На правах рукописи

ХАРИН Максим Алексеевич

РАЗРАБОТКА МОДЕЛЕЙ И МЕТОДОВ ВЕРИФИКАЦИИ И АНАЛИЗА ДОКУМЕНТОВ В ЭЛЕКТРОННОМ АРХИВЕ ЭНЕРГЕТИЧЕСКИХ ОБЪЕКТОВ

Специальность

05.13.12 – Системы автоматизации проектирования

(электротехника, энергетика)

Автореферат

диссертации на соискание ученой степени

кандидата технических наук

Иваново 2013

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего профессионального образования «Ивановский государственный энергетический университет имени В.И. Ленина»

Научный руководитель — **Кроль Татьяна Яковлевна,**
кандидат технических наук

Официальные оппоненты — **Шведенко Владимир Николаевич,**
доктор технических наук, профессор,
ФГБОУ ВПО «Костромской государственный технологический университет», заведующий кафедрой «Информационные технологии»

Ильичёв Николай Борисович,
кандидат технических наук, доцент,
ЗАО «СиСофт Иваново», главный специалист

Ведущая организация — ОАО «Зарубежэнергопроект», г. Иваново

Защита состоится 25 июня 2013 года в 14 часов на заседании диссертационного совета Д 212.064.02 при Ивановском государственном энергетическом университете по адресу: 153003, г. Иваново, ул. Рабфаковская, 34, корпус «Б», ауд. 301.

Отзывы на автореферат в двух экземплярах, заверенные печатью, просим присылать по адресу: 153003, г. Иваново, ул. Рабфаковская, 34, Учёный совет ИГЭУ. Тел.: (4932) 38-57-12, 26-98-61, факс: (4932) 38-57-01, e-mail: uch_sovet@ispu.ru

С диссертацией можно ознакомиться в библиотеке ФГБОУ ВПО «Ивановский государственный энергетический университет имени В.И. Ленина», автореферат размещён на сайте www.ispu.ru.

Автореферат разослан « 24 » мая 2013 г.

Учёный секретарь
диссертационного совета,
доктор технических наук, профессор — Тютиков Владимир Валентинович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Для предприятий энергетической отрасли важной задачей является создание единого информационного пространства путём перехода на безбумажный документооборот. При этом необходимо учитывать следующие особенности работы предприятий, занимающихся проектированием и монтажом энергетических объектов:

- территориальная распределённость (организации, занимающиеся проектированием, строительством, монтажом и эксплуатацией энергетических объектов всегда находятся на существенном расстоянии друг от друга);
- большое количество документации, которая должна поддерживаться в актуальном состоянии и быть доступной в сжатые сроки для оперативного принятия решений, особенно в аварийных ситуациях.

Эти особенности объективно требуют информационной интеграции процессов проектирования, монтажа и эксплуатации энергетических объектов. Ускорение информационных потоков необходимо для повышения эффективности и надёжности работы энергетических предприятий.

В организациях, занимающихся проектированием, строительством и реконструкцией энергетических объектов, обычно имеется архив технической документации порядка сотен тысяч документов. При этом организация может иметь распределённую структуру с филиалами в разных городах РФ, объекты строительства также могут быть удалёнными: от Нягани до Краснодарского края. В ходе строительных и особенно электромонтажных работ часто происходит изменение и дополнение проектной документации.

Поиск необходимой документации в "бумажном" архиве и её доставка (даже путём сканирования и электронной пересылки) в удалённые точки занимают много времени. Поэтому актуальна задача создания *системы электронного архива*: системы структурированного хранения проектной документации в электронном виде, обеспечивающей надёжность хранения, конфиденциальность и разграничение прав доступа, отслеживание истории использования документа, быстрый и удобный поиск, а также предоставляющей доступ к документации из любого места в любое время.

Особенности технической документации, которые необходимо сохранить при переходе к электронному архиву:

- соответствие ГОСТам серии СПДС;
- насыщенность символьными наименованиями (чертежи, объекты, устройства, материалы).

Рассмотрим комплект документации по некоторому энергетическому объекту (например, «Подстанция Мещанская»). Вся рабочая документация делится:

- по стадиям: проектная «ПД» и рабочая «РД»;

- по пусковым комплексам (ПК); 1 ПК – строительство подстанции; 2 ПК – строительство заходов КЛ 220 кВ на подстанцию; 3, 4 ПК – строительство дополнительных кабельных линий.
- по буквенной марке номера рабочей документации (в соответствии с требованиями ГОСТа), например: АЭВ, АЭП, РЗ, СС и др.

Практически каждый том рабочей документации содержит символьные наименования, например, спецификация оборудования содержит строки вида «Вентилятор 1U 48VDC для FOX515» или «Оптический лазер S1.1 LC SFP, 1310 нм». Также, в соответствии с ГОСТ 2.104-68*, в основной надписи на чертеже указываются фамилии и подписи лиц, выполняющих разработку, проверку, технологический контроль, нормоконтроль и утверждение документа.

В архиве необходимо предоставить возможность группировки документов в соответствии с приведённой классификацией, чтобы обеспечить доступ к единственной актуальной версии документа всем заинтересованным лицам: руководителю объекта (подстанции); проектировщикам из различных организаций, выполняющих проект; генподрядчику строительно-монтажных работ; начальнику монтажной бригады на объекте и т.д.

В унаследованной системе архива документы хранятся в бумажном или электронном виде на компакт-дисках в помещении в центральном офисе организации. При переходе на новую систему электронного архива бумажные документы должны быть отсканированы. Для обеспечения структурирования документов и их быстрого поиска в архиве должны храниться не только электронные образы (изображения) документов, но и их карточки (наборы атрибутов). Для формирования атрибутов документов применено распознавание отсканированных документов при помощи специализированных программных средств. При этом точность распознавания не всегда является стопроцентной, необходима *верификация*. Верификация – это процесс проверки правильности распознанных документов. Она производится человеком и заключается в сверке распознанного текста с графическим образом документа. Однако при большом потоке документов в силу монотонности работы увеличивается число ошибок верификации, что недопустимо для технической документации. В связи с этим актуальной является задача автоматизации процесса верификации для повышения скорости и уменьшения числа ошибок.

