

УДК 681.3

А.А. Филиппов, Д.О. Шалаев, В.В. Шеркунов

РЕАЛИЗАЦИЯ ВЕБ-ОРИЕНТИРОВАННОЙ СИСТЕМЫ АНАЛИЗА ЭЛЕКТРОННОГО АРХИВА ТЕХНИЧЕСКИХ ДОКУМЕНТОВ НА ОСНОВЕ ОНТОЛОГИИ¹

Филиппов Алексей Александрович, кандидат технических наук, доцент 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: al.filippov@ulstu.ru].

Шалаев Денис Олегович, магистрант кафедры «Информационные системы» УлГТУ, окончил факультет информационных систем и технологий УлГТУ. Имеет статьи в области интеллектуальных систем хранения и обработки информации. [e-mail: melges@post.ru].

Шеркунов Вячеслав Владимирович, аспирант кафедры «Информационные системы» УлГТУ, окончил факультет информационных систем и технологий УлГТУ. Имеет статьи в области интеллектуальных систем хранения и обработки информации. [e-mail: v.sherkunov@gmail.com].

Аннотация

В статье содержится описание и результаты экспериментов с веб-ориентированной системой формирования навигационной структуры электронного архива (ЭА) технических документов (ТД) на основе онтологических моделей. Предварительное индексирование документов выполняется с целью получения онтологических представлений. Онтология предметной области проектной организации представлена в формате OWL. Основными функциями данной системы являются: получение и хранение онтологических представлений ТД, а также формирование навигационной структуры ЭА с целью нахождения множества документов, схожих в семантическом смысле с определенным документом, представляющим интерес для проектировщика. Для хранения онтологии предметной области и онтологических представлений ТД используется графовая база данных Neo4j, взаимодействие с форматом OWL осуществляется с помощью библиотеки OWL API. Система реализована на платформе Java Enterprise Edition.

Ключевые слова: интеллектуальная система, структуризация, онтология, навигационная структура, электронный архив, техническая документация.

THE IMPLEMENTATION OF WEB-ORIENTED SYSTEM OF FORMATION OF NAVIGATION STRUCTURE OF ELECTRONIC ARCHIVE FOR TECHNICAL DOCUMENTATION ON THE BASIS OF ONTOLOGICAL MODELS

Aleksei Aleksandrovich Filippov, Candidate of Engineering; Associate Professor at the Department of Information Systems at Ulyanovsk State Technical University; graduated from the Faculty Information Systems and Technologies of Ulyanovsk State Technical University; an author of articles in the field of intelligent systems for data storage and processing. e-mail: al.filippov@ulstu.ru.

Denis Olegovich Shalaev, Candidate for the Master's Degree at the Department of Information Systems of Ulyanovsk State Technical University; graduated from the Faculty of Information Systems and Technologies of Ulyanovsk State Technical University and got the Bachelor's Degree in 2015; an author of articles in the field of intelligent systems of data storage and processing. e-mail: melges@post.ru.

Viacheslav Vladimirovich Sherkunov, Post-graduate Student at the Department of Information Systems at Ulyanovsk State Technical University; graduated from the Faculty graduated from the Faculty of Information Systems and Technologies of Ulyanovsk State Technical University; an author of articles in the field of intelligent systems of data storage and processing. e-mail: al.filippov@ulstu.ru.

Abstract

The article considers a description and experimental results of web-oriented system of formation of electronic archive

¹ Работа выполнена при частичной финансовой поддержке РФФИ (проект № 14-01-31080 мол_а).

(EA) for technical documentation (TD) on the basis of ontological models. Preliminary indexing of documents is realized for the purpose of receiving ontological views. Subject area ontology of design organization is represented in OWL files. The main functions of the system include receipt and storage TD ontological representations, formation of EA navigation structure with the purpose of searching numerous documents similar to the certain document (of interest to the designer) in semantics. In order to store the subject area ontology and TD ontological representations, the graph database Neo4j is used. The interaction with OWL format is executed through the use of OWL API library. The system is implemented on the basis of Java Enterprise Edition platform.

Key words: intelligent system, structuring, ontology, navigation structure, electronic archive, technical documentation.

ВВЕДЕНИЕ

В процессе интеллектуального анализа текстовых документов следует учитывать различные способы извлечения знаний из текста, например, поиск ключевых понятий по частым наборам слов, идентификацию фактов (события или отношения) и их характеристик и т. д. В задаче извлечения ключевых понятий из текста интерес представляют некоторые сущности, события и отношения. При этом извлеченные понятия анализируются и используются для вывода новых понятий.

