OSTIS-2011
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

ПРИМЕНЕНИЕ ОНТОЛОГИЧЕСКОГО ПОДХОДА И МУЛЬТИАГЕНТНОЙ ТЕХНОЛОГИИ ДЛЯ СОЗДАНИЯ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ УПРАВЛЕНИЯ ДОКУМЕНТАМИ

В.В. Ланин (*lanin@perm.ru*)
Пермский государственный университет, г. Пермь, Россия

Рассматриваются средства управления документами, позволяющие решать различные задачи, связанные с обработкой электронных документов в информационных системах. Описывается подход к анализу неструктурированных электронных документов с использованием онтологий и мультиагентной парадигмы. С помощью данного подхода может быть решен широкий класс задач, включающий как традиционные задачи обработки текстов, так и средства интеллектуального анализа документов при создании моделей предметных областей при проектировании информационных систем.

Ключевые слова: интеллектуальная обработка электронных документов, разработка информационных систем, онтологии, мультиагентные системы.

Введение

Задача создания информационных систем (ИС), позволяющих осуществлять гибкую настройку на потребности пользователей и бизнес-процессов, меняющиеся условия эксплуатации, – одна из критичных задач, так как возможности настройки определяют эффективность вложений в создание системы и стоимость ее эксплуатации, сопровождения. Свойство адаптирумости гарантирует возможность развития ИС, ее «живучесть» [Лядова, 2007]. Оно объединяет многие нефункциональные характеристики, обеспечить которые можно только при использовании для создания ИС инструментария, обладающего соответствующими возможностями. Как при создании системы, так и в ходе ее эксплуатации решаются задачи управления документами, их поиска, анализа и классификации, каталогизации и эффективного хранения, генерации и поддержания всего их жизненного цикла.

1. Задачи управления документами в динамически адаптируемых системах, управляемых моделями

Практически все задачи, решаемые при создании и эксплуатации системы, связаны с необходимостью работы с документами в различных форматах:

– *на этапе анализа предметной области происходит поиск и изучение документов, описывающих состояние предметной области*, свойства различных ее объектов, а также бизнес-процессы, которые должны быть автоматизированы, задающих условия функционирования создаваемой системы, предъявляемые к ней требования;

– *при создании системы каждый процесс должен быть документирован*, должна быть подготовлена программная документация, инструкции и руководства по эксплуатации ИС;

– в ходе эксплуатации системы пользователи получают отчеты, генерируют документы на основе информации, хранящейся в ИС и получаемой извне.

Включение средств BI (Business Intelligence), в частности, *средств подготовки отчетов* (*подсистем репортинга*) в состав ИС – требование к любой современной системе. Именно эти

средства позволяют представить информацию, получаемую пользователями, в удобном для них виде, визуализировать данные, результаты их обработки. Однако адаптируемая система должна обеспечить возможность настройки средств репортинга на любые изменения, которые могут произойти в ходе эксплуатации системы (дать возможность пользователям получать новые отчеты, менять формы документов и пр.). Желательно при этом снизить трудоемкость этой работы, минимизировать необходимость вмешательства разработчиков – позволить пользователям самим адаптировать ИС к своим потребностям.

Современные предметно-ориентированные средства создания ИС позволяют решить эти задачи, дать возможность принимать участие в создании системы и ее настройке в процессе эксплуатации пользователям-непрограммистам, являющимся специалистами в тех предметных областях, для которых разрабатываются ИС. В отличие от «традиционных систем», условия и правила функционирования которых не меняются в процессе эксплуатации, для систем, допускающих динамическую настройку, задачи анализа предметной области и разработки документации необходимо выполнять не только при создании системы, но и в ходе ее функционирования – каждый раз, как только возникает необходимость внесения изменений в «поведение» системы, адаптации ее к новым условиям и требованиям. Таким образом, *анализ документов и подготовка документации – задачи, которые выполняются в течение всего жизненного цикла ИС*. Решение этих задач требует автоматизации, использования средств, которые позволили бы снизить трудоемкость настройки системы, анализа изменений условий функционирования и потребностей пользователей и внесения изменений в правила функционирования ИС и их документирования.