Так как некоторые атрибуты документов распознаются лучше, другие хуже, а процесс верификации является последовательным, для ускорения следует использовать зависимость значений атрибутов друг от друга. Наибольшую скорость в данном случае обеспечивают продукционные правила, так как они используют простую модель «ключ-значение», что обеспечивает наиболее быстрый поиск нужного правила. Актуальной является задача извлечения подобных зависимостей из уже накопленного архива документации. Для этого необходимо адаптировать методы Data Mining для работы с электронным архивом технической документации. Существующие программные продукты извлечения знаний часто ориентированы на

конкретную предметную область (например, Deep Data Diver™ – в основном на медицинскую диагностику, AnswerTree – на маркетинговые исследования) и не предоставляют возможности интеграции с программами сканирования и верификации. Поэтому необходимо разработать специализированную систему, предназначенную для работы с техническими документами различной структуры.

Точность верификации необходима для быстрого поиска полного набора документов по заданному пользователем запросу. Например, пользователю может понадобиться комплект документации по релейной защите на некотором объекте или сводный сметный расчёт по подстанции. Необходимо предоставить возможность построения сложных запросов по тексту документа с учётом морфологических форм заданных слов. Актуальна также задача разработки методов и средств, обеспечивающих более высокую скорость поиска документов по сравнению с существующими системами. Причём, важен не только и не столько поиск конкретного документа, сколько поиск полного набора документов, удовлетворяющих набору условий. Для решения этой задачи необходимы дополнительное структурирование и группировка документов. Следовательно, актуальна задача автоматизации создания пакетов документов по каким-либо критериям.

В целом, основные требования к архиву могут быть сформулированы следующим образом:

- хранение больших объёмов документации (порядка сотен тысяч страниц);
- ориентация на техническую документацию, насыщенную символьными наименованиями, которые должны иметь единый вид;
- высокая скорость занесения документов в архив с учётом существующей системы хранения документов. Комплект документации по объекту объёмом 5000 страниц должен быть доступен в архиве не более чем за 2 недели, срочные документы должны быть занесены в архив в течение дня с учётом всех временных задержек;
- обеспечение возможности поиска документа по тексту;
- наличие средств автоматизированной группировки документов.

Анализ рынка программного обеспечения показал, что существующие системы не полностью удовлетворяют приведённым требованиям. Таким образом, актуальна задача разработки системы электронного архива, решающей данные задачи.

Современное архивоведение, в том числе и зарубежное, подробно рассмотрено в трудах Е.В. Старостина, Е.В. Булюлиной. Автоматизация архивного дела рассматривается в работах И.Н. Киселёва, В.И. Тихонова, Е.В. Бобровой. Задачам извлечения знаний из документов посвящены труды И.П. Норенкова, В.А. Дюка, Р. Михальски, К. Парсайе. В соответствии с ГОСТ 23501.101-87, электронный архив может быть отнесён к обслуживающим подсистемам САПР. Основы построения интеллектуальных САПР с применением технологий знаний рассмотрены в трудах И.П. Норенкова, П. Хилла, Дж. Джонса. Информационная интеграция и построение

корпоративных информационных систем рассматриваются в работах В.Н. Буркова, Н.Г. Твердохлеба, В.Н. Шведенко, Д. О'Лири, И.Д. Ратмановой, М.Г. Левина, А. Леона.

Работа выполнялась в ОАО «Электроцентромонтаж», занимающемся проектированием, строительством и реконструкцией энергетических объектов, монтажом и наладкой электрооборудования. Промышленное внедрение и эксплуатация выполнялись в 4-х филиалах этой же организации.

Диссертационная работа соответствует паспорту специальности 05.13.12 «Системы автоматизации проектирования (по отраслям)», так как затрагивает следующие вопросы:

- научные основы построения средств автоматизации проектирования, безбумажного документооборота и процессов работы электронных архивов технической документации (пункт 7 областей исследований в паспорте специальности);
- научные основы реализации жизненного цикла «проектирование – производство – эксплуатация», построения интегрированных средств управления и унификации прикладных протоколов информационной поддержки;
- разработка принципиально новых методов и средств взаимодействия «проектировщик – среда».

Цель работы. Целью работы является повышение скорости доступа к актуальной проектно-конструкторской и технической документации путём создания электронного архива документов, а также точности и скорости верификации документов при загрузке в архив путём использования уже накопленных в архиве знаний. При этом решались следующие задачи:

1. Разработка информационно-аналитической модели электронного архива, обеспечивающей хранение массивов технической документации объёмом порядка сотен тысяч документов, группировку документов в соответствии с ГОСТами серии СПДС, использующимися в энергетике.

2. Разработка метода анализа документов, позволяющего оптимизировать верификацию и структурировать документы путём извлечения и применения нечётких продукционных правил.

3. Разработка методов и средств поиска в электронном архиве, позволяющих построить полный набор документов по запросу пользователя при заданных ограничениях на время и общее количество документов.

4. Экспериментальная проверка разработанных моделей и методов путём реализации в программной системе электронного архива.

Методы исследования. Использовались методы Data Mining, нечёткой математики, теории баз данных, систем искусственного интеллекта.

Научная новизна результатов.

1. Разработана информационно-аналитическая модель электронного архива, позволяющая хранить документы и извлекать знания в виде нечётких продукционных правил. Она отличается от существующих моделей хранения тем, что позволяет варьировать набор атрибутов документа для разных типов, учитывать соответствие атрибутов типов и создавать на их основе продукционные правила.

2. Разработан метод анализа атрибутивного состава технической документации, основанный на разработанной модели метаданных и включающий в себя алгоритмы создания наборов правил-ассоциаций (справочников) и поиска последовательностей. Он отличается от существующих алгоритмов Data Mining, например FP-Growth, Apriori и их разновидностей, тем, что учитывает структуру хранения документов и особенности технических документов.

3. Разработан метод решения задачи кластеризации в архиве. Использование кластеризации позволяет группировать документы в соответствии с ГОСТами, либо по индивидуальным запросам пользователей. Метод отличается от традиционных алгоритмов агломеративной кластеризации тем, что вместо расстояния между точками использует разработанную модель метаданных, а также подготовленные на этапе анализа наборы продукционных правил. Это обеспечивает высокую скорость работы.

Практическая значимость работы.

1. На основе разработанной модели данных создана система электронного архива, позволяющая получать доступ к необходимым документам непосредственно с рабочих мест.

2. Применение методов извлечения знаний позволило сократить время верификации документов и увеличить скорость занесения документов в архив приблизительно на 25%, не увеличивая штат верификаторов. Метод позволяет извлекать знания с учётом того, что требуемые значения могут находиться в разных атрибутах, а также применять полученные знания при верификации без дополнительной интерпретации.