В [1] сказано, что извлечение ключевых понятий из текстовых документов можно рассматривать как фильтрацию больших объемов текста. Данный процесс включает в себя отбор документов из коллекции и пометку определенных термов в тексте. Существуют различные подходы к извлечению информации из текста. Примером может служить определение частых наборов слов и объединение их в ключевые понятия, другим подходом является идентификация в текстах и извлечение их характеристик. Фактами являются некоторые события или отношения. Идентификация производится с помощью набора образцов. Образцы представляют собой возможные лингвистические варианты фактов.

Кластеризация в процессе интеллектуального анализа текстовых документов является одним из этапов анализа данных, построения законченного аналитического решения. Аналитику часто легче выделить группы схожих объектов, изучить их особенности и построить для каждой группы отдельную модель, чем создавать одну общую модель для всех данных.

Онтология – артефакт, структура, описывающая значения элементов некоторой системы [2].

Неформально, онтология представляет собой некоторое описание взгляда на мир применительно к конкретной области интересов. Это описание состоит из терминов и правил использования этих терминов, ограничивающих их значения в рамках конкретной области.

На формальном уровне, онтология – это система, состоящая из набора понятий и набора утверждений об этих понятиях, на основе которых можно строить классы, объекты, отношения, функции и теории.

В задачах интеллектуального анализа онтология выступает в качестве описания предметной области [3–7], используется при решении задач кластеризации и классификации [8, 9].

При анализе существующих методов индексирования и кластеризации текстовых документов был выявлен ряд проблем:

- недостатки традиционных методов индексирования, при представлении текстовых документов как простого набора слов (избыточность, независимость и многозначность слов);

- большая вычислительная сложность традиционных методов кластеризации;

- отсутствие учета специфики технической документации в существующих методах структуризации текстовых ресурсов на основе кластеризации.

Также в настоящее время многие производители программных средств предлагают продукты и решения в области интеллектуального анализа текстовых документов:

- Intelligent Miner for Text (IBM);
- PolyAnalyst (Megaputer Intelligence);
- SAS Text Miner (SAS Institute);
- Rocket AeroText (Rocket Software);
- Oracle Text (Oracle);
- Rosette Linguistics Platform (Basis Technology);
- Cogito Categorizer (Expert System);
- NetOwl Text Mining (NetOwl);
- Semaphore (Smartlogic).

Данные программные системы позволяют производить процедуры индексирования, кластеризации и информационного поиска ТД, но при этом не учитывают специфику предметной области проектной организации и особенности технической документации. А также требуют покупки лицензий на право использования.

Таким образом, была разработана веб-ориентированная система, использующая адаптированные методы онтологического анализа для решения задач индексирования и структуризации ТД с учетом модели жизненного цикла (ЖЦ) и стандартов, применяемых на предприятии при проектировании.

1 АРХИТЕКТУРА СИСТЕМЫ

Система формирования навигационной структуры ЭА ТД на основе онтологических моделей является веб-ориентированным приложением, архитектура данной системы представлена на рисунке 1.

Как видно из рисунка, система разработана с использованием платформы Java Enterprise Edition и выполняется на сервере приложений Wildfly.

Ядро системы осуществляет основные функции авторизации, контроля прав доступа, управления настройками системы и реализует пользовательский интерфейс. В качестве хранилища данных ядра системы используется система управления базами данных PostgreSQL.