Можно отметить следующие требования к информационным системам, обладающим высокой степенью адаптируемости: наличие средств динамической (в ходе эксплуатации системы) настройки, максимально снижающих трудоемкость и позволяющих выполнять эту работу пользователям системы, для которых должна быть обеспечена возможность работы в привычных терминах предметной области ИС, в которой они работают и являются экспертами. Максимальная степень адаптируемости ИС достигается, если она основана на *метамоделировании* [Лядова, 2008] и функционирует в режиме *интерпретации моделей*, которые могут изменяться в ходе эксплуатации системы, описывающих предметную область ИС и условия ее работы.

Исследования проводятся в рамках создания CASE-технологии METAS, упрощенная структура подсистем работы с документами которой показана на рис. 1.

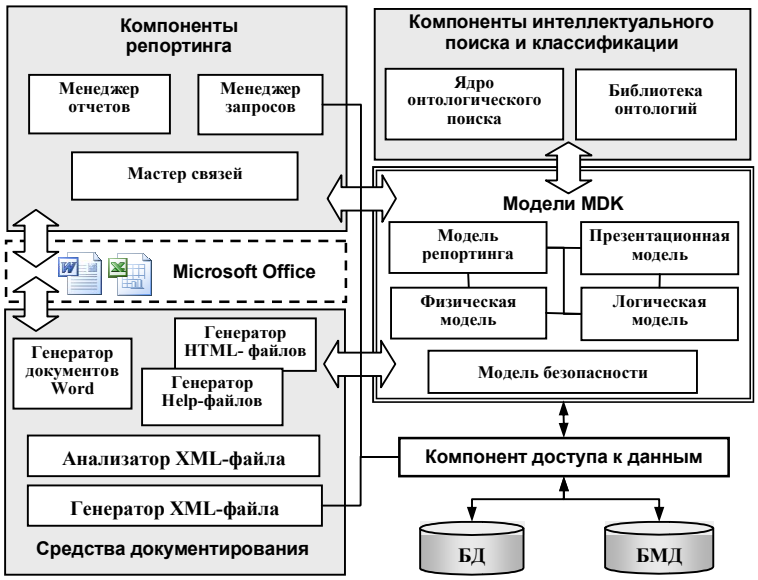

Рисунок 1 - Архитектура подсистемы управления документами CASE системы METAS

Основная часть документов, с которыми работает современное предприятие, – это неструктурированные документы. К этому типу относятся документы, представляющие

нормативно справочную информацию, распорядительные документы. Таким образом, основные знания о деятельности бизнес-системы располагаются именно в неструктурированных документах [Ландэ, 2005]. Если принять во внимание экспоненциальный рост количества документов [Ефремов, 2007] и отсутствие их семантического индексирования, проблема обработки неструктурированных документов становится чрезвычайно актуальной.

Обработка неструктурированных документов может потребоваться для решения следующих задач:

1) семантическое индексирование документов;
2) интеллектуальный поиск;
3) интеллектуальная классификация и каталогизация документов;
4) извлечение структурированной информации из неструктурированных документов;
5) автоматическое реферирование;
6) автоматизация работы аналитика при создании информационных систем;
7) ведение хронологии нормативно-справочной информации (НСИ).

Первые пять задач являются традиционными для обработки текстовой информации, но они могут получить более эффективное решение за счет применения подходов, основанных на явном представлении знаний. Две последние задачи являются менее традиционными, но не менее актуальными.

2. Модель электронного документа

Модель создает основу для описания документа, позволяющего подойти к решению задачи семантической индексации неструктурированных документов. Семантическая индексация является ключевым фактором в интеллектуальном управлении документами, т.к. позволяет обеспечить автоматическую классификацию документов, построение семантических связей между ними и автоматическое реферирование.

