На правах рукописи

Филиппов Алексей Александрович

ФОРМИРОВАНИЕ НАВИГАЦИОННОЙ СТРУКТУРЫ ЭЛЕКТРОННОГО АРХИВА ТЕХНИЧЕСКИХ ДОКУМЕНТОВ НА ОСНОВЕ ОНТОЛОГИЧЕСКИХ МОДЕЛЕЙ

Специальность 05.13.12 – Системы автоматизации проектирования (промышленность)

Автореферат

диссертации на соискание ученой степени кандидата технических наук

Ульяновск – 2013

Работа выполнена на кафедре «Информационные системы» в Ульяновском государственном техническом университете.

Научный руководитель: кандидат технических наук, доцент
Наместников Алексей Михайлович

Официальные оппоненты: **Соснин Петр Иванович**
доктор технических наук, профессор, УлГТУ, кафедра «Вычислительная техника», заведующий кафедрой

Стецко Александр Алексеевич
доктор технических наук, ФНПЦ ОАО «НПО «Марс», главный технолог-начальник службы главного технолога

Ведущая организация: **Институт проблем управления сложными системами РАН, г. Самара**

Защита диссертации состоится «6» ноября 2013 г. в 15:00 часов на заседании диссертационного совета Д212.277.01 при Ульяновском государственном техническом университете по адресу: 432027, г. Ульяновск, ул. Северный Венец, 32 (ауд. 211, Главный корпус).

С диссертацией можно ознакомиться в библиотеке Ульяновского государственного технического университета.

Автореферат разослан «_____» _____ 2013 г.

Ученый секретарь диссертационного совета, доктор технических наук, профессор: Смирнов Виталий Иванович

Общая характеристика работы

Актуальность работы

Современная крупная проектная организация обладает значительным по объему электронным архивом конструкторской и технической документации, большая часть которой представлена в текстовом неструктурированном виде. Фактически, такой электронный архив текстовой документации содержит в себе опыт и знания большого количества высококвалифицированных специалистов, которые на протяжении многих лет занимались разработкой и проектированием сложных систем. В работах таких исследователей, как Норенков И.П., Тарасов В.Б., Collins H. и др. отмечается, что при увеличении объема электронного архива затрудняется анализ документов по заранее заданным реквизитам, а от лиц, занимающихся проектированием сложных технических систем, требуются навыки в области семантической обработки большого объема технической документации, а также глубоких знаний предметной области. В результате, довольно часто важный опыт предыдущих разработок, зафиксированный в электронных архивах, остается невостребованным и, как следствие, увеличивается время выполнения цикла опытно-конструкторских работ.

Повышение конкурентоспособности современных изделий, выпускаемых предприятиями невозможно без сокращения сроков выполнения научно-исследовательских и опытно-конструкторских работ. В работах Малюх В.Н. и Норенкова И.П. отмечается, что именно на начальных этапах разработки новых систем принципиально важным является использование опыта предыдущих проектов, зафиксированных в технических документах. Решение указанной проблемы может основываться на применении интеллектуальных методов и алгоритмов анализа технических документов проектной организации с целью построения навигационной структуры электронного архива. Представление содержимого архива в виде иерархии кластеров, содержащих технические документы, близкие по тематике в контексте используемых стандартов проектируемых систем, позволяет сократить пространство поиска и тем самым ускорить процедуры нахождения требуемых документов по их содержанию.

Для эффективного применения методов интеллектуального анализа текстовой конструкторской и технической документации не достаточно рассматривать отдельный документ как набор терминов из ограниченной предметной области. Количество используемых терминов в документах настолько вели-

ко, что применение известных методов кластеризации текстовых документов является затруднительным по причине их невысокой эффективности. К тому же качество распределения технических документов по кластерам часто оставляет желать лучшего. Учет специфики проектных знаний приводит к необходимости формирования онтологии электронного архива особой структуры, включающей в себя систему понятий предметной области, семантические отношения между ними и функции интерпретации. Поскольку любая проектируемая система изменяет свое состояние в соответствие с жизненным циклом, электронный архив должен обладать функциями адаптации к различным этапам (стадиям) жизненного цикла. Таким образом, электронный архив проектной организации должен обладать свойствами интеллектуальной системы. Ведущие исследователи в области онтологических систем, такие как Хорошевский В.Ф., Загорулько Ю.А., Гаврилова Т.А., Соловьев В.Д., Лукашевич Н.В., Добров Б.В., Ландэ Д.В., Смирнов С.В., Gruber T.R., Berners-Lee T., Uschold M. и другие отмечают актуальность исследований, основанных на онтологическом подходе. В трудах данных исследователей отмечается важность использования онтологического инжиниринга в процессе проектирования. В настоящее время не существует математических методов и алгоритмов, позволяющих структурировать содержание электронного архива текстовых документов, основываясь на их содержании с учетом специфики предметной области проектной организации в контексте жизненного цикла проектируемых систем. Следовательно, актуальным является разработка моделей, методов и алгоритмов построения навигационной структуры электронного архива технической документации на основе предметно-ориентированной кластеризации документов.

Цель диссертационной работы

Целью диссертации является разработка и реализация моделей и алгоритмов структуризации электронного архива технической документации, обеспечивающих снижение времени выполнения процессов информационной поддержки в принятии проектных решений.

Предмет исследования

Модели, методы и средства поддержки принятия проектных решений при формировании навигационной структуры электронного архива технической документации.

Объект исследования

Объектом исследования является электронный архив технической документации крупной проектной организации.

Задачи исследования

В соответствии с целью работы актуальными являются следующие задачи исследования:

- Провести сравнительный анализ существующих современных методов, алгоритмов и систем структуризации содержимого электронных архивов, применяемых в проектных организациях. Рассмотреть их ограничения в контексте жизненного цикла проектируемых систем.
- Рассмотреть возможность применения методов онтологического анализа для решения задач структуризации технической документации с целью построения навигационной структуры электронного архива.
- Разработать формальную модель технического документа электронного архива в пространстве признаков, определяемых прикладной онтологией с учетом жизненного цикла проектируемых систем.
- Разработать адаптируемые к стадиям проектирования методы онтологически-ориентированного индексирования и кластеризации технических документов с целью формирования навигационной структуры электронного архива.
- Разработать необходимые программные средства, позволяющие структурировать содержание электронного архива технических документов, провести вычислительные эксперименты, доказывающие их эффективность, внедрить полученные результаты в практику проектной организации.