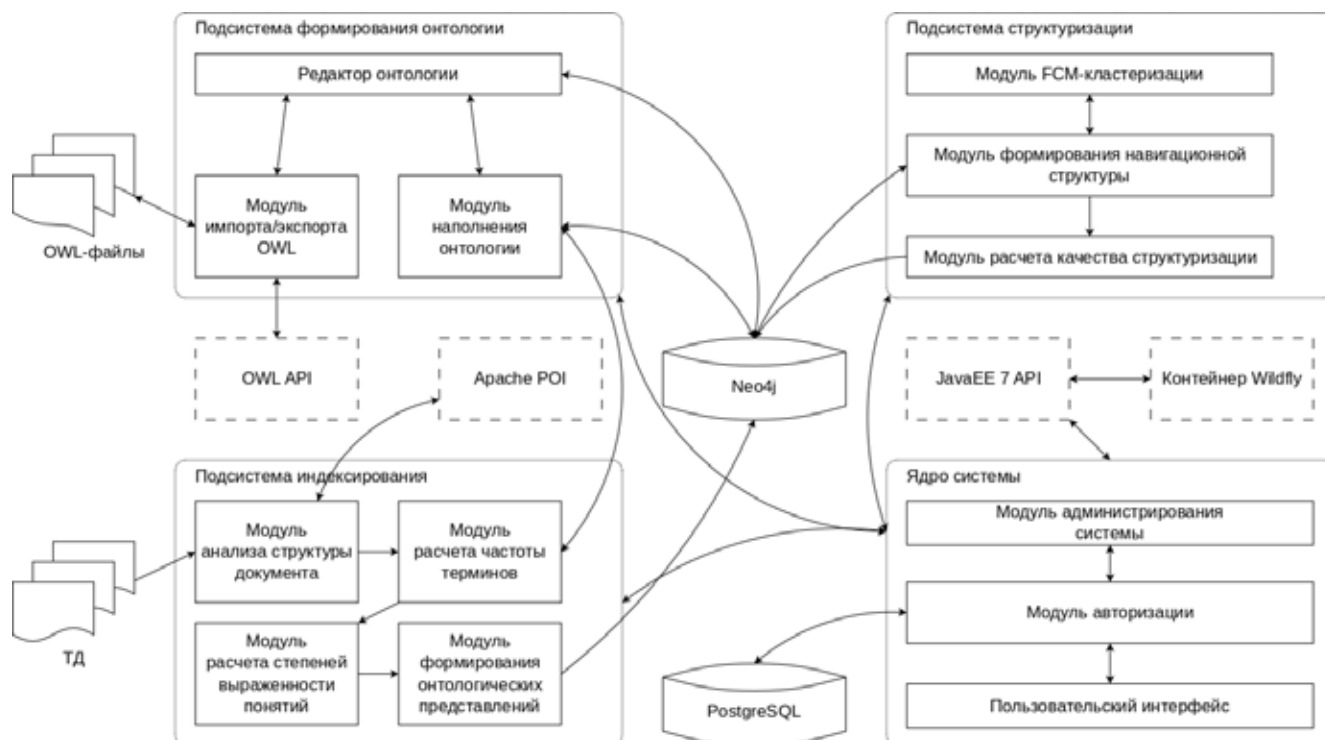


Рис. 1. Архитектура системы формирования навигационной структуры ЭА ТД на основе онтологических моделей

Подсистема формирования онтологии используется для автоматизированного построения онтологии предметной области проектной организации и осуществления операций импорта (экспорта) онтологии в файлы формата OWL. Взаимодействие с OWL-файлами производится с помощью библиотеки OWL API. Сформированная модель предметной области проектной организации хранится в графовой базе данных Neo4j в виде дерева.

При построении модели предметной области в виде прикладной онтологии для решения задач анализа ТД необходимо сформулировать основные требования к онтологии. Такие требования должны опираться на особенности предметной области и, кроме того, на особенности тех информационных ресурсов, которые подвергаются анализу. Цель применения онтологии заключается в привлечении дополнительных знаний об окружающей среде проектируемых средств при анализе документации для сокращения времени поиска необходимых документов. Фактически, для отдельно взятого информационного ресурса из ЭА ТД онтология задает новую систему координат, в которой кластер документов может рассматриваться как группа связанных между собой по смыслу ТД.

На основании вышесказанного формулируем основные требования к прикладной онтологии [10]:

1. Онтология ЭА проектной организации должна включать в себя описание применяемых в организации моделей ЖЦ проектируемых систем;
2. Структура онтологии должна основываться на использовании множества возможных серий стандартов, для каждой из которых определяется свой набор понятий и отношений между ними;
3. Множество понятий онтологии должно включать в себя те понятия, которые соответствуют уже реализо-

ванным проектам, результаты которых в документальном виде зафиксированы в ЭА.

4. Предметная область проектирования сложных систем накладывает определенные требования к структуре прикладной онтологии [10–13]. Жесткая привязка к применяемым на различных стадиях проектирования стандартам влечет за собой необходимость формирования онтологии, состоящей из множества уровней (рис. 2).

Определение: Терминологическое окружение понятия предметной области – множество терминов (слов) из ТД проектов в ЭА, которые наиболее близки с данным понятием в семантическом смысле.

Таким образом, подсистема формирования онтологии позволяет зафиксировать модель предметной области проектной организации в виде онтологии в формате OWL, удовлетворяющей всем требованиям, предъявленным выше. Также данная подсистема предоставляет возможность автоматизированного формирования терминологического окружения понятий (с использованием метода, описанного в [10, 14], для расчета частоты терминов называется соответствующий модуль из подсистемы индексирования), которые соответствуют уже реализованным проектам, результаты которых в документальном виде зафиксированы в ЭА.

Подсистема индексирования позволяет формировать онтологические представления ТД. Работа с документами в формате Microsoft Office осуществляется с помощью библиотеки Apache POI. Полученные в результате работы подсистемы онтологические представления ТД сохраняются в графовую базу данных Neo4j и представляют собой дерево или набор деревьев, который соответствует отдельным узлам онтологии предметной области.