2.1. Гиперграфовая модель электронного документа

Электронный документ представляет собой *набор структурных элементов*, называемых в данной работе *фрагментами*. Примерами фрагментов могут служить таблица, заголовок, реквизиты углового бланка и т.д. Таким образом, документ может быть представлен тройкой вида:

$$d = (S(F, R), C).$$

Здесь $S(F, R)$ – ориентированный гиперграф, вершинам которого сопоставлены элементы множества F (множество F – это множество фрагментов документа, а R – это множество ребер графа, соответствующее связям между фрагментами); элементы множества C представляют информационное содержание документа (его контент).

Рассмотрим подробнее описанные выше множества.

Гиперграф $S(F, R)$ задает *отношение между фрагментами документа*. Ориентированность графа необходима, например, для отслеживания связей «часть-целое» между фрагментами. Вершины, входящие в ребро, пронумерованы, что позволяет установить порядок следования фрагментов в тексте документа. Очевидно, что ребро, включающее все вершины, соответствует документу в целом.

Фрагменты могут быть двух видов: *элементарные* фрагменты представляют простейшие неделимые элементы, такие как заголовок или дата составления документа, а *составные* фрагменты содержат в себе другие фрагменты.

Определим формально *фрагмент* как пару вида:

$$\boldsymbol{f} = (\boldsymbol{stat}, \boldsymbol{inf}), \boldsymbol{inf} = \begin{bmatrix} \boldsymbol{F}^*, \boldsymbol{F}^* \subseteq \boldsymbol{F}; \\ \boldsymbol{c}, \boldsymbol{c} \in \boldsymbol{C}. \end{bmatrix}$$

где *stat* – это статическая часть фрагмента, она может быть представлена текстом, изображением, ссылкой, каким-либо специальным символом, кроме того, здесь может содержаться и информация для представления фрагмента; *inf* – это часть фрагмента, которая либо указывает место для размещения элемента содержания c, $c \in C$, либо содержит множество фрагментов F^*. Пример графа документа d с фрагментами $F = \{f_1, f_2, f_3, f_4, f_5, f_6\}$ и ребрами $R = \{r_1, r_2, r_3\}$ показан на рис. 2.

Традиционно для представления документа используются обычные графы, чаще всего деревья (например, формат XML). Древовидная структура описания значительно упрощает работу с документом, но, вместе с тем, вносит и существенные ограничения. Выбор гиперграфа в качестве структуры данных для представления структуры документа обосновывается возможностями гиперграфов представлять произвольные связи между фрагментами документа и их множествами.

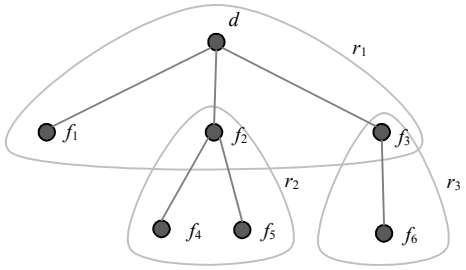

Рисунок 2 - Гиперграф документа

В описанных выше обозначениях *шаблон документа* можно определить как $t = (S(F, R), C_0)$, где C_0 – *первичный контент* (например, стандартные заголовки, реквизиты и т.д.).

Для выполнения различных операций над документом необходима операция выделения произвольной части документа, назовем ее *операцией получения диапазона Rng*. Входным параметром является произвольное множество вершин графа, а результатом – подграф, порожденный заданным в качестве параметра множеством вершин.

Операция расшифровки – наложение структуры на фрагмент (вершину графа).

Представленное выше формальное описание включает структуру и содержание документа. Помимо структуры и содержания в большинстве приложений важную роль играют *визуальное оформление документа* и его представление в определенном формате. В терминах предлагаемой модели *представление документа в определенном формате* должно определить функцию, задающую соответствие между фрагментами документа и некоторым множеством форматов, элементы которого задают правила отображения фрагмента документа.

Операция *поиска* применима к различным составляющим документа: структуре, содержанию и представлению. Результатом операции будут фрагменты документа, удовлетворяющие критериям поиска.

