МУЛЬТИАГЕНТАЯ СИСТЕМА ДЛЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДОКУМЕНТОВ

Вячеслав Ланин

***Аннотация**: В статье представлены промежуточные результаты реализации комплексного подхода к разработке подсистемы управления электронными документами в CASE-системе METAS, предназначенной для создания распределенных информационных систем, допускающих динамическую настройку на меняющиеся условия эксплуатации и потребности пользователей. Предлагается значительно увеличить эффективность работы с электронными документами за счет их автоматизированного интеллектуального анализа. В предлагаемом решении для анализа документов используются агентный и онтологический подходы. Онтологии позволяют в явном виде представить семантику и структуру документа. Использование агентов позволяет упростить процесс анализа, сделать его расширяемым и масштабируемым. Результаты интеллектуального поиска и обработки документов, получаемых из гетерогенных источников, могут быть использованы не только для автоматической классификации и каталогизации документов в информационной системе в удобной для пользователя форме, но и для снижения трудоемкости выполнения этапа анализа предметной области информационной системы, ее проектирования, а также для интеллектуализации процессов создания отчетных документов на основе информации, размещенной в базе данных системы.*

***Ключевые слова**: онтология, агент, мультиагентные системы, интеллектуальный поиск, анализ документов, адаптируемые информационные системы, CASE-технология.*

***ACM Classification Keywords**: D.2 Software Engineering: D.2.2 Design Tools and Techniques – Computer-aided software engineering (CASE); H.2 Database Management: H.2.3 Languages – Report writers; H.3.3 Information Search and Retrieval – Query formulation.*

***Conference**: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Введение

К настоящему времени разработано большое количество CASE-систем, автоматизирующих наиболее трудоемкие этапы разработки информационных систем (ИС), связанные с программированием бизнес-операций и созданием интерфейса. Самым продолжительным и трудоемким становится *этап анализа предметной* области, который обычно не автоматизируется CASE-системами. Таким образом, одним из перспективных направлений развития CASE-систем является автоматизация этого процесса. В CASE-системах, ориентированных на создание ИС с динамической адаптацией во время их использования, где стадия анализа предметной области «растягивается» на все время функционирования системы, эта задача становится особенно актуальной. Если учесть, что на стадии эксплуатации таких систем задача реинжиниринга возложена (хотя бы частично) на пользователей – специалистов в предметных областях, но не в области информационных технологий, то средства автоматизации анализа становятся важнейшими компонентами. Другими словами, если ставить задачу динамической настройки информационной системы на меняющиеся условия, то основа реализации средств ее динамической адаптации – средства реструктуризации данных в базе данных (БД) ИС. А эти средства позволяют вносить изменения в модель данных на основе результатов анализа предметной области, нормативно-справочных и распорядительных документов, регламентирующих деятельность в этой области. Отсюда следует необходимость поддержки в динамически адаптируемых системах одного из самых сложных и

трудоемких этапов разработки ИС – этапа анализа. Источником информации для анализа могут служить документы различного вида, т.к. деятельность любой бизнес-системы строится именно на основе нормативных документов. Поддержка бизнес-операций средствами ИС требует отражения в модели данных системы норм, закрепленных в нормативно-справочных данных, распорядительных документах, в виде ограничений, налагаемых на данные (атрибуты, свойства объектов предметной области, информация о которых хранится в БД, а также связи между ними) и операции, выполняемые над ними [1].

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 и ее БД, позволяющей получать результаты обработки данных, хранящихся в БД ИС, в удобной для пользователей форме, но и становится основой средств разработки ИС – средств реструктуризации данных.

Описание документов с помощью онтологий

Для повышения эффективности обработки электронный документ требует наличия метаданных, описывающих структуру и семантику данных. Одним из возможных подходов к описанию информации, заложенной в документе, является подход на основе онтологий. Под онтологией понимается база знаний специального типа, которая может «читаться» и пониматься, отчуждаться от разработчика и/или физически разделяться ее пользователями [4]. Онтологический подход обладает такими преимуществами, как

- удобство восприятия человеком;
- отсутствие необходимости в специальной квалификации пользователя при разработке онтологии;
- возможность описания одного документа различными онтологиями.