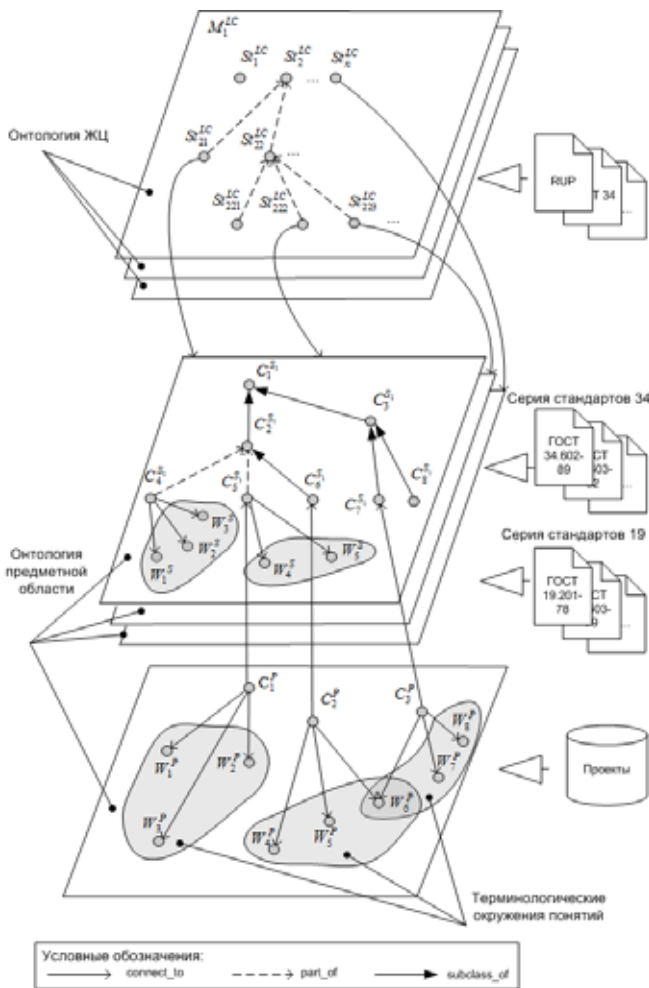


Рис. 2. Структура прикладной онтологии ЭА

В основе онтологического индексирования ТД лежит следующая функция:

$$F_{ov} : ch_j^d \rightarrow oV_j^d,$$

где ch_j^d – j -й раздел ТД d ;

oV_j^d – онтологическое представление j -го раздела ТД d .

На первом шаге работы подсистемы индексирования используется модуль анализа структуры документа для выделения разделов ТД.

На втором шаге работы подсистемы индексирования с помощью модуля расчета частоты терминов определяется терминологическая составляющая каждого раздела ТД.

Терминологическая составляющая j -го раздела ТД записывается в виде множества пар «термин–частота»:

$$\left\{ (w_{1j}^d, f_{1j}^j), (w_{2j}^d, f_{2j}^j), \dots, (w_{ij}^d, f_{ij}^j), \dots, (w_{l_j}^d, f_{l_j}^j) \right\},$$

где l_j – количество определенных терминов в j -м разделе ТД после фильтрации стоп-слов.

В основе метода расчета нормализованного веса термина w_{ij}^d в составе j -го раздела ТД лежит следующая зависимость:

$$f_i^j = 1 + \log \left(tf_{w_{ij}^d} \right) \cdot \log \left(\frac{N}{dt} \right) \cdot \frac{1}{\sqrt{tf_{w_{1j}^d}^2 + tf_{w_{2j}^d}^2 + \dots + tf_{w_{nj}^d}^2}},$$

$$1 \leq i \leq n,$$

где f_i^j – нормализованный вес термина w_{ij}^d в j -м разделе документа;

$tf_{w_{ij}^d}$ – частота встречаемости термина w_{ij}^d ;

N – общее количество документов;

dt – количество документов, содержащих термин w_{ij}^d ;

n – количество терминов в j -м разделе документа.

Определение: Степень выраженности понятия онтологии интеллектуального ЭА – степень совпадения терминологического окружения понятия с набором терминов некоторого фрагмента ТД при условии, что в терминологическое окружение включены термины, наиболее близкие в семантическом отношении с понятием.

На третьем шаге работы подсистемы индексирования используется модуль расчета степеней выраженности понятий для каждого раздела ТД. Вычисление степеней выраженности понятий уровня онтологии производится с применением аппарата нечетких соответствий [10, 15].