3. Разработанный метод поиска в архиве, использующий оригинальную схему взаимодействия компонент и дополнительные средства СУБД, обеспечивает построение полного списка документов по пользовательскому запросу при заданных временных ограничениях. Наличие атрибутивного и полнотекстового поиска позволяет учитывать многообразие технической документации и выполнять поиск только нужных пользователю документов.

4. Применение методов кластеризации позволяет более наглядно группировать документы в пакеты, что упрощает работу пользователям, например, при подготовке отчётов или комплектов технической документации по определённому объекту.

Апробация работы. Материалы диссертационной работы докладывались и обсуждались на следующих конференциях:

1) I Международная конференция «Автоматизация управления и интеллектуальные системы и среды (АУИСС - 2010)»;

2) XVI Международная открытая научная конференция «Современные проблемы информатизации» (2011);

3) конференция «Спецпроект: анализ научных исследований» (30-31.05.2011г);

4) конференция «Наука в информационном пространстве – 2011» (29-30.09.2011г.).

Публикации. По результатам работы опубликованы 2 статьи в изданиях, рекомендованных ВАК, 6 статей в научных журналах, 5 тезисов докладов на конференциях, получено 1 свидетельство о государственной регистрации программы для ЭВМ.

Личный вклад. Выносимые на защиту модели и методы разработаны автором лично. В созданной системе электронного архива автором разработаны система шаблонов Flexi Capture, система конфигурирования, мастер загрузки документов, компоненты, реализующие описанные в диссертации методы.

Внедрение. Система ДокПрофи™ зарегистрирована в Реестре программ для ЭВМ, номер свидетельства 2011610409. Успешно внедрена и применяется в ОАО «Электроцентромонтаж» для оперативного доступа сотрудников предприятия к актуальной технической документации. Тем самым заложена основа для единого информационного пространства предприятия.

Структура и объем работы. Диссертация состоит из введения, четырёх глав, заключения, списка литературы из 101 наименования и включает 138 страниц основного текста, 36 рисунков, 3 таблицы, 8 формул. В приложении приведены 4 акта о внедрении и 1 свидетельство о государственной регистрации программы для ЭВМ.

СОДЕРЖАНИЕ РАБОТЫ

Во введении определены цели и задачи исследования, обоснована актуальность выбранной темы, сформулированы полученные научные результаты, перечислены основные положения, выносящиеся на защиту.

Первая глава посвящена анализу существующих систем электронного архива, а также существующих средств и методов извлечения знаний. На основе требований, предъявляемых к системе электронного архива в энергетической отрасли, был проведён анализ существующих систем. Однако найти готовый продукт, полностью удовлетворяющий данным требованиям, не удалось. Многие продукты не ориентированы на техническую документацию (Docs Fusion, LanDocs), не предоставляют возможности сканирования и распознавания (Staff ware, MS Sharepoint Portal Server). Есть адаптируемые платформы, с помощью которых можно реализовать

требуемый функционал, например SmartPlant Foundation. Однако стоимость их адаптации превысит стоимость покупки в 3 – 5 раз (средние статистические данные рынка). Средний срок адаптации и внедрения подобных систем составляет 1,5 – 2 года. Среди недостатков также можно отметить ресурсоёмкий поиск по содержанию документа и недостаточно высокую скорость занесения новых документов в архив. Также существуют системы NormaCS и TDMS, ориентированные на техническую документацию, однако их связь с программами сканирования пока в стадии разработки. В связи с этим актуальной является разработка специализированного программного обеспечения, удовлетворяющего приведённым требованиям.

Для ускорения занесения документов в архив важной является задача извлечения зависимостей значений атрибутов из ранее загруженных документов для повышения точности и скорости верификации. Программное средство извлечения знаний должно выполнять следующие функции и удовлетворять условиям:

- анализ и поиск закономерностей в архиве;
- высокая скорость работы с сохранением точности;
- учёт нечёткости совпадений;
- гибкая настройка на документы с переменным набором атрибутов;
- интеграция со справочниками, например справочником организаций 1С;
- возможность интеграции с электронным архивом и программами сканирования и верификации.

Был проведён анализ существующих средств извлечения знаний. Существующие российские (комплекс AT-Технология, аналитическая платформа Deductor, система Deep Data Diver™ и др.) и зарубежные (аналитический модуль AnswerTree, система WizWhy, система See5/C5.0) разработки имеют свои достоинства и недостатки. Однако многие из них ориентированы на работу с конкретной предметной областью и не предоставляют возможности интеграции с программами сканирования и верификации. Так как названные системы не удовлетворяют всем требуемым условиям, необходима разработка специализированных компонентов, предназначенных для работы с электронным архивом документов и интегрируемых с программой верификации.

Таким образом, в результате анализа выявлено, что актуальной является разработка специализированной системы электронного архива, обеспечивающей основу для создания единого информационного пространства предприятия энергетической отрасли. Такая система обеспечит хранение проектной, технической, строительной, эксплуатационной, нормативной, юридической информации на всех этапах жизненного цикла энергетического объекта, предотвращая его информационный износ. Как показывает анализ, главной причиной аварий является именно информационный, а не физический износ. Соответственно, внедрение электронного архива позволит повысить эффективность работы предприятия в целом и уменьшить вероятность аварийных ситуаций.

Вторая глава посвящена решению первой задачи построения электронного архива – разработке информационно-аналитической модели. Разработанная модель метаданных, интегрируемая с моделями хранения документов и учитывающая соответствие атрибутов документов различных типов, позволяет реализовать поиск знаний в документах с учётом особенностей технической документации.

Основные принципы разработанной модели данных:

1. Данные, содержащиеся в документах архива, хранятся в трёх основных блоках: заголовок документа; атрибуты документа; файлы документа.
2. Система должна обеспечивать версионность документов. Для этого используются «таблицы-версии» и «таблицы-двойники». Для атрибутов и файлов документов создаются исторические таблицы-двойники, и для документа в целом ведётся таблица версий.
3. В системе используются связи между значениями атрибутов в виде продукционных правил. Это обусловлено линейным характером корреляции атрибутов документа, а также линейностью процесса верификации: при подтверждении значения одного атрибута необходимо подставить значение другого.

Информационная модель электронного архива должна обеспечивать поддержку хранения и поиска документов, поиска закономерностей в документах.