Представленная модель позволяет формализовать алгоритмы интеллектуальной обработки электронных документов в информационных системах. Модель предоставляет широкие возможности для интеграции документа с онтологическими ресурсами.

Практическая значимость модели находит подтверждение в современных форматах, таких как OpenXML и OpenDocument Format (ODF).

2.1. Описание документа с помощью онтологии

В настоящее время существуют различные подходы, модели и языки, ориентированные на интегрированное описание данных и знаний. Наиболее перспективным и универсальным, по мнению автора, представляется онтологический подход.

Согласно общепринятому определению под онтологией (в широком смысле) понимается база знаний специального типа, которая может «читаться» и пониматься, отчуждаться от разработчика и/или физически разделяться ее пользователями [Хорошевский, 2001]. Учитывая специфику решаемых в данной работе задач, можно конкретизировать понятие онтологии: онтология – это спецификация некоторой предметной области, которая включает в себя словарь терминов (понятий) предметной области и множество связей между ними, которые описывают, как эти термины соотносятся между собой.

В рассматриваемом подходе предполагается наличие трех *типов онтологий*:

- онтология предметной области конкретной информационной системы (ИС);
- онтология как база знаний (БЗ) интеллектуального агента;
- онтология как описание документа.

Рассмотрим назначение каждого из перечисленных типов онтологий.

Онтологии предметной области имеют наиболее типичное применение, они используются для описания понятий предметной области ИС. Например, школьное образование, социальная помощь гражданам или инновационное развитие регионов. В этой онтологии описывается связь понятий, языковые единицы для их выражения, аксиомы предметной области. Онтология предметной области используется для семантического индексирования и анализа всех документов системы [Weal, 2007].

Для анализа документов используется мультиагентный подход. Интеллектуальные агенты, руководствуясь *онтологией как базой знаний* (второй тип онтологий), производят поиск и анализ конкретных понятий документа. Каждая из вершин такой онтологии имеет определенный прототип, интерпретация которого известна агенту. Таким образом, агент использует онтологию как определенную программу своих действий. Вершинами онтологии данного типа могут являться понятия из онтологии предметной области.

Третий тип онтологий используется для *описания структуры и содержания документов*. Этот тип онтологий включает в себя два класса («плоскости») вершин. К первому классу относятся вершины, описывающие структуру документа. Например: таблица, дата, должность и т.д. (они представляют собой общие понятия, не зависящие от конкретной предметной области). Другим типом будут являться вершины, содержащие понятия документа. Первый тип вершин будем называть *структурные вершины*, второй тип – *семантический вершины*. Благодаря такому подходу из документа можно получить требуемые данные: известно, где искать данные, и как они могут быть интерпретированы.

Если представлять документ с использованием онтологий, то задача сопоставления онтологии и анализируемого документа сводится к задаче поиска понятий онтологии в документе. Как следствие, системе необходимо ответить на вопрос: описывает ли данная онтология документ или нет. На последний вопрос можно ответить утвердительно, если в процессе сопоставления в документе были найдены все понятия, включенные в онтологию. Таким образом, исходная задача сводится к задаче поиска в тексте документа общих понятий на основе формальных описаний. На основе онтологии может быть получен фрейм, слоты которого заполняются в процессе анализа документа. В качестве слотов фрейма выступают понятия онтологии, а значения этих фреймов заполняются данными анализируемого документа. Таким образом, из неструктурированного документа может быть получен структурированный документ-фрейм.

Онтологии располагаются на *трех уровнях репозитария*. На первом уровне расположены онтологии, описывающие объекты, используемые в конкретной системе и учитывающие ее особенности. На втором уровне описываются объекты, инвариантные к предметной области. Объекты третьего уровня описывают наиболее общие понятия и аксиомы, с помощью которых описываются объекты нижележащих уровней.

3. Агентный подход

Для решения проблемы выделения общих понятий на основе формальных описаний предлагается агентный подход [Тарасов, 2002]. Агент рассматривается как система, направленная на достижение определенной цели, способной к взаимодействию со средой и другими агентами. Для определения агента необходимо задать способ описания базы знаний, характера взаимодействия со средой и способа сотрудничества между агентами.