В качестве подхода к решению описанной выше задачи был выбран онтологический подход [1], в котором онтология описывает как структуру, так и содержание документа. В соответствии с предлагаемым подходом *онтология используется для описания семантики данных документа и его структуры.* Учитывая специфику решаемых в данной работе задач, конкретизируем понятие онтологии: будем считать, что *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 конкретной предметной области.

Для построения иерархии понятий онтологии используются следующие базовые типы отношений:

- “is_a” («экземпляр – класс», гипонимия);
- “part_of” («часть – целое», меронимия);
- “synonym_of” (синонимия).

Следует учесть, что данные типы отношений являются базовыми и не зависят от онтологии, но необходимо предоставить пользователю возможность добавления новых отношений, которые бы учитывали специфику описываемой предметной области.

Ректору Иванову И.И.
студентки Сидоровой А.А.

заявление.

Прошу освободить меня от занятий 10.02.08 для участия в спортивных соревнованиях.

10.02.2008 Сидорова А.А.

а)

б)

Рис. 1. Пример простого документа «Заявление» (а) и онтологии, описывающей класс документов «Заявление» (б)

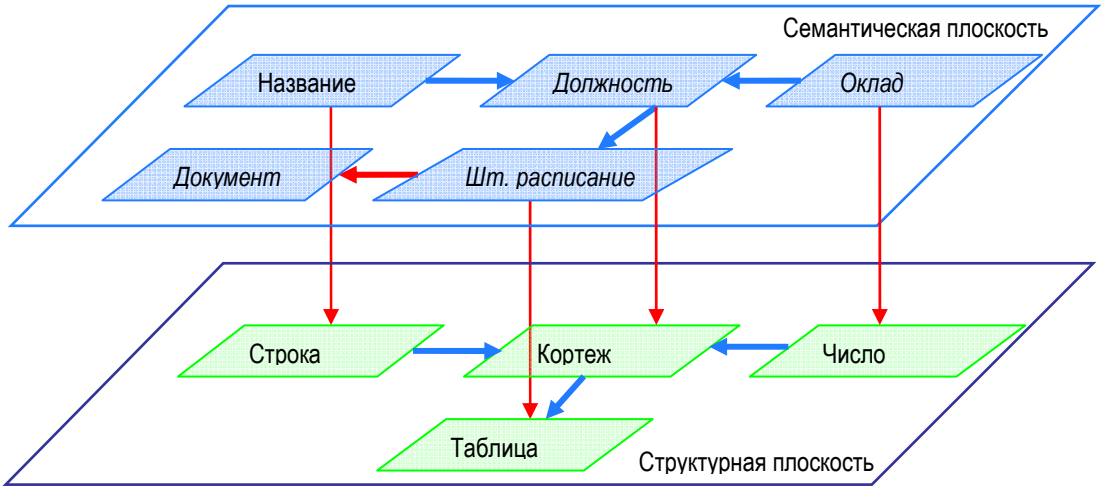
Приказ № 1
от 01.11.2005

Установить с 01.12.2005 следующее штатное расписание:

Ассистент	4 000 рублей
Старший преподаватель	6 000 рублей
Доцент	10 000 рублей
Заведующий кафедрой	15 000 рублей

Ректор Иванов И.И.

а)



б)

Рис. 2. Пример документа «Приказ» (а) и разбиение вершин онтологии для документа на две плоскости (б)

Кроме отношений онтология включает в себя два типа вершин. К первому типу отнесем 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*. На рис. 1 структурные вершины имеют темный оттенок, а семантические вершины изображены более светлым оттенком.

Фактически в данном контексте *онтология – это иерархическая понятийная основа рассматриваемой предметной области*. Онтология документа используется для анализа документа, благодаря ей 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 имеющегося 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. Прежде, чем производить поиск вершин, содержащих понятия документа, необходимо провести поиск вершин, описывающих структуру документа. Таким образом, исходная задача сводится к задаче поиска в тексте документа общих понятий на основе формальных описаний.

В приведенном примере (рис. 2, б) вершины онтологии разбиты на две плоскости, что учитывается при сопоставлении документа (рис. 2, а) и его онтологии.