Далее рассмотрена модель, описывающая хранение документов и извлекаемых закономерностей. Она имеет следующий вид:

$$M = (AA, DT, D, SA, SS),$$

где $AA = \{aa_1, aa_2, ..., aa_{|AA|}\}$ – множество возможных атрибутов документов. Каждый атрибут имеет имя и тип;

$DT = \{dt_1, dt_2, ..., dt_{|DT|}\}$ – множество типов документов. Каждый тип документа dt_i представляется в виде набора $dt_i = \{type\_name, pt, TA\}$, где type_name – название типа, $pt \in DT$ – родительский тип для поддержки вложенных типов, $TA = \{ta_1, ta_2, ..., ta_{|TA|}\}$ – множество атрибутов типа. Каждый атрибут типа ta_i представляет собой набор $ta_i = \{aname, atype, forder, fsort\_order, fsort\_type, uniqueness\_check, req\}$, где aname – наименование атрибута, atype – тип атрибута (строка, число, дата и др.), forder – номер по порядку, fsort_order и fsort_type – порядок при сортировке и её тип, uniqueness_check – признак задания уникальности документа, req – обязательность;

$D = \{d_1, d_2, ..., d_{|D|}\}$ – множество документов. Каждый документ d_i представляет собой следующий набор $d_i = \{dt, doc\_name, reg\_date, reg\_number, owner\_filial, active\_doc, DA\}$, где $dt \in DT$ – тип документа, doc_name – имя документа, reg_date и reg_number – регистрационные дата и номер, owner_filial – филиал, в котором он был создан, active_doc – признак активности документа, $DA = \{da_1, da_2, ..., da_{|DA|}\}$ – множество

атрибутов документа. Атрибут документа da_i представляет собой набор $da_i = \{ta, value, active\_attribute\}$, где $ta \in TA$ – атрибут типа, value – его значение, active_attribute – признак активности атрибута. В отличие от атрибутов типа, задающих некоторые шаблоны документов, атрибуты документа представляют собой конкретные значения для конкретных документов;

$SA = \{sa_1, sa_2, ..., sa_{|SA|}\}$ – множество правил-ассоциаций, связывающих значения атрибутов в одном документе. Ассоциация представляет собой правило вида «*Если* $ta_1 = s$, *то* $ta_2 = s_k$ *с вероятностью* x_k ». Пусть $S(ta_1)$ – множество всех значений атрибута ta_1, $S(ta_2)$ – множество всех значений атрибута ta_2. Тогда множество ассоциаций (справочник) представляет собой декартово произведение множеств:

$$S(ta_1) \times S(ta_2) = \{(s, s_k) : s \in S(ta_1), s_k \in S(ta_2)\}, \quad (1)$$

причём каждой паре (s, s_k) соответствует значение вероятности ее появления в соответствующих атрибутах документа x_k. Пример ассоциации: Если атрибут «Номер договора» равен «11-РП-11», то атрибут «Объект» равен «ПС Примерная» с вероятностью 95%, «Подстанция Примерная» с вероятностью 4% и «Пример» с вероятностью 1%;

$SS = \{ss_1, ss_2, ..., ss_{|SS|}\}$ - множество правил-последовательностей, связывающих значения атрибутов в разных документах. Введём обозначение $S_d(da_i)$ для значения атрибута da_i в документе d. Последовательности ищутся для каких-либо типов T_1 и T_2, которые имеют соответствующие множества атрибутов TA_1 и TA_2. Пусть $TA_1 \cap TA_2 = TA$. Последовательности имеют смысл, когда мощность множества $|TA| \geq 2$. Тогда $\forall A_1, A_2 \in TA$ необходимо найти вероятность x_{12} истинности выражения

$$S_{d_1}(A_1) = S_{d_2}(A_1) \Rightarrow S_{d_1}(A_2) = S_{d_2}(A_2). \quad (2)$$

Таким образом, множество последовательностей представляет собой декартово произведение $TA \times TA = \{(A_1, A_2) : A_1, A_2 \in TA\}$, где каждой паре (A_1, A_2) соответствует значение вероятности x_{12}. Очевидные свойства вероятностей: $x_{ii} = 1$ и $x_{ij} = x_{ji}$ – следуют из определения последовательности. Пример последовательности: Если в документах типов «Рабочая документация» и «Акт освидетельствования работ» совпадают значения атрибутов «Объект», то значения атрибутов «Адрес объекта» совпадут с вероятностью 95%. Подобные правила следует использовать, когда документы создаются последовательно, один на основе другого.

Введено также отношение соответствия атрибутов типов. Атрибут ta_1 типа dt_1 называется соответствующим атрибуту ta_2 типа dt_2, если в этих атрибутах может содержаться одинаковая по сути информация. Для таких атрибутов будем использовать обозначение

$$ta_1 \leftrightarrow ta_2. \quad (3)$$

Пусть, например, D – комплект документов по энергетическому объекту, где помимо всего прочего используются шкафы телемеханики. Это сочетание в разных типах документов dt_i может находиться в разных атрибутах ta_j . Например, в чертежах оно содержится в наименовании документа, в рабочей и сметной документации – в списках оборудования, в акте о приёмке работ – в списке сданных работ. Соответственно, можно сказать, что наименование чертежа соответствует оборудованию в рабочей документации и сданным работам в акте о приёмке работ. Правила продукции могут быть применены не только к атрибутам, непосредственно указанным в них, но и ко всем соответствующим атрибутам.

Рассмотрена операционная семантика модели. На этапе конфигурирования архива задаются множество возможных атрибутов документов AA и множество типов документов DT. При создании типа dt_i задаётся множество его атрибутов на основе элементов множества AA. Если созданные атрибуты и типы документов ещё не были использованы в архиве, то их можно изменить или удалить. Документы d_i создаются в процессе работы с архивом. При создании экземпляра документа задаётся его тип dt, на основе атрибутов типа создаются атрибуты документа DA. Значения атрибутов заполняются в процессе верификации документа. Если документ занесён в архив, он не может быть удалён, может быть изменён только признак активности. Набор типов может иметь иерархическую структуру. Это даёт возможность группировки документов в пакеты в соответствии с обозначениями по ГОСТу, а также в соответствии с желанием пользователя.

Первоначально составление продукционных правил выполняется по существующим документам архива с помощью методов, приведённых в главе 3. Далее при занесении новых документов производится корректировка правил, то есть выполняется самообучение. Отметим также, что простая пара «ключ - значение» обеспечивает наиболее высокую скорость поиска подходящих правил во время верификации по сравнению с другими моделями представления знаний.

Для обеспечения быстрого поиска документов по атрибутам или тексту с возможностью искать документы по фразе, синонимам, а также на разных языках необходимо разработать информационную основу. Разработанная подмодель использует метаданные СУБД, поисковые индексы, таблицы с двоичным индексом и HASH-таблицы для увеличения скорости поиска.