Методы исследования

В диссертационной работе применяются методы онтологического анализа, теории графов, теории нечетких систем и мягких вычислений, кластерного анализа, объектно-ориентированного программирования.

Научная новизна

Научная новизна результатов исследования заключается в следующем:

1. Предложена новая формальная модель онтологии электронного архива технической документации, отличающаяся многоуровневой структурой и позволяющая описывать состояние содержимого электронного архива в контексте выполненных проектов, применяемых стандартов проектирования и жизненных циклов систем.
2. Разработана формальная онтологическая модель технического документа как ресурса электронного архива проектной организации, позволяющая решать задачу семантической структуризации содержимого электронного архива.

3. Предложен метод структуризации технических документов, отличающийся способом адаптации FCM-метода кластеризации и позволяющий формировать иерархическую навигационную структуру содержимого электронного архива проектной организации, учитывая жизненный цикл проектируемой системы.
4. Предложен алгоритм генетической оптимизации, позволяющий производить параметрическую настройку весов семантических отношений интеллектуального электронного архива технической документации на основе результатов экспертной классификации фрагмента содержимого электронного архива.

Практическая значимость работы

Созданная программная система онтологически-ориентированной структуризации текстовых технических документов электронного архива практически применяется в процессе проектирования автоматизированных систем и позволяет достичь улучшенных технико-экономических показателей объектов проектирования за счет сокращения времени выполнения опытно-конструкторских работ.

Разработанные модели и алгоритмы реализованы в форме программной системы и внедрены в деятельность ФНПЦ ОАО «НПО «Марс» (г. Ульяновск). Практическое использование результатов диссертационной работы подтверждено соответствующими документами о внедрении.

Основания для выполнения работы

Данная научная работа выполнялась в рамках тематического плана научных исследований Федерального агентства по образованию в 2009 и 2010 годах, была поддержана грантами РФФИ № 10-07-00064-а в 2010, 2011 и 2012 годах, № 12-01-97010-р_поволжье_а в 2012 году.

Достоверность результатов диссертационной работы

Достоверность научных положений, выводов и рекомендаций подтверждена результатами математического моделирования, результатами экспериментов и испытаний, а также результатами использования материалов диссертации в проектных подразделениях организации.

Основные положения, выносимые на защиту

1. Модель прикладной онтологии электронного архива технической документации является адекватной и эффективной для решения задачи построения навигационной структуры содержимого электронного архива.
2. Онтологическая модель технического документа является достаточной для представления содержания информационного ресурса электронно-

го архива в контексте жизненного цикла для задачи онтологически-ориентированой структуризации.

3. Адаптация метода структуризации технических документов электронного архива, заключающаяся в применении новой меры расстояния между документами и учитывающая состояние предметной области, представленное в онтологии, является эффективной.
4. Разработанный комплекс программ как подсистема электронного архива технической документации в полной мере реализует все описанные теоретические положения и позволяет сократить время поиска технических документов.

Апробация работы

Основные положения и результаты диссертации докладывались, обсуждались и получили одобрение на 43-й научно-технической конференции УлГТУ (г. Ульяновск, 2009 г.); всероссийской конференции с элементами научной школы для молодежи «Проведение научных исследований в области обработки, хранения, передачи и защиты информации» (г. Ульяновск, 2009 год); семинаре с международным участием «Интеллектуальный анализ временных рядов» (г. Ульяновск, 2010 год); 45-й научно-технической конференции УлГТУ (г. Ульяновск, 2011 г.); VI-ой международной научно-технической конференции «Интегрированные модели и мягкие вычисления в искусственном интеллекте» (г. Коломна, 2011 г.); всероссийской школе-семинаре «ИМАП-2011» (г. Ульяновск, 2011 г.); молодежной научно-технической конференции «Автоматизация процессов управления» (г. Ульяновск, 2011 г.); 46-й научно-технической конференции УлГТУ (г. Ульяновск, 2012 г.); 4-й всероссийской научно-технической конференции аспирантов, студентов и молодых ученых «ИВТ-2012» (г. Ульяновск, 2012 г.); 1-м международном симпозиуме «Гибридные и синергетические интеллектуальные системы: теория и практика» (г. Калининград, 2012 г.); 13-ой национальной конференции по искусственному интеллекту с международным участием «КИИ-2012» (г. Белгород, 2012 г.); всероссийской школе-семинаре «ИМАП-2012» (г. Ульяновск, 2012 г.); 47-й научно-технической конференции УлГТУ (г. Ульяновск, 2013 г.); III-й международной научно-технической конференции «OSTIS-2013» (г. Минск, 2013 г.); VII-ой международной научно-практической конференции «Интегрированные модели и мягкие вычисления в искусственном интеллекте» (г. Коломна, 2013 г.).

Научные публикации

По результатам работы было опубликовано 18 статей, из которых 4 в

журналах из перечня ВАК, и 2 тезиса докладов. Получены свидетельства о государственной регистрации программ для ЭВМ № 2012617586 (2012 г.), 2012617589 (2012 г.).

Структура и объем диссертации

Диссертация состоит из введения, четырех глав, заключения, списка использованной литературы и приложений. Основное содержание работы изложено на 168 страницах, включая 35 рисунков и 11 таблиц. Список использованных источников состоит из 131 наименования.

Личный вклад

Все результаты, составляющие содержание диссертации, получены автором самостоятельно.

Краткое содержание работы

Во введении рассмотрена актуальность работы, определена ее цель и задачи, сформированы положения, выносимые на защиту, их научная новизна и практическая ценность. Представлены основания для выполнения работы, ее апробация и структура.

В **первой главе** содержится анализ методов индексирования и кластеризации текстовых документов. Рассмотрена модель документа «множество слов» и модели, учитывающие взаимное положения слов. Приведена классификация существующих методов индексирования текстовых документов. Представлена классификация существующих методов индексирования и кластеризации текстовых документов в задачах интеллектуального анализа. Рассмотрено применение методов мягких вычислений в задачах индексирования и кластеризации. Описаны особенности технических документов как ресурсов электронного архива. Приведен анализ существующих программных систем управления электронными архивами технических документов.