На четвертом шаге работы подсистемы индексирования с использованием модуля формирования онтологических представлений получаем первоначальное онтологическое представление каждого раздела документа в виде:

$$oV_j^d = \langle ch_j, \{ \hat{C}_j^P \cup \hat{C}_j^S \} \rangle, \hat{C}_j^P \subseteq C^P, \hat{C}_j^S \subseteq C^S |_{S_k^{LC}},$$

где \hat{C}_j^P, \hat{C}_j^S – первоначальные (ориентировочные) наборы понятий уровней проектов и стандартов соответственно, которые требуют уточнения.

Заключительным шагом в формировании онтологического представления ТД является уточнение набора понятий уровня стандартов, которое опирается на найденное подмножество понятий в ТД уровня проектов онтологии [10].

Реализуя вышеуказанные процедуры, получаем окончательные онтологические представления для каждого j -го раздела документа в следующем виде:

$$oV_j^d = \langle ch_j, \{ C_j^P \cup C_j^S \} \rangle, C_j^P \subseteq C^P, C_j^S \subseteq C^S |_{S_k^{LC}}.$$

Подсистема структуризации формирует навигационную структуру ЭА ТД путем нахождения расстояния между документами. Рассмотрим данную меру в контексте понятий онтологии, относящихся к уровню стандартов, представив каждое онтологическое представление документа в качестве дерева (иерархии) понятий предметной области. Такая иерархия определяется путем нахождения минимального дерева, включающего все понятия из онтологического представления [10, 16].

Редакционное расстояние между иерархиями определяется на основе вычисления стоимости редакционной операции, которая определяется отдельно для каждого типа семантического отношения. Так, для отношения обобщения редакционную операцию будем записывать

как $\varphi_{S_i}(R_G^D)$, а для отношения «часть–целое» – $\varphi_{S_i}(R_C^D)$. Индекс S_i показывает принадлежность значения редакционной операции к i -й группе стандартов. Фактически, в задаче кластеризации ТД под редакционной операцией будем понимать вес соответствующего отношения, принимающий значение в диапазоне от 0 до 1 и имеющий различные значения в рамках каждой группы стандартов.

Итоговое редакционное расстояние между иерархиями вычисляется по следующей формуле:

$$\tau_{ov}^* = \max_i \left(\sum_{s=1}^m \varphi_{S_i}(R_G^D)_s + \sum_{l=1}^n \varphi_{S_i}(R_C^D)_l \right),$$

где i – номер группы стандартов;

s – номер добавляемого отношения обобщения;

l – номер добавляемого отношения «часть–целое».

Итоговое редакционное расстояние вычисляется как наибольшее из редакционных расстояний, определенных для каждой группы стандартов.

Коэффициент нормализации T_{ov}^* рассчитываем исходя из всех семантических отношений обобщенной иерархии. Таким образом, меру расстояния между онтологическими представлениями ТД будем определять с помощью следующего выражения:

$$\|OV^{d_1} - OV^{d_2}\| = \frac{\tau_{ov}^*}{T_{ov}^*}.$$

Для выполнения процесса формирования навигационной структуры в виде вложенного набора кластеров ТД необходимо решить задачу настройки весов семантических отношений между понятиями онтологии на уровне стандартов. Ранее весовые коэффициенты для отношения обобщения и для отношения «часть–целое» были определены как $\varphi_{S_i}(R_G^D)$ и $\varphi_{S_i}(R_C^D)$ соответственно.

Поскольку указанные отношения применяются между понятиями в онтологии для различных групп стандартов, будем предполагать, что их оптимальные значения для каждой такой группы (в рамках своей иерархии понятий) в общем случае будут различными. Сформулируем принцип оптимальности для весовых коэффициентов семантических отношений онтологии.

Пусть $\{OV^{d_i}\}^*$ – множество онтологических представлений документов, входящих в эталонную выборку (для которой известно экспертное разделение документов на классы). Справедливо следующее соотношение:

$$\{OV^{d_i}\}^* \subset \{OV^{d_i}\},$$

где $\{OV^{d_i}\}$ – полное множество онтологических представлений ТД интеллектуального ЭА. В онтологии предметной области на уровне стандартов определены отношения обобщения и отношения «часть–целое» на множестве понятий с соответствующими весовыми коэффициентами $\varphi_{S_i}(R_G^D)$ и $\varphi_{S_i}(R_C^D)$, где S_i – i -я группа стандартов, используемая при создании онтологии.