Для представления слов в поисковом индексе используется следующая модель. Имеются несколько словарей Dict = (Name, Lang, Words), где Name – наименование словаря, Lang – язык, Words – набор слов. Слово (Word) имеет следующие характеристики: Word = (SearchWord, Flags, Synonyms), где SearchWord – слово для поиска, Flags – набор флагов, характеризующий аффиксы для данного слова, Synonyms – множество синонимов типа Word. Данная модель позволяет учитывать различные формы слов. Например, для поиска сметных расчётов по разделу ЭО-ВР

необязательно знать точное наименование документа. Достаточно задать поиск по словам «Раздел ЭО-ВР» и получить необходимые документы.

Итак, разработанная модель данных в виде набора основных сущностей электронного архива и продукционных правил, связывающих значения атрибутов документов, является информационной основой для работы с документами, пакетами документов и типами документов электронного архива. Она предоставляет возможности создания, обновления и получения всех версий документа, настройки типов и пакетов документов. Модель правил позволяет извлекать закономерности из документов и применять их при верификации и группировке. Подмодель поиска документов является информационной основой атрибутивного и полнотекстового поиска документов электронного архива. На её базе возможен поиск с широким набором параметров.

Третья глава посвящена решению следующей задачи: разработке методов и алгоритмов извлечения и применения закономерностей в технических документах на основе созданной модели данных. Как отмечалось в главе 2, для описания корреляций между атрибутами документов архива наиболее подходящими моделями являются продукционные правила. Такая форма обеспечивает простую форму правил «ключ -

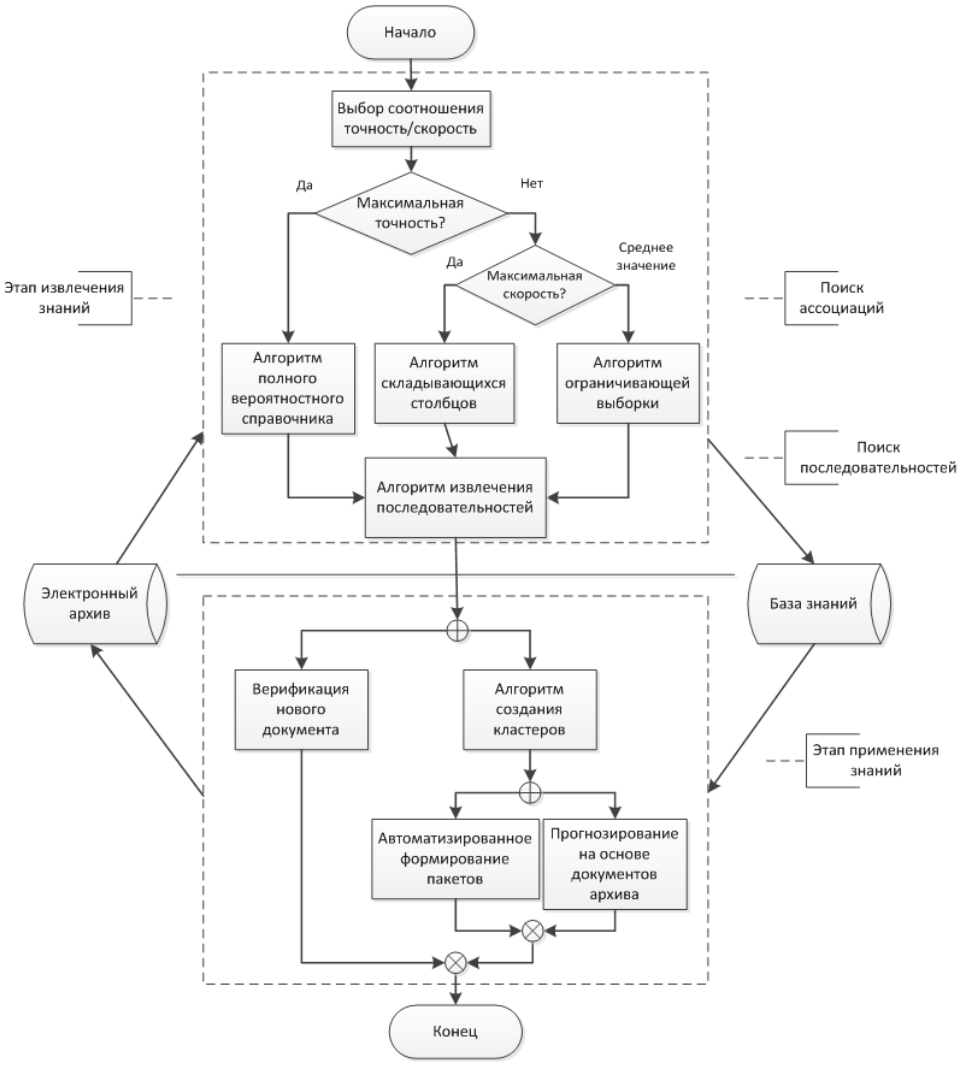

Рис. 1. Общая блок-схема метода

значение», что позволяет находить нужные правила максимально быстро. Полученные правила могут быть применены при верификации новых документов, а также при создании кластеров для автоматизированного формирования пакетов документов и прогнозирования значений отдельных атрибутов документов. Таким образом, применение данных правил позволит решить одну из задач работы: ускорение занесения документов в архив и автоматизация группировки документов по пакетам.

Общая блок-схема разработанного метода приведена на рис. 1 и состоит из двух этапов: этап извлечения поиска закономерностей и этап применения закономерностей.

В работе предложены следующие алгоритмы, указанные на схеме.