Агентный подход к анализу документов

К процессу поиска документов предъявляется ряд требований:

- высокая скорость обработки больших объемов данных;
- отказоустойчивость;
- масштабируемость и
- настраиваемость на потребности пользователей и меняющиеся условия.

Для решения проблемы выделения общих понятий на основе формальных описаний предлагается агентный подход [2]. Здесь *под агентом понимается система, направленная на достижение определенной цели, способная к взаимодействию со средой и другими агентами* [3]. Данный подход будет удовлетворять требованиям, предъявляемым к процессу поиска, если при построении системы будут реализованы все преимущества мультиагентных систем.

При использовании данного подхода *для каждой вершины онтологии, содержащей общее понятие, создается агент*, *который проводит поиск данного конкретного понятия*. Для признания агента интеллектуальным необходимым условием является *наличие у него базы знаний*. Таким образом, чтобы определить агентов, действующих в системе, необходимо выбрать способ для описания базы знаний (БЗ), характер взаимодействия со средой и сотрудничества. Средствам представления базы знаний агентов посвящен следующий раздел статьи.

Одним из важнейших свойств агентов является *социальность или способность к взаимодействию* [2]. Как было сказано ранее, 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:

1. Весь имеющийся список шаблонов понятия он разбивает на отдельные компоненты и запускает более простых агентов для поиска структурных вершин.
2. Производит сборку результатов из всех списков, полученных агентами более низкого уровня.

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». Данная модель 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».

Архитектура мультиагентной системы и процесс анализа документа представлены на рис. 3.

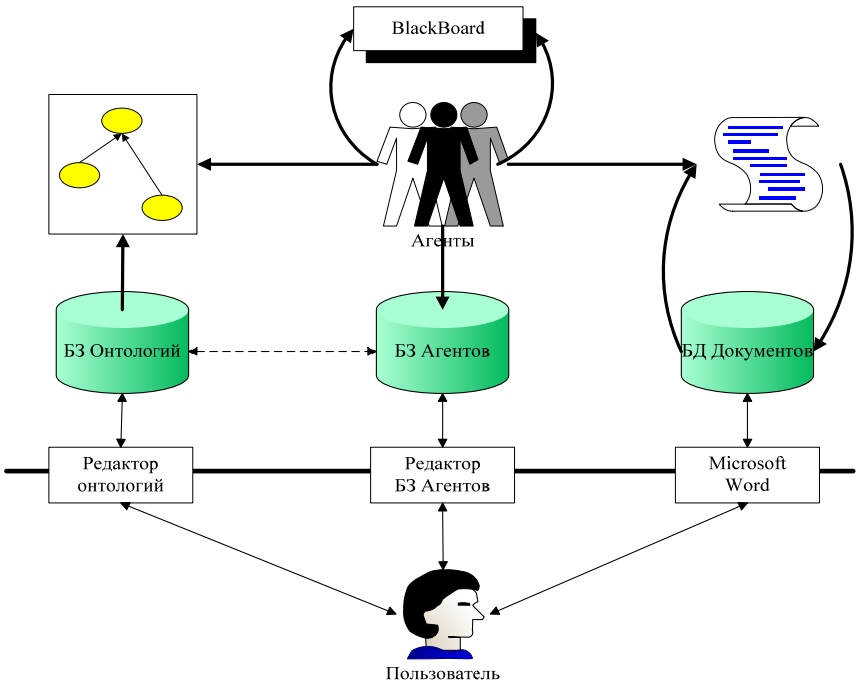


Рис. 3. Архитектура системы SemanticDoc

Представление базы знаний агентов

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 (рис. 4). 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.

Содержимое анализируемого документа представлено в виде специальной *объектной модели*, за основу которой была взята объектная модель документа Microsoft Word. Для доступа к этой объектной модели разработаны API-функции, позволяющие оперировать одинаковыми понятиями при работе с документами в различных форматах. В состав API-функций включены функции по синтаксическому разбору приложений, функции для вычисления различных метрик между понятиями, функции для извлечения информации о структуре документа. Если для поиска понятия вершины необходимы дополнительные действия, они могут быть описаны с помощью скрипта с использованием упомянутых выше API-функций. В скрипте также могут быть использованы обращения к объектной модели самой ИС.