Во **второй главе** описываются онтологические модели и методы структуризации электронного архива на основе кластеризации.

При построении модели предметной области в виде прикладной онтологии для решения задач анализа технической документации 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 сокра-

щения времени поиска необходимых документов. Фактически, для отдельно взятого информационного ресурса из электронного архива технической документации онтология задает новую систему координат, в которой кластер документов может рассматриваться как группа связанных между собой по смыслу технических документов.

Предметная область проектирования сложных систем (к которым можно отнести класс автоматизированных систем (АС)) накладывает определенные требования к структуре прикладной онтологии. Жесткая привязка к применяемым на различных стадиях проектирования стандартам влечет за собой необходимость в формировании онтологии, состоящей из множества уровней (рисунок 1).

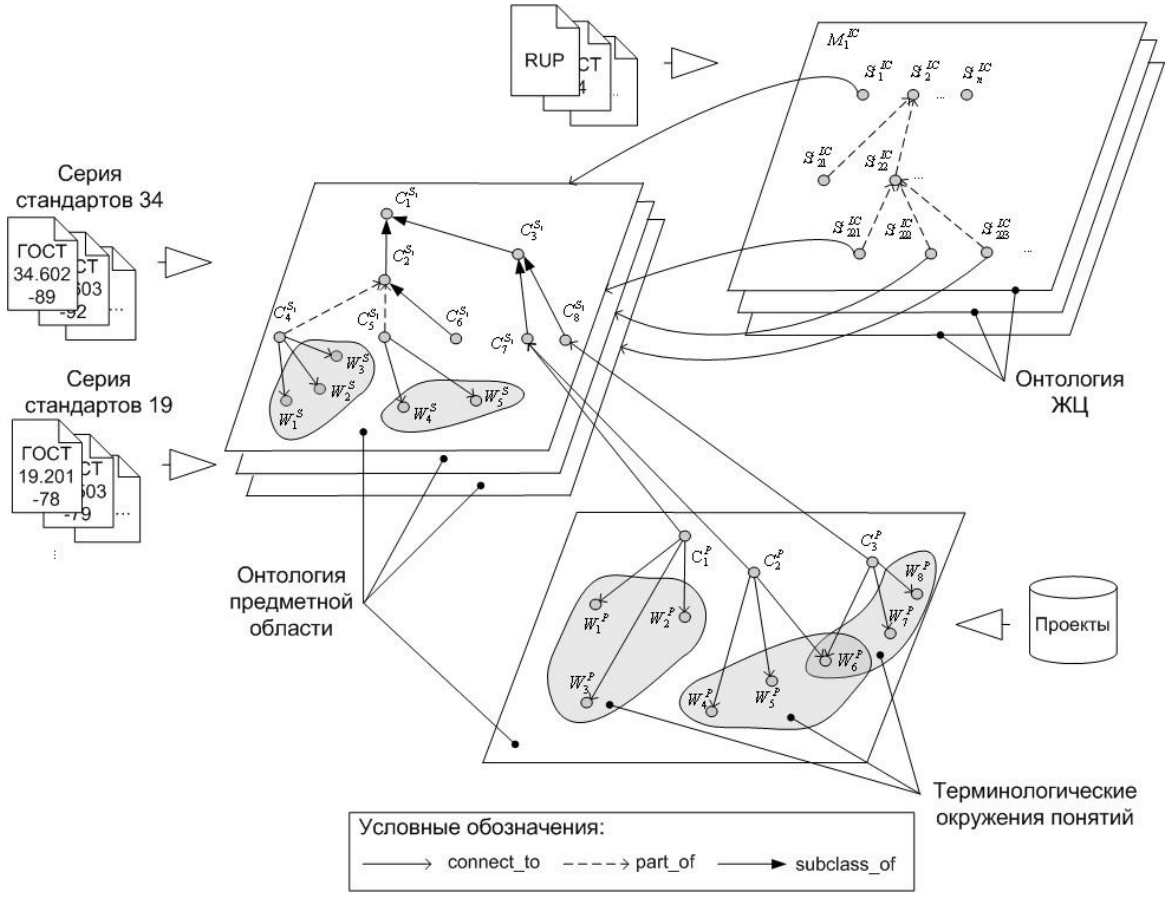

Рис. 1. Структура прикладной онтологии электронного архива

Формально онтология электронного архива технической документации проектной организации состоит из двух онтологических компонент и записывается как кортеж вида:

$$O = \langle O^D, O^{LC}, R_A \rangle, \tag{1}$$

где O^D – компонент онтологии предметной области; O^{LC} – онтология ЖЦ

проектируемых систем; R_A – отношение однонаправленной ассоциации между компонентами онтологии.

Онтологию предметной области запишем в виде кортежа:

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

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

$$R^D = \{R^D_G, R^D_C, R^D_A\},$$

где R^D_G – антисимметричное, транзитивное, нерефлексивное бинарное отношение обобщения («subclass_of»); R^D_C – бинарное транзитивное отношение композиции («part_of»); R^D_A – бинарное отношение однонаправленной ассоциации.

Множество понятий C записывается следующим образом:

$$C = \left(C^{S_1} \cup C^{S_2} \cup \ldots \cup C^{S_k}\right) \cup C^P,$$

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

Множество интерпретирующих функций представлено в виде:

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

где $F^D_{WC} : \{W\} \to \{C\}$ – функция, сопоставляющая набору терминов подмножество понятий предметной области и задается алгоритмически; $F^D_{C^P C^S} : \{C^P\} \to \{C^S\}$ – функция интерпретации подмножества понятий на проектном уровне онтологии, позволяющая осуществить переход на уровень понятий (концептов), определенных в стандартах.

Онтология ЖЦ как компонента кортежа (1) записывается следующим образом:

$$O^{LC} = \langle M^{LC}, St^{LC}, R^{LC} \rangle,$$

где M^{LC} – множество моделей ЖЦ проектируемых изделий; St^{LC} – множество стадий (этапов) ЖЦ. Отношение R^{LC} имеет вид «часть-целое (part_of)» и позволяет декомпозировать стадии жизненного цикла проектируемой системы в онтологии на этапы и т. д.