1. *Алгоритмы создания справочника (поиска ассоциаций):*

   - *алгоритм полного вероятностного справочника.* Пусть D – множество документов d некоторого типа dt , $A_1, A_2 \in TA$ – некоторые атрибуты типа. Последовательным перебором документов находим все встречающиеся пары значений атрибутов $S_d(A_1)$ и $S_d(A_2)$ и количество повторений пар в документах. Затем, разделив полученные количества на общее число документов, получаем справочник. Так как в этом методе дважды используется бинарный поиск, эффективность алгоритма – $O(\log_2^2 n)$, где n – количество документов;

   - *алгоритм складывающихся столбцов.* Пусть $d_1, d_2 \in D$ – пара документов. Тогда если $S_{d1}(A_1) = S_{d2}(A_1)$, но $S_{d1}(A_2) \neq S_{d2}(A_2)$, то в справочник записывается значение некоторой функции сложения $F(S_{d1}(A_2), S_{d2}(A_2))$. Эта функция может быть по-разному определена на разных типах значений. В результате получаем справочник, в котором все вероятности равны 100%. Эффективность данного алгоритма – $O(\log_2 n)$;

   - *алгоритм ограничивающей выборки.* Данный метод комбинирует два предыдущих с учётом особенностей архива. В качестве примера выборки возьмём выборку по периодам. Каждый документ архива имеет в качестве обязательного атрибута дату регистрации. Общее множество документов можно разбить на непересекающиеся подмножества по каким-либо периодам (например, месяцы, годы), далее построить справочники методом складывающихся столбцов. Полученные справочники можно объединить в один, который будет показывать, сколько раз то или иное значение встречается в справочниках. Эффективность данного алгоритма находится в интервале $(O(\log_2 n); O(\log_2^2 n))$ и зависит от размера периода разбиения. При уменьшении периода алгоритм стремится к алгоритму полного вероятностного справочника, при увеличении – к алгоритму складывающихся столбцов. Целесообразно использовать данный метод и, варьируя значение периода, подобрать оптимальное соотношение скорости и точности. В качестве примера, анализируя

паспорта теплоизоляционных плит, можно выделить следующий набор ассоциаций: «Если марка плиты N-III-60-L, то коэффициент теплопроводности при 25^0C в Вт/(м·K) равен 0,0324 с вероятностью 95% и 0,0035 с вероятностью 5%».

2. *Алгоритм извлечения последовательностей.* Для каждого значения $S_{d_1}(A_1)$ ищутся все документы d, у которых $S_{d_1}(A_1) = S_d(A_1)$. Далее подсчитывается количество документов, где совпадают значения вторичного атрибута, то есть $S_{d_1}(A_2) = S_d(A_2)$. Пример извлекаемой последовательности: «Если в акте освидетельствования работ наименование проектной документации равно номеру чертежа, то наименование материалов изделий в этих документах совпадёт с вероятностью 98%».

Приведённые методы работы с ассоциациями и последовательностями извлекают знания из архива в виде связей между значениями атрибутов документов одного или разного типов. Это позволяет без дополнительной интерпретации применять их при верификации новых документов, при этом исправляется большое количество ошибок распознавания. Полученные правила могут быть применены не только к атрибутам, на которых строились правила, но и к соответствующим им по формуле *(3)* атрибутам. Благодаря этому значения атрибутов приводятся к единому виду. Это особенно важно для технической документации, где необходимо единообразное наименование технических средств в разных документах.

3. *Алгоритм кластеризации.* Он позволяет объединять документы архива в группы на основе значений какого-либо атрибута. Сначала строится справочник одним из описанных выше методов. Далее для каждого значения основного атрибута A_1 выбираются значения вторичного атрибута A_2, если они встречались в каком-либо правиле справочника. Аналогично для каждого полученного значения вторичного атрибута A_2 выбираются значения основного атрибута A_1. Такая процедура называется шагом кластеризации. Если полученное множество значений основного атрибута C отличается от исходного, то шаг кластеризации повторяется. Например, справочник представляет собой некоторую таблицу пар «Наименование организации – Адрес организации». Начинаем работу с некоторого конкретного адреса. Найдём все наименования организации, соответствующие этому адресу. Затем всем этим наименованиям найдём соответствующие адреса, которые могут не совпадать по написанию с ранее учтёнными. Если был найден новый вариант адреса, то повторим процедуру. В результате в кластер отбираются все варианты наименований и адресов одной и той же организации. Отметим, что справочники по разным парам атрибутов могут задавать разные разбиения на кластеры, поэтому в методе используются только пары атрибутов. Данный алгоритм используется для автоматизированной группировки документов, например, при разбиении документов по папкам в соответствии с буквенной маркой по ГОСТу. При этом рассматриваются пары атрибутов «Название документа» – «Номер документа», например, «Огнезащита металлоконструкций. Спецификация материалов» – «05-М-08-01-ОЗ1.СО изм. 1».

Таким образом, на основе созданной модели данных разработаны алгоритмы, позволяющие извлекать закономерности из документов и применять их при верификации и анализе. Эти алгоритмы отличаются от существующих алгоритмов Data Mining следующим:

- направлены на работу с электронным архивом, так как базируются на специализированной модели хранения данных;
- учитывают специфику энергетической отрасли, так как ориентированы на быстрый анализ больших объёмов информации, включая специфические символьные обозначения, и позволяют приводить неструктурированный набор документов к сложным структурированным спискам, требуемым ГОСТом.

Реализация данных алгоритмов в системе позволяет ускорить загрузку документов в архив и обеспечить их автоматизированную группировку.

В четвертой главе рассматриваются вопросы реализации модели и методов, приведено краткое описание системы электронного архива. Была поставлена задача перевода архива технической документации ОАО «Электроцентромонтаж» порядка 1 миллиона страниц в электронно-структурированный вид. Для этого разработана система ДокПрофи™, её общая архитектура представлена на рис. 2 в виде диаграммы развёртывания.

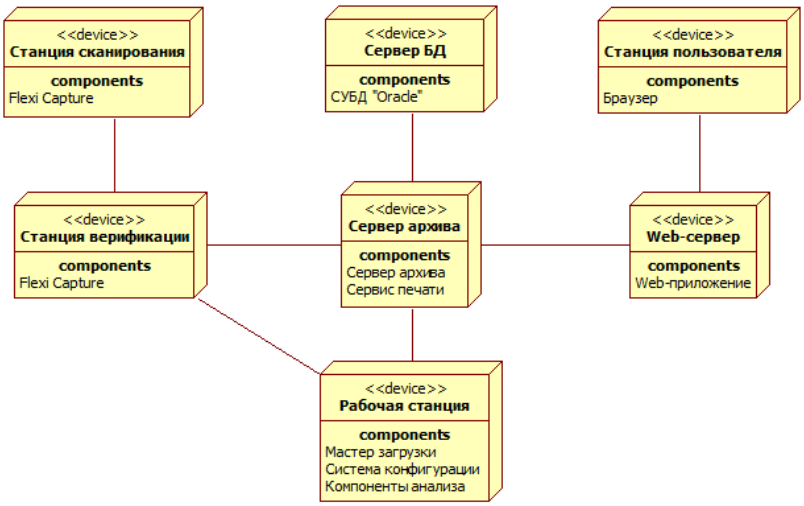

Рис. 2. Диаграмма развертывания архива

Данные хранятся в базе данных на основе реализованной модели хранения документов. Для доступа используются виртуальные представления (view) и хранимые процедуры. Приложения архива обращаются к ним унифицировано посредством сервера электронного архива. Применяемая архитектура даёт такие преимущества, как повышение скорости и надёжности работы, распределённость и простота доступа к сервису архива, масштабируемость, многоуровневая защита данных. Каждый компонент системы может быть изменён независимо от других: есть возможность сменить платформу СУБД, использовать нужные языки программирования для реализации клиентских приложений, добавлять и удалять новые компоненты архива.