Вторым подходом является подход с использование *регулярных выражений*. Последние позволяют легко учитывать различные формы слова и работать с большими объемами информации [5]. 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. Например, эквивалентом к «\d{5}» является «пятизначное число» и т.д. Кроме того, желательна реализация функции построения регулярного выражения «по образцу». Это означает, что по примерам, приведенным пользователем, возможно автоматическое построение регулярного выражения. Например, пользователь в качестве примеров предложил две даты: «1.12.08» и «15.07.2006». Система должна построить регулярное выражение, которое бы соответствовало обоим форматам представления дат: «(\d{1,2}).(\d{1,2}).((\d{4})|(\d{2}))».

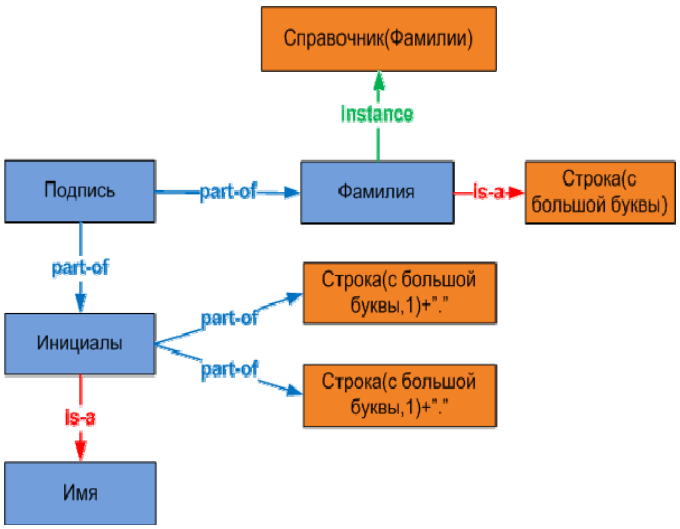


Рис. 4. Представление базы знаний агента с помощью онтологии

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.*

Заключение

На данный момент результатом работы стала реализация на платформе .NET системы SemanticDoc, представляющей собой мультиагентную систему, которая проводит сопоставление документа и онтологии.

В информационном поиске для сравнения качества результатов были введены две характеристики: *точность* (*precision*) и *полнота* (*recall*) [6]. Подобные характеристики можно ввести и для системы сопоставления документа и онтологии. Под *точностью* (P) будем понимать долю правильно проведенных соответствий документа и онтологии по отношению ко всем сделанным системой соответствиям. Под *полнотой* (R) – долю правильно проведенных соответствий по отношению ко всем соответствиям документа и онтологии.

Пусть N – число существующих соответствий между документом и онтологией, M – число проведенных системой сопоставлений, A – число правильно проведенных системой сопоставлений. Тогда:

$$P = A/M \quad \text{и} \quad R = A/N.$$

Обычно эти два критерия «конфликтуют» и на практике стопроцентная точность и полнота недостижимы.

Работы по оценке пока не проводились, следующим этапом исследование станет оценка величин P и R при проведении экспериментов на реальных документах.

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, 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 «обратной связи» и интеллектуального анализа документов.

В рамках данной работы разрабатывается также формальная модель электронного документа и онтологии применительно к решаемой задаче, а на ее основе уточняется существующая объектная модель ИС, метаданных и алгоритмы управления документами.

Благодарности

Работа выполнена при поддержке гранта РФФИ № 08-07-90006-Бел_а.

Библиографический список

[1] Ланин В.В. Интеллектуальное управление документами как основа технологии создания адаптируемых информационных систем // Труды международной научно-технической конференций «Интеллектуальные системы» (AIS'07). Т. 2 / М.: Физматлит, 2007. С. 334-339.

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

[3] Рассел С. Искусственный интеллект: современный подход. М.: Издательский дом «Вильямс», 2006.

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

[5] Фридл Дж. Регулярные выражения. СПб.: Питер, 2003.

[6] 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.

Сведения об авторе

Вячеслав Ланин *– Пермский государственный университет, аспирант кафедры математического обеспечения вычислительных систем; Россия, г. Пермь, 614990, ул. Букирева, д. 15; e-mail: lanin@psu.ru*