В основе онтологического индексирования ТД лежит следующая функция:

$$F_{oV} : ch_j^d \to oV_j^d, \tag{2}$$

где ch_j^d – j-ый раздел технического документа d, oV_j^d – онтологическое представление j-го раздела технического документа d.

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

Семантический коэффициент отношения (семантическое расстояние) между понятием и термином определяется следующим образом:

$$S\left(c_i^{P(S)}, w_j\right) = \frac{\sum_{occur(c_i^{P(S)}, w_j)} \frac{1}{exp(sentence \cdot (paragraph+1))}}{num(occur(c_i^{P(S)}, w_j))} \cdot \\ \cdot \frac{num(paragraph - cooccur(c_i^{P(S)}, w_j))}{num(totalparagraph)},$$

где $c_i^{P(S)}, w_j$ – i-е понятие уровня проектов (стандартов) и j-й термин соответственно; $sentence$ – расстояние, выраженное в количестве предложений между понятием и термином; $paragraph$ – расстояние, выраженное в количестве абзацев между понятием и термином; $num(occur(c_i^{P(S)}, w_j))$ – количество совпадений $c_i^{P(S)}$ и w_j; $num(paragraph - cooccur(c_i^{P(S)}, w_j))$ – количество абзацев, где существует совместная встречаемость $c_i^{P(S)}$ и w_j; $num(totalparagraph)$ – число абзацев в документе.

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

$$\{(w_{1j}^d, f_1^j), (w_{2j}^d, f_2^j), \ldots, (w_{ij}^d, f_i^j), \ldots, (w_{l_j j}^d, f_{l_j}^j)\},$$

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:

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

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 электронного архива – 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.*

Вычисление степеней выраженности понятий онтологии для каждого раздела ТД производиться с применением аппарата нечетких соответствий описанного в работах Берштейна и Боженюка. *Нечетким соответствием* между множествами W (терминологические окружения понятий на уровне проектов (стандартов)) и $C^{P(S)}$ (множество понятий прикладной онтологии на уровне проектов (стандартов)) называется тройка множеств $\tilde{\Gamma} = \left(W, C^{P(S)}, \tilde{O}\right)$, в которой W и $C^{P(S)}$ – четкие множества, а $\tilde{O}$ – нечеткое множество в $W \times C^{P(S)}$. Множество W есть область отправления, множество $C^{P(S)}$ – область прибытия, а $\tilde{O}$ – нечеткий график нечеткого соответствия.

Назовем носителем нечеткого соответствия $\tilde{\Gamma} = \left(W, C^{P(S)}, \tilde{O}\right)$ четкое соответствие $\Gamma = \left(W, C^{P(S)}, O\right)$, у которого график O является носителем нечеткого графика $\tilde{O}$. В контексте онтологии график O определяет экземпляры однонаправленных ассоциаций R_A^D между понятиями проектов и терминами в онтологии.

Образом множества $\tilde{W^d}$ (множество терминов ТД d) при соответствии $\tilde{\Gamma}$ есть нечеткое множество $\tilde{\Gamma}(\tilde{W^d})$ в $C^{P(S)}$, определяемое выражением:

$$\tilde{\Gamma}(\tilde{W^d}) = \{\langle \mu_{\Gamma(W^d)}(c^{P(S)}), c^{P(S)} \rangle \mid c^{P(S)} \in C^{P(S)}\}, \quad (3)$$

где $\mu_{\Gamma(W^d)}(c) \vee_{w^d \in W^d} \left(\mu_{\Gamma(W^d)}(w^d) \wedge \mu_O \langle w, c^{P(S)} \rangle\right)$.

Для нахождения значения доминирования концептов используем метод сравнения терминологического окружения каждого понятия в онтологии предметной области уровня проектов с анализируемым текстом. Минимальным фрагментом анализируемого текста является отдельное предложение, а

максимальным – текстовый документ в целом, так как в различных частях (фрагментах) документа делается акцент на разных понятиях предметной области.

Степень выраженности $\mu_{S_p^d}(c^{P(S)})$ понятий $c^{P(S)} \in C^{P(S)}$ в p-м фрагменте ТД d вычисляется как образ множества терминов по формуле (3).

Алгоритм вычисления степени доминирования понятия в текстовом фрагменте состоит из следующих шагов:

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

$$\hat{\mu}_{S_p^d}\left(c^{P(S)}\right) = \max_{c}\left(\mu_{S_p^d}\left(c^{P(S)}\right)\right).$$

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

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

где $c_i^{P(S)} \in c^{P(S)} - c_k^{P(S)}$; $c_k^{P(S)} = \arg\max_{c^{P(S)}}\left(\mu_{S_p^d}\left(c^{P(S)}\right)\right)$; n – количество концептов с ненулевой степенью выраженности для текстового фрагмента S_p^d.

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

$$\Delta_{S_p^d}\left(c^{P(S)}\right) = \hat{\mu}_{S_p^d}\left(c^{P(S)}\right) - \tilde{\mu}_{S_p^d}\left(c^{P(S)}\right). \quad (4)$$

Выражение (4) фактически определяет качество выделения текстового фрагмента в ТД с целью ограничения в тексте определенного понятия предметной области, которое зафиксировано в онтологии интеллектуального электронного архива.

В основе метода определения доминирующего понятия в текстовом фрагменте технического документа лежит модифицированный генетический алгоритм, целью которого является нахождение такой последовательности текстовых фрагментов, которая соответствует минимальному значению целевой функции

$$F\left(S^d\right) = \frac{1}{s}\sum_{p}\left(1 - \Delta_{S_p^d}\left(c^{P(S)}\right)\right) \to \min,$$

$p = \overline{1,s}$, где s – количество текстовых фрагментов; $s = \overline{1,m}$, где m – количество предложений в индексируемом документе.

Проведенные эксперименты с выделенными фрагментами ТД на основе генетической оптимизации показали, что в среднем около 30% понятий в сумме дают 70% от общей степени выраженности всех понятий фрагмента ТД. Учитывая данный факт, первоначальные наборы понятий $\hat{C}_j^P$ и $\hat{C}_j^S$ j-го раздела документа расширяются наиболее значимыми понятиями каждого фрагмента.