Оценка эффективности разработанного в диссертации метода оценена следующим способом. Отслеживалась динамика скорости занесения в архив документов по нескольким объектам для нескольких верификаторов. В начале внедрения, когда количество правил было минимальным, средняя скорость составляла 100 страниц в день для одного человека. По мере занесения документов в архив происходила корректировка и внесение новых правил, что увеличило среднюю скорость занесения до 120 страниц в день. При начале работы с новым объектом скорость занесения уменьшалась до 110 страниц в день, так как документация по новому объекту подразумевает новые закономерности в атрибутах. Однако по мере работы скорость вновь увеличивалась до 120–125 страниц в день. Таким образом, использование метода позволяет увеличить количество верифицируемых документов в день в среднем на 20–25%. Это позволяет внедрять архив в других филиалах и организациях при сохранении существующего штата верификаторов.

Также в четвертой главе описана реализация возможности автоматизированного формирования пакетов документов на основе алгоритма кластеризации, предложенного в главе 3. Это позволяет провести реорганизацию пакетов, что упрощает поиск документов и подготовку отчётов на их основе. Время подготовки комплекта документации по объекту после внедрения системы сократилось на 50%, отчёта по работам на объекте на основе актов КС-2 и справок КС-3 – на 70%.

Совместная работа с документами и удалённый доступ реализуются с помощью Web-приложения (рис. 3). Оно реализует функции поиска и просмотра нужных документов. Представлены три вида поиска документов: поиск по атрибутам, полнотекстовый поиск и смешанный поиск. Атрибутивный поиск реализован на разработанной модели данных в виде параметризованных запросов. Для

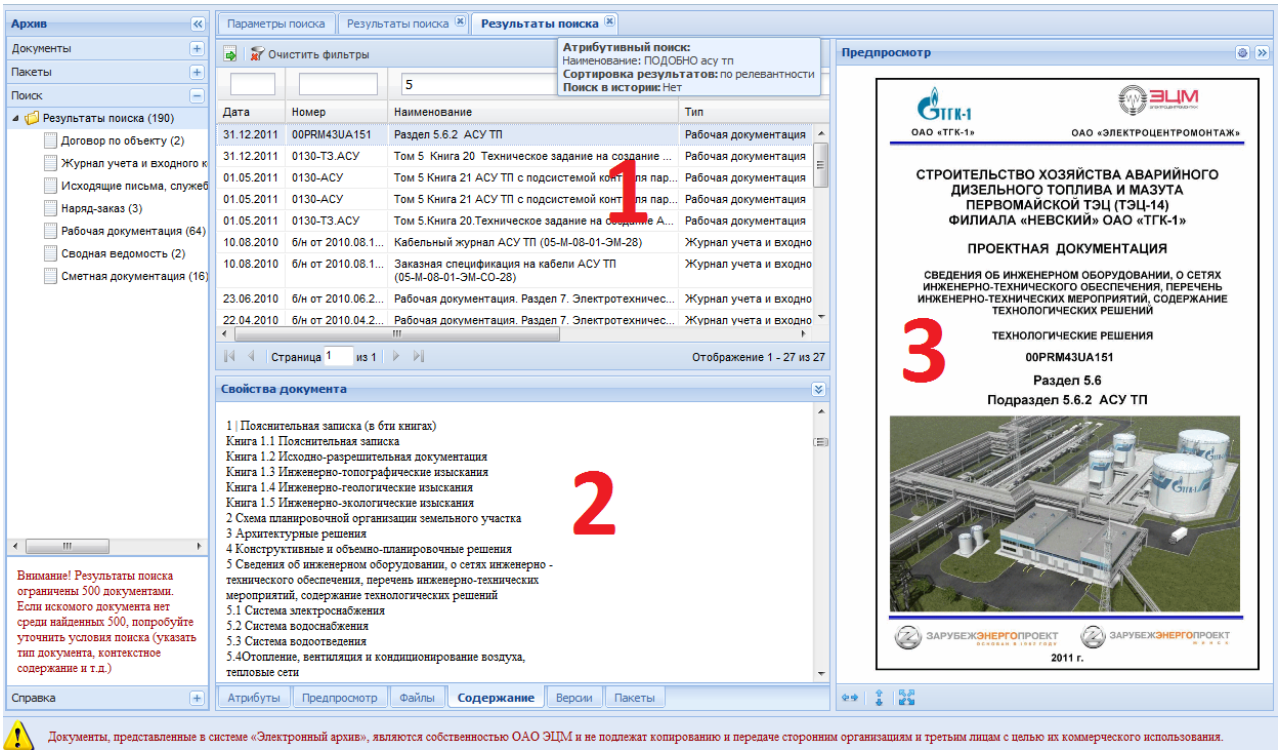

Рис. 3. Web-приложение архива

полнотекстового поиска использовалась разработанная схема взаимодействия компонент, что позволило сократить время поиска до требуемого ограничения. Параметры поиска задаются на первой вкладке (рис.3, область 1). На следующей вкладке выводятся результаты поиска по заданному условию. Также имеется возможность быстрой фильтрации списка документов. В области 2 отображаются свойства выбранного документа: атрибуты, файл предпросмотра, другие файлы, полнотекстовое содержание, версии документа, пакеты, в которые он входит. В области 3 также отображается файл предпросмотра документа. Экспериментальная проверка показала, что после внедрения системы электронного архива время поиска документов сократилось с 30–60 до 3–5 минут.

Разработанная система успешно применяется в ОАО «Электроцентромонтаж»: в архив занесена основная часть технической документации по различным объектам, она активно используется сотрудниками различных отделов. Проведённое анкетирование показало, что большинство пользователей системы отмечают удобство работы и поиска документов и наполненность архива всеми нужными документами. Отмеченные пользователями недостатки, например отсутствие нужных типов документов, были исправлены средствами системы. Реализованная система позволила привести техническую документацию к единому электронному виду, предоставлять доступ к ней сотрудникам предприятия непосредственно с рабочего места, осуществлять поиск документов по заголовкам и содержанию, а также группировать документы по пакетам в соответствии с задачами пользователей.