Множество $\{OV^{d_i}\}^*$ состоит из двух подмножеств: $\{OV^{d_i}\}_+^* \cup \{OV^{d_i}\}_-^*$, которые соответствуют экспертному разбиению документов на два заранее определенных

класса. Задача оптимизации весовых коэффициентов семантических отношений будет состоять в нахождении такого множества коэффициентов

$$\left\{ \langle \varphi_{S_1}^*(R_G^D), \varphi_{S_1}^*(R_C^D) \rangle, \langle \varphi_{S_2}^*(R_G^D), \varphi_{S_2}^*(R_C^D) \rangle, \dots, \langle \varphi_{S_n}^*(R_G^D), \varphi_{S_n}^*(R_C^D) \rangle \right\},$$

при которых качество кластеризации, определяемое выражением:

$$F^* = \frac{\max(\bar{K}_+ + \bar{K}_-, \hat{K}_+ + \hat{K}_-)}{N} \rightarrow \min, \quad (1)$$

было бы наилучшим. В выражении (1) \hat{K}_- и \bar{K}_- – множества отсутствующих документов соответственно в первом и во втором кластерах, \hat{K}_+ и \bar{K}_+ – множества лишних документов соответственно в первом и во втором кластерах, N – количество документов.

При наличии экспертного разбиения выборки ТД подсистема структуризации с помощью выражения 1 может рассчитать качество полученной навигационной структуры, для чего используется соответствующий модуль.

2 РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Для проведения экспериментов с разработанной системой экспертом была подготовлена выборка, состоящая из 5017 ТД и сгруппированная по разным основаниям классификации:

- по классу документации – 4 группы (ЕСКД, ЕСПД, ЕСТД, Нормативные);
- по виду документации – 52 группы (ГОСТ 2.601, 2.602, 2.102, 2.701 и 3.1201);
- по разделу документации – 27 групп (ГОСТ 2.106 и сложившаяся на предприятии система присвоения документам десятичных номеров);
- по тематике работ – 28 групп (изделия, рассматриваемые в документах).

При этом данная выборка ТД является смешанной и примерно на 1/20 часть состоит из нормативных документов.

При анализе результатов работы системы была использована онтология предметной области, содержащая две серии стандартов:

1. Информационные технологии. Взаимосвязь открытых систем (ГОСТ 34). Содержит 108 понятий онтологии уровня стандартов.

2. Единая система программной документации (ГОСТ 19). Содержит 111 понятий онтологии уровня стандартов.

Уровень проектов онтологии сформирован на основе выборки, состоящей из 5017 ТД. Данный уровень содержит 81 понятие и 10078 уникальных терминов, составляющих терминологическое окружение понятий уровня проектов.

Таким образом, онтология предметной области имеет в своем составе 300 понятий: 219 понятий на уровне стандартов и 81 понятие на уровне проектов, а также 10078 уникальных терминов на уровне проектов.

Таблица 1

Оценка качества структуризации тестового множества ТД
(больше – лучше)

Вид разбиения	Oracle Text	Традиционное представление	Онтологическое представление	Онтологическое представление с учетом ЖЦ
Класс документации	0,42	0,40	0,39	0,48
Раздел документации	0,33	0,25	0,33	0,39
Тематика работ	0,20	0,12	0,27	0,34
Вид документа	0,50	0,05	0,17	0,26

Результаты работы системы сравнивались с результатами экспертного разбиения и результатами работы системы Oracle Text, так как данный программный продукт, в соответствии с текстом Oracle Technology Network Developer License Terms, может быть использован для некоммерческих целей в рамках разработки, тестирования, демонстрации и анализа прототипа приложения.

Для проведения экспериментов по оценке качества структуризации был построен индекс, содержащий в своем составе онтологические и традиционные представления ТД (множество пар «термин-частота»). На следующем шаге полученный индекс был подвергнут различным вариантам структуризации с последующим расчетом качества:

- структуризация средствами системы Oracle Text (FCM-алгоритм) традиционных представлений ТД;
- структуризация средствами разработанной системы с использованием FCM-алгоритма кластеризации традиционных представлений ТД;
- структуризация средствами разработанной системы с использованием модифицированного FCM-алгоритма кластеризации онтологических представлений ТД;
- структуризация средствами разработанной системы с использованием модифицированного FCM-алгоритма кластеризации онтологических представлений ТД с учетом моделей ЖЦ.

Оценка качества структуризации тестовой выборки ТД представлена в таблице 1. Лучшие значения оценочной функции представлены для онтологических представлений с учетом моделей ЖЦ, лишь для структуризации по виду документа система Oracle Text показала лучшие результаты. Отдельное внимание следует уделить разбиению тестовой выборки ТД по тематике работ, так как данный вариант структуризации наиболее близко соответствует основной задаче разработанной системы – структуризации ЭА ТД по содержанию отдельных документов. По данным оценочной функции результаты структуризации онтологических представлений ТД с учетом моделей ЖЦ примерно на 40% лучше по сравнению с результатами системы Oracle Text.