Заключительным шагом в формировании онтологического представления ТД является применение интерпретирующей функции $F^D_{C^P C^S} : \{C^P\} \to \{C^S\}$, которая позволяет уточнить набор понятий уровня стандартов, опираясь на найденное подмножество понятий в ТД уровня проектов онтологии с учетом моделей жизненного цикла.

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

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

Формальную меру расстояния между документами рассмотрим 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.

Редакционное расстояние между иерархиями определяется на основе вычисления стоимости редакционной операции. Для отношения обобщения редакционную операцию запишем 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 серии стандартов. Фактически, в задаче структуризации ТД под редакционной операцией понимается вес соответствующего отношения, принимающий значение в диапазоне от 0 до 1 и имеющий различные значения в рамках каждой серии 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 серии стандартов; s – номер добавляемого отношения обобщения; l – номер добавляемого отношения «часть-целое».

Коэффициент нормализации T_{oV} есть сумма весов всех семантических отношений обобщенной иерархии. Мера расстояния между онтологическим представлениями ТД определяется с помощью следующего выражения:

$$\left\| oV^{d_1} - oV^{d_2} \right\| = \frac{\tau^*_{oV}}{T_{oV}}$$

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. Данная задача решается с помощью модифицированного генетического алгоритма, задача которого состоит в нахождении такого множества коэффициентов семантических отношений при которых качество структуризации, определяемое выражением:

$$F^* = \frac{\max\left(\bar{K}_+ + \bar{K}_-, \hat{K}_+ + \hat{K}_-\right)}{N} \to \min \tag{5}$$

было бы наилучшим. В выражении (5) $\bar{K}_-$ и $\hat{K}_-$ – множества отсутствующих документов соответственно в первом и во втором кластерах, $\bar{K}_+$ и $\hat{K}_+$ – множества лишних документов соответственно в первом и во втором кластерах, N – количество документов. Поскольку указанные отношения применяются между понятиями в онтологии для различных серий стандартов, их оптимальные значения для каждой такой серии (в рамках своей иерархии понятий) в общем случае будут различными.

В **третьей главе** представлено описание разработанной программной системы интеллектуального электронного архива технических документов.

Интеллектуальный электронный архив – программная система, предназначенная для систематизации и автоматизации работы со множеством электронных ТД, учитывающая состояние онтологии предметной области. Структура ИЭА представлена на рисунке 2.

Основные функции ИЭА:

1. Онтологически-ориентированное индексирование технических документов. Формирование онтологических представлений ТД.
2. Хранение онтологии предметной области и онтологических представлений ТД.
3. Онтологически-ориентированная структуризация технических докумен-

тов с учетом состояния онтологии предметной области.

4. Построение навигационной структуры электронного архива, учитывающей использование моделей ЖЦ, с целью сокращения пространства поиска ТД.
5. Поиск в навигационной структуре документов близких в семантическом смысле к документу указанному пользователем.