К перспективам развития системы можно отнести возможность реализации клиентских приложений архива, взаимодействующих с используемыми САПР. Это позволит работать с архивом (добавлять и извлекать документы) непосредственно в среде разработки.

В заключении подведены итоги работы и сделаны основные выводы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. Разработана информационно-аналитическая модель электронного архива, обеспечивающая хранение массивов технической документации объёмом порядка сотен тысяч документов, группировку документов в соответствии с ГОСТами серии СПДС, использующимися в энергетике.
2. Разработан метод анализа документов, позволяющий оптимизировать верификацию и структурировать документы путём извлечения и применения нечётких продукционных правил. Разработан метод кластеризации, позволяющий автоматизировать создание пакетов документов на основе полученных правил.
3. Разработаны методы и средства поиска в электронном архиве, позволяющие построить полный набор документов по запросу пользователя при заданных ограничениях на время и общее количество документов.
4. Разработана клиент-серверная архитектура приложения и схема взаимодействия компонент архива. На основе этого реализована тиражируемая система электронного архива и достигнута требуемая скорость поиска документов.

Приведённая архитектура в перспективе позволит разработать клиентские приложения для существующих САПР в целях доступа к документам из сред разработки.

5. Разработанные модели и методы показали свою эффективность при реализации на предприятии энергетической отрасли. Время поиска документа сократилось до 3–5 минут, скорость занесения в архив увеличилась на 20%, время подготовки комплекта документации по объекту сократилось на 50–70%. Предложенные методы могут быть применены при реализации архива, ориентированного на схожие предметные области. Тем самым решена задача оперативного доступа сотрудников ОАО «Электроцентромонтаж» к актуальной проектно-конструкторской и технической документации путём создания электронного архива документов.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ РАБОТЫ

Публикации в изданиях, рекомендованных ВАК РФ:

1. Кроль Т.Я., Харин М.А., Евдокимов П.В. Методы создания справочника на основе электронного архива / Т.Я. Кроль, М.А. Харин, П.В. Евдокимов // Известия «КБНЦ РАН». – 2011. – №1. С. 154 – 158.

2. Кроль Т.Я., Харин М.А. Опыт построения и реализации электронного архива на базе системы сканирования и распознавания Flexi Capture / Т.Я. Кроль, М.А. Харин // Программная инженерия. – 2012. – №6. – С. 35 – 42.

3. Свидетельство о государственной регистрации программы для ЭВМ «ДокПрофи» / Т.Я. Кроль, А.С. Карпов, Д.В. Иванов, А.С. Капитонихин, А.С. Угрюмов, М.А. Харин, Е.А. Воробьёв. – № 2011610409; дата 11.01.2011.

Публикации в прочих изданиях:

4. Харин, М.А. Обзор средств автоматизированного извлечения знаний и их применение в электронных архивах документов / М.А. Харин // Молодой учёный. — 2010. — №5. Т.1. — С. 106 – 108.

5. Харин, М.А. Электронные архивы документов и средства автоматизированного извлечения знаний / М.А. Харин // Информационные технологии моделирования и управления. – 2010. – № 2. – С. 242 – 246.

6. Кроль, Т.Я. Методы решения задачи кластеризации и прогнозирования в электронном архиве / Т.Я. Кроль, М.А. Харин // Молодой учёный. — 2011. — №6. Т.1. — С. 135 – 137.

7. Кроль, Т.Я. Методы поиска в электронном архиве / Т.Я. Кроль, М.А. Харин, Н.В. Никоноров, Д.В. Иванов // Информационные технологии моделирования и управления. – 2011. – № 6. – С. 702 – 709.

8. Кроль, Т.Я. Модели данных для реализации поиска и прав доступа к документам / Т. Я. Кроль, М.А. Харин, Д.В. Иванов, Н.В. Никоноров // Молодой учёный. — 2011. — №11. – С. 79 – 84.

9. Кроль, Т.Я. Использование методов кластеризации для автоматизированного формирования пакетов документов / Т.Я. Кроль, М.А. Харин // Молодой учёный. — 2012. — №10. – С. 93 – 95.

Труды конференций:

10. Кроль Т.Я., Харин М.А. Проблема верификации при занесении документов в электронный архив // Мат-лы I междунар. конф. «Автоматизация управления и интеллектуальные системы и среды (АУИСС - 2010)». Россия, Приэльбрусье, 20-27 декабря 2010 г. [Электронный ресурс]. – Режим доступа: http://www.iipru.org/docs/auiss2010_tom2.pdf

11. Кроль Т.Я., Харин М.А. Особенности занесения документов в электронный архив. Мат-лы XVI междунар, откр. науч. конф. «Современные проблемы информатизации»); публ. с 01 по 31 января 2011 г. (Конференция проводится в дистанционном режиме). [Электронный ресурс]. – Режим доступа: http://www.sbook.ru

12. Кроль Т.Я. Схема наполнения электронного архива документами / Т.Я. Кроль, М.А. Харин, П.В. Евдокимов // Мат-лы I междунар. Конф. «Автоматизация управления и интеллектуальные системы и среды». Терскол, 20-27.12.2010. Т. IV. – Нальчик, 2010. – С. 53 – 56.

13. Кроль Т.Я., Харин М.А. Использование последовательностей при занесении документов в электронный архив // Мат-лы конф. «Спецпроект: анализ научных исследований», 30-31.05.2011г. [Электронный ресурс]. – Режим доступа: http://www.confcontact.com/20110531/tn8_krol.htm

14. Кроль Т.Я., Харин М.А. Расширение модели документа электронного архива с целью извлечения и использования накопленных знаний // Мат-лы конф. «Наука в информационном пространстве – 2011», 29-30.09.2011г. [Электронный ресурс]. – Режим доступа: http://www.confcontact.com/20110929/tn_hrol.htm

ХАРИН Максим Алексеевич

РАЗРАБОТКА МОДЕЛЕЙ И МЕТОДОВ ВЕРИФИКАЦИИ И АНАЛИЗА ДОКУМЕНТОВ В ЭЛЕКТРОННОМ АРХИВЕ ЭНЕРГЕТИЧЕСКИХ ОБЪЕКТОВ

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата технических наук

Подписано в печать _____. Формат 60x84 1/16.
Печать плоская. Усл. печ. л. 1,16. Тираж 100 экз. Заказ №

ФГБОУВПО «Ивановский государственный энергетический университет имени В. И. Ленина».
Отпечатано в УИУНЛ ИГЭУ
153003, г. Иваново, ул. Рабфаковская, 34.