Одним из важнейших свойств агентов является *способность к взаимодействию*. Для каждой вершины онтологии, содержащей общее понятие (семантическая вершина), создается агент. Согласно принятой классификации агентов он является *интенциональным*. Данный агент нацелен на решение двух задач: весь имеющийся список шаблонов понятия он разбивает на отдельные компоненты и запускает более простых агентов для поиска структурных вершин (1), производит сборку результатов из всех списков, полученных агентами более низкого уровня (2).

Упомянутые выше агенты более низкого уровня являются *рефлекторными*. Они получают шаблон, и их целью становится отыскание в тексте фрагментов, попадающих под этот шаблон.

Важным вопросом становится коммуникация агентов. *Механизмы коммуникации агентов* делятся на непосредственные и опосредованные. Примером реализации *непосредственной*

коммуникации может служить модель взаимодействия «*заказчик – подрядчик*» (*contract network*). Механизм *опосредованной коммуникации* реализуется с помощью архитектуры «*доски объявлений*» (*blackboard*):

Модель «заказчик – подрядчик» предполагает деление всего множества агентов системы на два класса – класс *заказчиков* и класс *подрядчиков*. Суть данной модели взаимодействия заключается в решении различных задач путем направления их на выполнение наиболее подходящим для этого агентам. За распределение задач ответственны агенты – заказчики. Потенциальные подрядчики анализируют выставленные заказчиками заявки, анализируют их на предмет возможности реализации и, в случае положительного результата анализа, подают заявку заказчику. Модель «доска объявлений». Blackboard-архитектура основана на модели классной доски, на которой представлено текущее состояние системы, в рамках которой оперируют агенты. Агенты постоянно анализируют информацию на доске, пытаясь найти применение своим возможностям. В случае, если в некоторый момент времени агент обнаруживает возможность внесения своего вклада в процесс решения текущих задач, он оставляет на доске информацию о начале работы в данном направлении, а по окончанию работы помещает результат на доску. Учитывая особенности решаемой задачи, реализована комбинация двух моделей коммуникации «заказчик – подрядчик» и «доски объявлений».

Одним из наиболее важных вопросов в системе является вопрос представления БЗ агента. К настоящему моменту представление БЗ агента возможно тремя различными способами: с использованием онтологий, с помощью регулярных выражений и на базе продукций.

Преставление знаний агента с помощью онтологии – наиболее выразительный способ, использующий все преимущества явного представления знаний. Достоинством данного способа является то, что для «доказательства» вершины онтологии мы можем применить различные средства. Например, это может быть простое совпадение ключевой фразы или обращение к БД ИС. Онтологии позволяют описать различные ситуации в случае, если не удается найти точное соответствие. Мы может найти обобщающее или конкретизирующее понятие и т.п.

Вторым подходом является подход с использование *регулярных выражений*. Последние позволяют легко учитывать различные формы слова и работать с большими объемами информации. Однако необходимо учитывать, что иногда, особенно для неквалифицированных пользователей, задача правильного построения регулярного выражения становится достаточно сложной. С целью ее упрощения предполагается наличие в системе специального редактора, позволяющего работать с регулярными выражениями на естественном языке.

Недостатком регулярных выражений является то, что при поиске они не позволяют учитывать местонахождение искомого слова/фразы. Для устранения данного недостатка возможно совместное использование регулярных выражений и *правил продукционного типа*, которые являются третьим способом представления БЗ агента.

Продукции, в основном, используются для анализа структуры документа. Введены специальные понятия, которые могут быть использованы при задании условий. Например, правило находящее заголовки в тексте может быть сформулировано следующим образом: «***Если*** *(шрифт абзаца отличен от абзаца до и абзаца после) и (абзац выровнен по центру),* ***то*** *данный абзац является заголовком*».