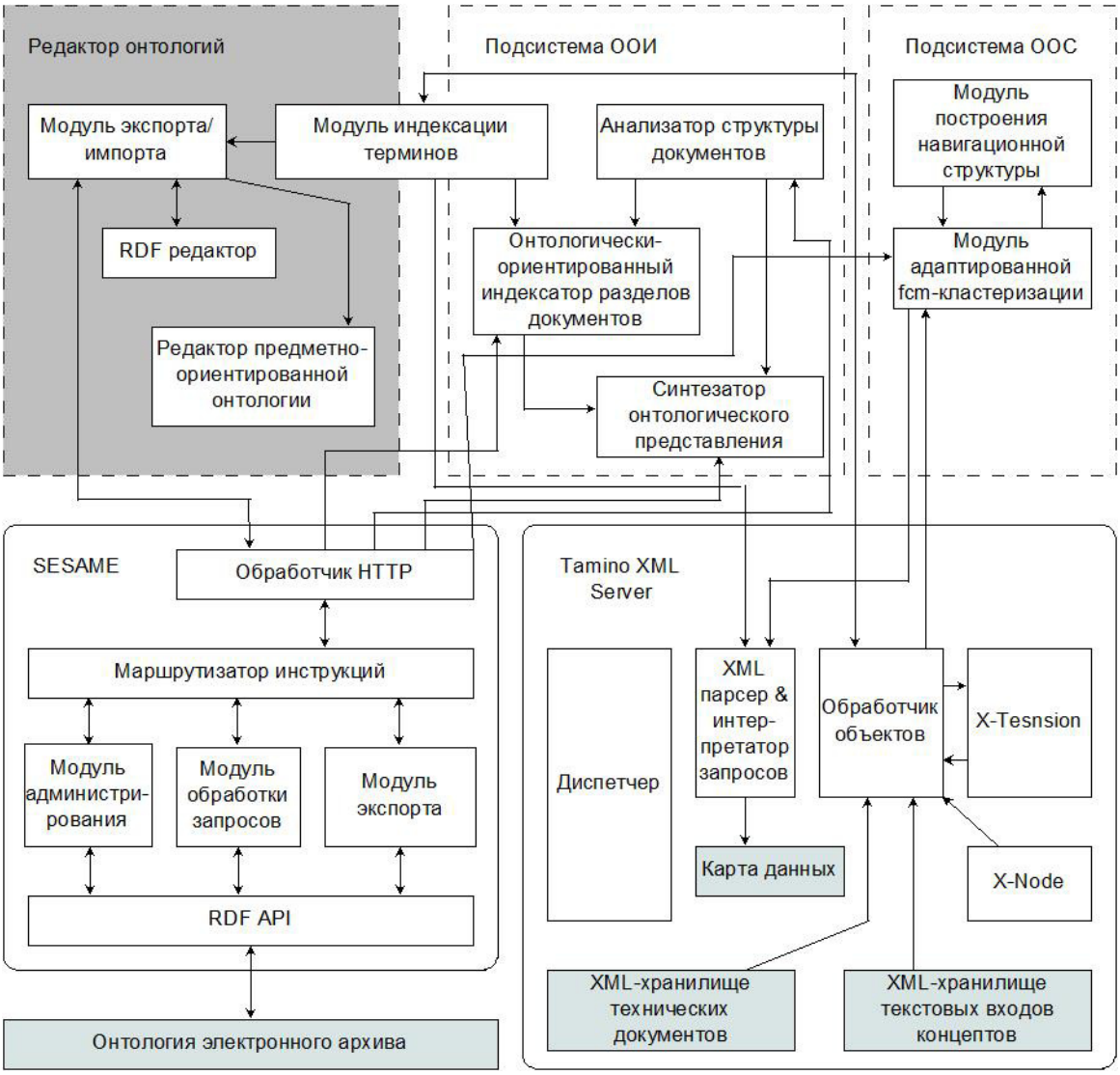

Рис. 2. Структура интеллектуального электронного архива

Разработанная система включена в состав электронного архива ФНПЦ ОАО «НПО «Марс». ИЭА использует хранилище электронного архива НПО «Марс» для получения входных данных, которые в дальнейшем используются в процессах индексирования и структуризации.

В качестве хранилища исходных и проиндексированных документов ИЭА выступает XML-ориентированная СУБД Tamino. Хранилищем онтологий в ИЭА является сервер онтологий Sesame. Программная система реализована

на языке программирования Java в среде разработки Eclipse.

В **четвертой главе** проводится анализ адекватности разработанных моделей и методов на основе вычислительных экспериментов.

Для осуществления анализа временных затрат на процесс индексирования множества ТД был проведен ряд экспериментов. Учитывалось общее время индексирования, а также временные затраты на процесс построения онтологического и традиционного представлений ТД.

Исходя из результатов экспериментов сделаем вывод – основное влияние на длительность процесса индексирования документов оказывает формирование онтологических представлений ТД, а именно операция определения доминирующих понятий в текстовых фрагментах ТД (≈90% от общего времени процесса индексирования).

Для осуществления анализа результатов работы подсистем онтологически-ориентированного индексирования и структуризации на множестве документов электронного архива ФНПЦ ОАО «НПО «Марс» была использована онтология предметной области, которая имеет в своем составе 300 понятий: 219 понятий на уровне стандартов и 81 понятие на уровне проектов, а также 10078 уникальных термов на уровне проектов. Экспертом ФНПЦ ОАО «НПО «Марс» была подготовлена выборка состоящая из 5017 ТД и сгруппированная по разным основаниям классификации. На первом шаге эксперимента был построен индекс, содержащий в своем составе онтологические и традиционные представления ТД. На следующем шаге полученный индекс был подвергнут различным вариантам структуризации с последующим расчетом качества структуризации:

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

Лучшие значения оценочной функции для онтологических представлений с учетом ЖЦ получены при структуризации выборки ТД по тематике работ, так как данный вариант структуризации наиболее близко соответствует основной задаче ИЭА – структуризации ЭА ТД по содержанию отдельных документов. Для структуризации по виду документа лучшие результаты пока-

зала система Oracle Text (рисунок 3). По результатам экспериментов качество структуризации онтологических представлений ТД с учетом ЖЦ примерно на 40% лучше по сравнению с результатами системы Oracle Text.

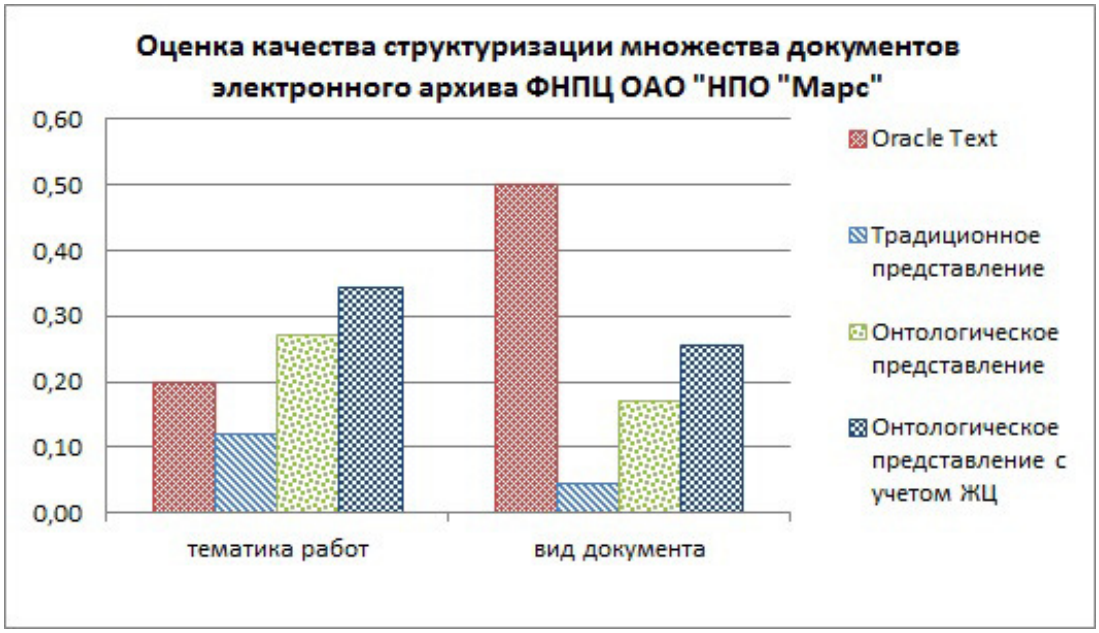

Рис. 3. Оценка качества структуризации множества документов электронного архива ФНПЦ ОАО «НПО «Марс»

Из результатов экспериментов видно, что при увеличении количества документов и кластеров время структуризации традиционных представлений ТД значительно возрастает. Разница в скорости структуризации между традиционными и онтологическими представлениями ТД, является значительной и зависит от размера выборки ТД, количества понятий уровня стандартов онтологии предметной области, количества терминов в составе редуцированного множества терминов всей выборки ТД и количества кластеров (рисунок 4).

Для проведения экспериментов по оценке снижения времени выполнения проектных процедур с использованием навигационной структуры электронного архива в рамках опытно-конструкторской работы ФНПЦ ОАО «НПО «Марс» был использован существующий проект. В выборке, состоящей из 5017 ТД, был произведен поиск ТД похожего по содержанию на техническое задание из состава рассматриваемого проекта, с последующим подсчетом суммарного времени поиска документа (рисунок 5). Использование навигационной структуры электронного архива, вместо традиционных методов разбиения множества ТД на определенное число кластеров, сократило суммарное время поиска документов в среднем на 13%.