ЗАКЛЮЧЕНИЕ

Онтологическая модель ТД, учитывающая модели ЖЦ, является эффективным средством при решении задачи структуризации множества ТД. Сравнение с традиционной моделью ТД показало, что онтологическая модель ТД

показывает лучшие результаты качества структуризации. По данным экспериментов, результаты структуризации онтологических представлений ТД с учетом моделей ЖЦ примерно на 40% лучше по сравнению с результатами системы Oracle Text.

Таким образом, использование онтологически ориентированных методов индексирования и структуризации, учитывающих применяемые в процессе проектирования серии стандартов и модели ЖЦ, предпочтительно при необходимости более качественного построения навигационной структуры ЭА.

Разработанная программная система может применяться в проектных организациях для более эффективной организации ЭА технической документации.

СПИСОК ЛИТЕРАТУРЫ

1. Анализ данных и процессов: учеб. пособие / А.А. Баргесян [и др.]; – 3-е издание. – СПб. : БХВ-Петербург, 2009. – 512 с.
2. Онтологии и тезаурусы : учеб. пособие / В.Д. Соловьев [и др.]. – М. : Бином, 2006. – 157 с.
3. Ермаков А.Е. Автоматизация онтологического инжиниринга в системах извлечения знаний из текста // Матер. междунар. конф. «Диалог 2008». – М., 2008. – С. 154–159.
4. Загоруйко Ю.А., Кононенко И.С., Сидорова Е.А. Семантический подход к анализу документов на основе онтологии предметной области. – URL: <http://www.dialog21.ru/digests/dialog2006/materials/html/SidorovaE.html>.
5. Конотоп Д.И., Зинченко В.П. Оптимальное проектирование сложных технических объектов с использованием онтологического подхода // Онтология проектирования. – 2011. – № 1 (2). – С. 44–53.
6. Соснин П.И., Маклаев В.А. Инструментальные средства для спецификации концептуализаций в проектировании автоматизированных систем // Онтология проектирования. – 2012. – № 1 (3). – С. 32–52.
7. Mathieu d'Aquin, Gabriel Kronberger, Mari Carmen Suárez-Figueroa Combining Data Mining and Ontology Engineering to enrich Ontologies and Linked Data // Proceedings of the First International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data. – Heraklion, Greece, 2012. – URL: <http://ceur-ws.org/Vol-868/paper3.pdf>.
8. Dejing Dou, Hao Wang, Haishan Liu Semantic Data Mining: A Survey of Ontology-based Approaches // IEEE

International Conference Semantic Computing (ICSC) – 2015. – URL: http://ix.cs.uoregon.edu/~dou/research/papers/icsc15_invited.pdf.

9. Sathya Priya J. Clustering Technique in Data Mining for Text Documents // *International Journal of Computer Science and Information Technologies*. – 2012. – 1 (3). – pp. 2943–2947.

10. Наместников А.М., Филиппов А.А. Формирование навигационной структуры электронного архива технической документации на основе онтологии // *Радиотехника*. – 2014. – №11. – С. 108–117.

11. Норенков И.П. Основы автоматизированного проектирования : учебник для вузов. – 4-е изд., перераб. и доп. – М. : Изд-во МГТУ им. Н.Э. Баумана, 2009. – 432 с.

12. Токмаков Г.П. Онтологии и их применение для интеграции информационных ресурсов // *Автоматизация процессов управления*. – 2010. – № 1 (19). – С. 37–49.

13. Куприянов А.А., Мельниченко А.С. Модели структуризации и формализации онтологии предметной области на стадиях проектирования автоматизированных систем // *Автоматизация процессов управления*. – 2010. – № 2 (20). – С. 70–75.

14. Physical and Semantic Relations to Build Ontologies for Representing Documents / J. Serrano-Guerrero, J.A. Olivas, J. de la Mata, P. Garces // *Fuzzy logic, Soft Computing and Computational Intelligence (Eleventh International Fuzzy Systems Association World Congress IFSA) / Beijing, China*. Tsinghua University Press, 2005. – Vol. I. – pp. 503–508.

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

16. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск : ИМ СО РАН, 1999. – 270 с.

REFERENCES

1. Bargesyan A.A. et al. *Analiz dannykh i protsessov: uchebnoe posobie 3-e izdanie* [Data and Process Analysis: Textbook. The 3d Edition]. St. Petersburg, BKHV-Peterburg Publ., 2009. 512 p.

2. Solovev V.D. et al. *Ontologii i tezaury: uchebnoe posobie* [Ontologies and Thesauri: Textbook]. Moscow, Binom Publ., 2006. 157 p.