Описанные выше средства представления БЗ агента позволяют конечному пользователю модифицировать существующих и добавлять новых агентов. Агентный подход оптимизировать процесс анализа документов путем параллельной работы агентов, что существенно для больших объемов документов.

4. Подсистема генерации отчетов

Средства репортинга METAS включают два основных компонента: «Менеджер запросов» и «Менеджер отчетов» [Lanin, 2008].

Одно из основных требований к подсистемам создания запросов и отчетов – это возможность их разработки пользователями-непрограммистами. Такое требование может быть выполнено только за счет введения дополнительного *семантического слоя*, основой которого могут быть метаданные, уже присутствующие в системе. Это дает пользователю возможность

работы с данными в соответствии с терминологией, принятой в конкретной предметной области ИС, позволяет абстрагироваться от физической структуры данных в БД.

Благодаря реализованной возможности хранения электронных документов в БД (создан *специальный тип данных*), сгенерированные отчеты становятся частью данных ИС, над ними можно выполнять разрешенные в системе операции.

5. Генерация документации

Этап создания документации является необходимым при разработке любой информационной системы. Руководство программиста необходимо для обеспечения сопровождения системы. Пользовательская документация необходима для эффективного обучения пользователей работе с новой ИС. Однако очень часто разработчики программных систем и комплексов игнорируют данный этап в связи с большим количеством времени, необходимого для создания качественной документации. Причём чем больше сложность системы, тем сложнее создание документации. Поэтому возникают проблемы при эксплуатации ИС, ее сопровождении. Особенно эти проблемы обостряются для систем, допускающих динамическую адаптацию, т.к. при настройке системы появляются расхождения ее «поведения» с описаниями, данными в документации. Однако при этом, если система работает в режиме интерпретации, программный код интерпретатора не изменяется, следовательно, модифицировать при настройке системы необходимо только пользовательскую документацию[Tsybin, 2008].

6. Процесс построения системы документов

Как было сказано выше, одно из применений разрабатываемой системы – поиск зависимостей и установление связей между документами, регламентирующими деятельность системы. В результате анализа должна быть построена *система взаимосвязанных документов*:

- относящихся к определенным направлениям деятельности бизнес-системы (к определенным понятиям, объектам предметной области);
- отражающих связи между этими понятиями (с каждым понятием может быть связан документ или совокупность документов, связи между документами отражают связи между понятиями);
- содержащих нормативную информацию, которая также может быть выделена на основе анализа содержания документов.

На основе построенной системы взаимосвязанных документов можно частично автоматизировать процесс анализа изменений предметной области и внесения изменений в модель предметной области ИС (т.е. реализовать поддержку процесса разработки и адаптации ИС). Таким образом, система управления документами становится не только «надстройкой» над ИС, позволяющей получать результаты обработки данных, хранящихся в БД ИС, в удобной для пользователей форме, но и становится основой средств разработки ИС.

Результатом анализа документов должно стать автоматическое построение онтологии, вершинами которой будут сами анализируемые документами и их понятия. Схематически процесс показан на рис. 3.

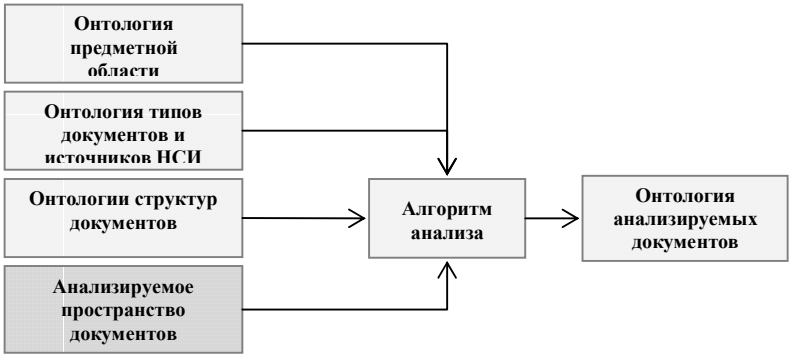

Рисунок 3 - Схема построения системы взаимосвязанных документов

Рассмотрим пример связи двух документов (рис. 4). Предположим, что «Документ 2» является обновленной версией «Документа 1». «Документ 1» состоит из двух разделов, в «Документе 2» появился новый раздел и один раздел был изменен. В результате анализа система свяжет два документа отношением «класс-подкласс», разделы документа будут связаны с документом отношением «часть-целое». У «Документа 2» вершина, связанная со вторым разделом, будет переопределять соответствующую вершину «Документа 1», первый раздел будет получен по иерархии наследования, а третий раздел будет собственным атрибутом.