Рис. 4. Временные затраты на процесс структуризации множества документов электронного архива ФНПЦ ОАО «НПО «Марс»

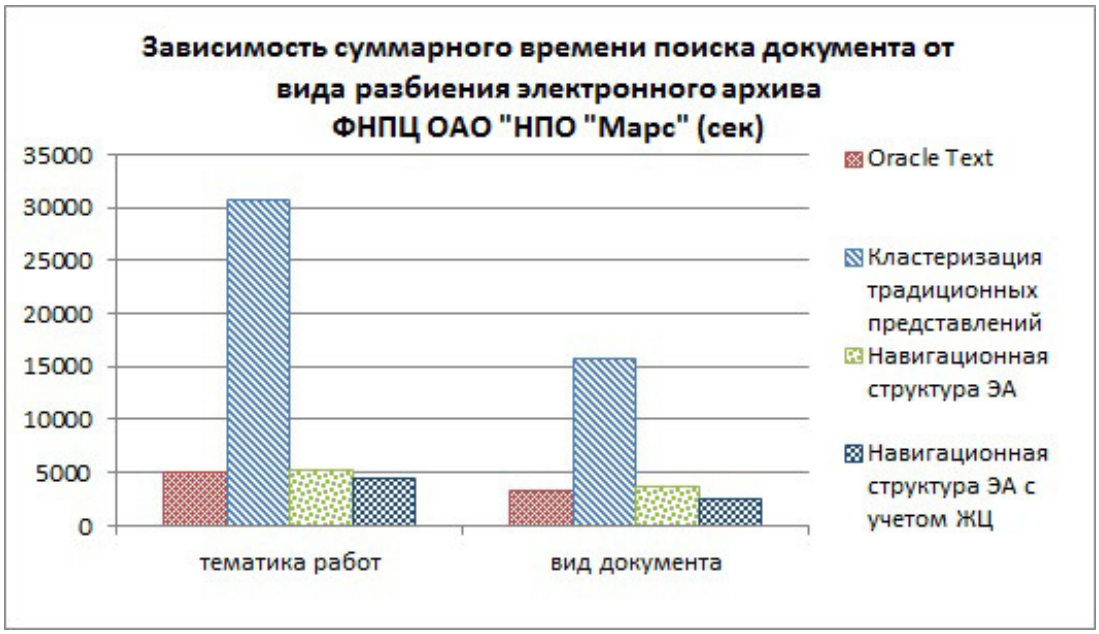

Рис. 5. Зависимость суммарного времени поиска документа от вида разбиения множества ТД

В заключении приведены основные результаты исследований, полученные в диссертационной работе:

1. Предложена новая формальная модель онтологии электронного архива технической документации, отличающаяся многоуровневой структурой и позволяющая описывать состояние содержимого электронного архива в контексте выполненных проектов, применяемых стандартов проектирования и жизненных циклов систем.

2. Разработана формальная онтологическая модель технического проектного документа как ресурса электронного архива проектной организации, позволяющая решать задачу семантической структуризации содержимого электронного архива.
3. Предложен метод структуризации технических документов, отличающийся способом адаптации FCM-метода кластеризации и позволяющий формировать иерархическую навигационную структуру содержимого электронного архива проектной организации, учитывая жизненный цикл проектируемой системы.
4. Предложен алгоритм генетической оптимизации, позволяющий производить параметрическую настройку весов семантических отношений интеллектуального электронного архива технической документации на основе результатов экспертной классификации фрагмента содержимого электронного архива.
5. Разработана программная система интеллектуального анализа электронного архива технический документов.
6. Проведены вычислительные эксперименты:
   - оценивалось качество структуризации множества документов электронного архива ФНПЦ ОАО «НПО «Марс». По данным экспериментов результаты структуризации онтологических представлений ТД с учетом ЖЦ примерно на 40% лучше по сравнению с результатами системы Oracle Text.;
   - выполнялся анализ временных затрат на процессы индексирования и структуризации электронного архива технической документации. Общие временные затраты на процессы индексирования и структуризации онтологических представлений технических документов, в среднем, на 7% меньше суммарных временных затрат на процессы индексирования и структуризации традиционных представлений ТД. Онтологический подход к индексированию и структуризации множества ТД дает возможность произвести структуризацию электронного архива за меньшее время, при этом основные временные затраты будут отнесены к процессу индексирования документов;
   - оценивалось снижение времени выполнения проектных процедур с использованием навигационной структуры электронного архива. Использование навигационной структуры электронного архива с учетом ЖЦ, вместо традиционных методов разбиения множества ТД на определенное число кластеров, сократило суммарное время поиска

документов в среднем на 13%.

Таким образом, использование онтологически-ориентированных методов индексирования и структуризации, учитывающих применяемые в процессе проектирования серии стандартов и модели жизненного цикла, предпочтительно при необходимости более качественного построения навигационной структуры электронного архива.

Навигационная структура ЭА строится на основе данных о содержимом документов, тем самым необходимо использовать данные, содержащиеся в существующей системе управления ЭА ФНПЦ ОАО «НПО «Марс», для улучшения качества поиска необходимых документов. Использование дополнительной информации о содержимом ЭА, при построении навигационной структуры данного архива, позволит сузить границы поиска и повысить качество структуризации. Необходимо рассматривать предлагаемую в работе онтологическую систему структуризации как подсистему ЭА ФНПЦ ОАО «НПО «Марс».

В приложении А представлен акт об использовании результатов диссертации. **В приложении Б** представлены исходные коды основных алгоритмов программной системы. **В приложении В** представлены RDF-схема и фрагмент онтологии предметной области. **В приложении Г** представлены фрагменты предобработанного ТД и онтологического представления ТД.