3. Ermakov A.E. Avtomatizatsiia ontologicheskogo inzhiniringa v sistemakh izvlecheniia znaniia iz teksta [Automation of Ontology Engineering in the Systems for Text Knowledge Extraction]. *Mater. mezhdunar. konf. "Dialog 2008"* [Proc. of the Intern. Conf. "Dialog-2008"], Moscow, 2008, pp. 154–159.

4. Zagorulko Yu.A., Kononenko I.S., Sidorova E.A. *Semanticheskii podkhod k analizu dokumentov na osnove ontologii predmetnoi oblasti* [A Semantic Approach to the Document Analysis on the Basis of Domain Ontology]. Available at: www.dialog21.ru/digests/dialog2006/materials/html/SidorovaE.html.

5. Konotop D.I., Zinchenko V.P. Optimalnoe proektirovanie slozhnykh tekhnicheskikh obektov s ispolzovaniem ontologicheskogo podkhoda [An Optimal Designing of Complex Technical Objects Using the

Ontological Approach]. *Ontologiiia proektirovaniia* [Ontology of Designing], 2011, no. 1(2), pp. 44–53.

6. Sosnin P.I., Maklaev V.A. Instrumentalnye sredstva dlia spetsifikatsii kontseptualizatsii v proektirovanii avtomatizirovannykh sistem [Tools for the Specification of the Conceptualization in Automated System Designing]. *Ontologiiia proektirovaniia* [Ontology of Designing], 2012, no. 1(3), pp. 32–52.

7. Mathieu d'Aquin, Gabriel Kronberger, Mari Carmen Suárez-Figueroa. Combining Data Mining and Ontology Engineering to Enrich Ontologies and Linked Data. *Proc. of the First International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data*. Heraklion, Greece, 2012. Available at: <http://ceur-ws.org/Vol-868/paper3.pdf>.

8. Dejing Dou, Hao Wang, Haishan Liu. Semantic Data Mining: A Survey of Ontology-based Approaches. *IEEE International Conference Semantic Computing (ICSC) – 2015*. Available at: http://ix.cs.uoregon.edu/~dou/research/papers/icsc15_invited.pdf.

9. Sathya Priya J. Clustering Technique in Data Mining for Text Documents. *International Journal of Computer Science and Information Technologies*. 2012, no. 1(3), pp. 2943–2947.

10. Namestnikov A.M., Filippov A.A. Formirovanie navigatsionnoi struktury elektronogo arkhiva tekhnicheskoi dokumentatsii na osnove ontologii [Formation of Navigation Structure of Electronic Archive of Technical Documentation on the Basis of Ontology]. *Radiotekhnika* [Radioengineering], 2014, no. 11, pp. 108–117.

11. Norenkov I.P. *Osnovy avtomatizirovannogo proektirovaniia: uchebnik dlia vuzov*. 4-e izd., pererab. i dop. [Fundamentals of Computer-Aided Design: Textbook, the 4d revised and enlarged Edition]. Moscow, MGTU im. N.E. Bauman Publ., 2009. 432 p.

12. Tokmakov G.P. Ontologii i ikh primenenie dlia integratsii informatsionnykh resursov [Ontologies and their Use for Information Resource Integration]. *Avtomatizatsiia protsessov upravleniia* [Automation of Control Processes], 2010, no. 1(19), pp. 37–49.

13. Kupriyanov A.A., Melnichenko A.S. Modeli strukturizatsii i formalizatsii ontologii predmetnoi oblasti na stadiakh proektirovaniia avtomatizirovannykh sistem [Models of Structuring and Formalization for Ontology of Application Domain of Computer-Aided System Design]. *Avtomatizatsiia protsessov upravleniia* [Automation of Control Processes], 2010, no. 2(20), pp. 70–75.

14. Serrano-Guerrero J., Olivas J. A., de la Mata J., Garces P. Physical and Semantic Relations to Build Ontologies for Representing Documents. *Fuzzy logic, Soft Computing and Computational Intelligence (Eleventh International Fuzzy Systems Association World Congress IFSA)*, Beijing, China. Tsinghua University Press, 2005, vol. 1, pp. 503–508.

15. Bershtein L.S., Bozhenuk A.V. *Nechetkie grafy i gipergrafy* [Fuzzy Graphs and Hypergraphs]. Moscow, Nauchnyi Mir Publ., 2005. 256 p.

16. Zagoruyko N.G. *Prikladnye metody analiza dannykh i znaniia* [Applied Approaches to Data and Knowledge Analysis]. Novosibirsk, IM SO RAN Publ., 1999. 270 p.