_q}, \overline{I_d}$ - концептуальные индексы запроса и ТД соответственно, c - концепты онтологии, где $0 \le \gamma(\overline{I_q}, \overline{I_d}) \le 1$.

6. Вычислительные эксперименты.

В ходе научно-исследовательской работы были разработаны следующие интеллектуальные подсистемы: подсистема автоматизированного формирования онтологической сети с использованием Wiki-ресурсов и подсистема онтологического поиска ТД, основанная на использование информационных потребностях пользователя. С данными подсистемами были проведены вычислительные эксперименты, которые показали следующие результаты. В подсистему автоматизированного формирования онтологической сети был загружен концепт «СУБД». Данная подсистема сформировала концептуальную сеть (рисунок 2).

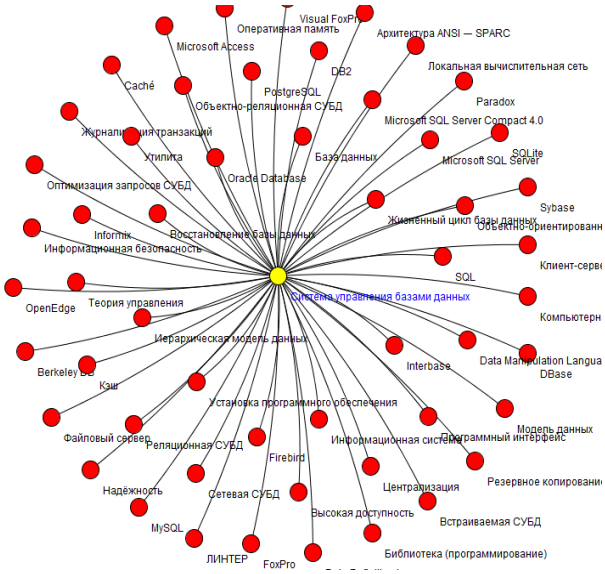

Рисунок 2. Сформированная концептуальная сеть

Были проведены вычислительные эксперименты с подсистемой онтологического поиска. Для оценки качества информационного поиска ТД использовались величина F-мера. Данная величина учитывает в себе два параметра: точность и полнота. F-мера 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:

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

Полученные результаты сравнивались с результатами следующих систем информационного поиска: Яндекс Персональный поиск (ЯПП), Архивариус 3000 (А300), AOL Desktop Search (AOL),Copernic Desktop Search (CDS). Сравнительные результаты эксперимента представлены на рисунке 3.

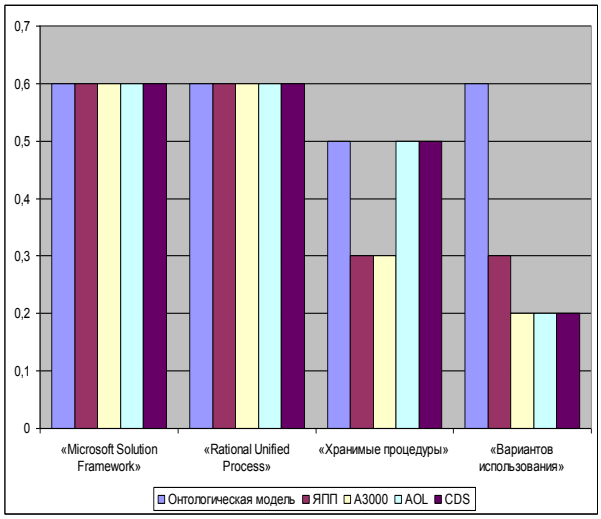

Рисунок 3. Результаты эксперимента

Как видно из рисунка, лучшее качество поиска онтологическая модель показывает в том случае, если в запросе присутствует высокая неопределенность в определении термина к конкретной тематике.

Заключение

В данной работе предлагается использовать интеллектуальные агенты в процессе поиска ТД. Интеллектуальный агент пополняет пользовательский профиль, обрабатывает поисковые запросы и выводит список ранжированных документов. В процессе научного исследования были разработаны подсистемы онтологического поиска и подсистема автоматизированного формирования онтология, с которыми были проведены вычислительные эксперименты.

Данная работа выполнена при частичной финансовой поддержки РФФИ, проект №14-01-31086.

Библиографический список

[Берштейн и др, 2005.] Берштейн Л.С., Боженюк А.В. Нечеткие графы и гиперграфы. М.: Научный мир, 2005 – 256 с.

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

[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.

[Маркарова, 1996] Маркарова Т.С., Конструкции с предикатами, выражающими отношения "Часть-целое" в современном русском языке. : автореф. дис. на соиск. учен. степ.канд.филол. наук (10.02.01);МГУ – М., 1996 - 23 с.

[Найханова, 2008] Найханова Л.В., Технология создания методов автоматического построения онтологий с применением генетического и автоматного программирования. – Улан-Удэ.:БНЦ СО РАН, 2008. – 237 с.

[Рассел и др., 2006] Рассел С., Норвиг П., Искусственный интеллект. Современный подход. – М.:Вильямс, 2006. – 1408 с.

[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.

ONTOLOGICAL APPROACH TO THE FORMATION OF PROJECT REQUESTS OF INTELLIGENT AGENT

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

** Ulyanovsk State Technical University, Ulyanovsk, Russia*

nam@ulstu.ru

subkhangulov-ruslan@yandex.ru

Introduction

In this paper is considered the use of intelligent agents in the formation of information requests to the electronic archive of technical documents. Intelligent agent contains an ontological description of the profile of the designer, which is expressed as a fragment of domain ontology. Tasks intellectual agent are addition information on user needs and finding text documents requested by the user of an electronic archive.

Main Part

In this article we propose to use intelligent agents in process of search for technical documents. Intelligent agent in process of search using user profiles and ontological model of search. The user profile contains a fragment of domain ontology. Thus, the search engine has information about preferences of user. This information helps to improve the quality of the search of documents

In the process of scientific research have been developed subsystem of ontological search and subsystem of automated formation of ontology.

Conclusion

Computational experiments were performed with the subsystem ontological search. Experiments showed that the ontological model search shows better results than the traditional model