Список публикаций

Публикации в изданиях, рекомендованных ВАК:

1. Филиппов А.А., Наместников А.М. Концептуальная индексация проектных документов // Автоматизация процессов управления №2(20). – 2010. С. 34-39.
2. Филиппов А.А., Наместников А.М. Реализация системы кластеризации концептуальных индексов проектных документов // Автоматизация процессов управления №3(25). – 2011. С. 46-50.
3. Филиппов А.А., Наместников А.М., Субхангулов Р.А. Разработка инструментария для интеллектуального анализа технической документации // Известия Самарского научного центра Российской академии наук № 4, Том 13. – 2011. С. 984-990.
4. Филиппов А.А. Формирование навигационной структуры электронного архива технических документов на основе онтологических моделей // Автоматизация процессов управления, № 3(33), 2013. С. 61-68.

Публикации в прочих изданиях:

1. Филиппов А.А., Наместников А.М. XML репозиторий проектных документов // Всероссийская конференция с элементами научной школы для молодежи «Проведение научных исследований в области обработки, хранения, передачи и защиты информации», 1-5 декабря 2009 г. Россия, Ульяновск: сборник научных трудов. В 4 т. Т. 4. – Ульяновск : УлГТУ, 2009, С. 254-256.
2. Филиппов А.А., Наместников А.М. Концептуальная индексация проектных документов // Интеллектуальный анализ временных рядов: сборник научных трудов семинара с международным участием «Интеллектуальный анализ временных рядов» по результатам НИР, поддержанной ФЦП, проект № 02.740.11.5021, г. Ульяновск, 15 июня 2010 г. – Ульяновск : УлГТУ, 2010. С. 69-77.
3. Филиппов А.А., Наместников А.М. Нечеткая кластеризация концептуальных индексов проектных документов // Интегрированные модели и мягкие вычисления в искусственном интеллекте. Сборник научных трудов 6-й Международной научно-технической конференции (Коломна, 16-19 мая 2011 г.). В 2-х томах. Т2. - М. : Физматлит, 2011. С. 958-968.
4. Филиппов А.А., Наместников А.М. Метод онтологической кластеризации документов в интеллектуальном проектном репозитории // Гибридные и синергетические интеллектуальные системы: теория и практика : материалы 1-го международного симпозиума / под ред. проф. А.В. Колесникова. – Калининград : Изд-во БФУ им. И. Канта, 2012. С.205-213.
5. Филиппов А.А., Наместников А.М., Субхангулов Р.А. Система кластеризации и полнотекстового поиска проектных документов на основе прикладной онтологии // Тринадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2012 (16-20 октября 2012 г., г. Белгород, Россия): Труды конференции. Т.2.– Белгород : Изд-во БГТУ, 2012. С. 104-111.
6. Филиппов А.А., Наместников А.М. Метод генетической оптимизации онтологических представлений проектных документов в задаче индексирования // Тринадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2012 (16-20 октября 2012 г., г. Белгород, Россия): Труды конференции. Т.4. – Белгород : Изд-во БГТУ, 2012. С. 84-91.
7. Филиппов А.А. Индексирование и кластеризация проектных документов на основе графовой модели онтологии // Информатика, моделирование,

автоматизация проектирования: сборник научных трудов / под. ред. Н. Н. Войта. – Ульяновск : УлГТУ, 2011. С. 367-372.

8. Филиппов А.А. Нечеткая кластеризация концептуальных индексов проектных документов // Автоматизация процессов управления: сборник докладов Молодежной научно-технической конференции, Ульяновск, 13-14 декабря 2011 г. / под общ. ред. А.А. Емельянова. – Ульяновск : ФНПЦ ОАО «НПО «Марс», 2011. С. 116-122.
9. Филиппов А.А. Онтологически-ориентированное индексирование проектных документов. XML-сервер Tamino как ядро интеллектуального проектного репозитория // Вузовская наука в современных условиях : сборник материалов 46-й научно-технической конференции (23-28 января 2013 года). В 3 ч. Ч.2. – Ульяновск : УлГТУ, 2012. С. 154-157.
10. Филиппов А.А. Онтологически-ориентированная кластеризация проектных документов // Информатика и вычислительная техника: сборник научных трудов 4-й Всероссийской научно-технической конференции аспирантов, студентов и молодых ученых ИВТ-2012. В 2 т. / под ред. Н. Н. Войта. – Ульяновск : УлГТУ, 2012. С.323-331.
11. Филиппов А.А. Реализация онтологически-ориентированных подсистем индексирования и кластеризации проектных документов // Информатика, моделирование, автоматизация проектирования: сборник научных трудов / под ред. Н. Н. Войта. – Ульяновск : УлГТУ, 2012. С.389-397.
12. Филиппов А.А. Анализ временной сложности онтологически-ориентированных методов индексирования и кластеризации проектных документов // Вузовская наука в современных условиях : сборник материалов 47-й научно-технической конференции (28 января – 2 февраля 2013 года). В 3 ч. Ч.2. – Ульяновск : УлГТУ, 2013. С. 174-177.
13. Филиппов А.А., Наместников А.М., Субхангулов Р.А. Онтологически-ориентированная система кластеризации и полнотекстового поиска проектных документов // Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2013): материалы III Междунар. научн.техн. конф. (Минск, 21-23 февраля 2013г.) / редкол. : В. В. Голенков (отв. ред.) [и др.]. – Минск : БГУИР, 2013. С. 219-224.
14. 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, 20-22

мая 2013 г.). В 3-х томах. Т3. – М. : Физматлит, 2013. С. 1278-1289.

Свидетельства:

1. Свидетельство о государственной регистрации программы для ЭВМ №2012617586. Онтологически-ориентированный индексатор проектных документов / А.А.Филиппов, А.М.Наместников. - 2012 г. — М.: Роспатент, 2012.
2. Свидетельство о государственной регистрации программы для ЭВМ №2012617589. Онтологически-ориентированный кластеризатор проектных документов / А.А.Филиппов, А.М.Наместников. - 2012 г. — М.: Роспатент, 2012.

Филиппов Алексей Александрович

Формирование навигационной структуры электронного архива технических документов на основе онтологических моделей

Автореферат

Подписано в печать __.__.2013. Формат 60x84/16.
Бумага писчая. Усл. печ. л. 1,17. Уч.-изд. л. 1,00.
Тираж 100 экз. Заказ _____

Типография УлГТУ, 432027, г.Ульяновск, ул. Сев. Венец, д. 32.