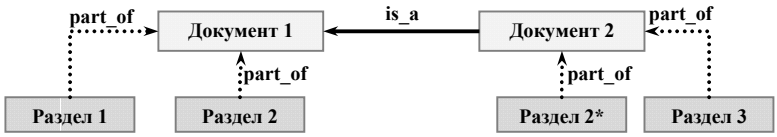

Рисунок 4 - Схема построения системы взаимосвязанных документов

Набор отношений может быть расширен пользователем путем описания шаблонов или реализации программных компонентом.

Кроме автоматического построения системы взаимосвязанных документов, возможно «ручное» создание схемы документов пользователем. В офисные приложения встраиваются специальные модули расширения, которые позволяют связать фрагменты документа с понятиями и отслеживать связи между ними.

Средства анализа документов могут быть использованы как для снижения трудоемкости работы пользователей с документами, так и для поддержки решения задачи анализа предметной области разработчиками. В данном случае предлагается глубокая интеграция функциональных подсистем ИС, включающих как средства разработки, так и средства, с которыми работают «конечные пользователи». Это дает возможность создания CASE-технологии, предназначенной для создания динамически настраиваемых ИС, обладающих уникальными возможностями адаптации к меняющимся условиям эксплуатации на основе средств поддержки «обратной связи» и интеллектуального анализа документов.

Заключение

Описанные выше средства реализованы в рамках создания CASE-технологии METAS, они прошли апробацию при реализации нескольких проектов. Средства анализа документов могут быть использованы как для снижения трудоемкости работы пользователей с документами, так и для поддержки решения задачи анализа предметной области разработчиками.

Библиографический список

[Ефремов, 2007] Ефремов В. Search 2.0: огонь по «хвостам» // Открытые системы. СУБД №08 (134), 2007.

[Ландэ, 2005] Ландэ Д. Поиск знаний в Internet. Профессиональная работа. М.: Издательский дом «Вильямс», 2005.

[Лядова, 2008] Лядова Л.Н. Метамоделирование и многоуровневые метаданные как основа технологии создания адаптируемых информационных систем // Advanced Studies in Software and Knowledge Engineering / International Book Series “Information Science & Computing”, Number 4. Volume 2, 2008. Institute of Information Theories and Applications FOI ITHEA, Sofia, 2008. P. 125-132.

[Тарасов, 2002] Тарасов В.Б. От многоагентных систем к интеллектуальным организациям: философия, психология, информатика. М.: Эдиториал, УРСС, 2002.

[Хорошевский, 2001] Хорошевский В.Ф., Гаврилова Т.А. Базы знаний интеллектуальных систем. СПб.: Питер, 2001.

[Lanin, 2008] Lanin V. Architecture and Implementation of Reporting Means in Adaptive Dynamically Extended Information Systems // International Journal “Information Technologies & Knowledge” / Sofia (Bulgaria) – Vol. 2/2008, Number 3, P. 273-277.

[Tsybin, 2008] Tsybin A., Lyadova L. Software Testing and Documenting Automation // International Journal “Information Technologies and Knowledge” / Sofia (Bulgaria) – Vol. 2/2008, Number 3. P. 267-272.

[Weal, 2007] Weal M.J., Kim S., Lewis P.H., Millard D.E., Sinclair P.A.S., De Roure D.C., Nigel R. Ontologies as facilitators for repurposing web documents / Shadbolt. Southampton, 2007.