

Онтологии как Основа Функционирования Систем Обработки Электронных Документов

Ланин Вячеслав

*Кафедра Математического Обеспечения Вычислительных Систем, Пермский
Государственный Университет, ул. Букирева, д. 15, г. Пермь, 614990, Россия*

lanin@psu.ru

Аннотация. В работе описывается подход к анализу неструктурированных электронных документов с использованием онтологий. С помощью данного подхода может быть решен широкий класс задач, включающий в себя как традиционные задачи обработки текстов, так и средства интеллектуального анализа документов при создании моделей предметных областей при проектировании информационных систем.

Ключевые слова: электронный документ, онтология, агенты, интеллектуальный поиск.

1 Введение

Анализ темпов роста количества электронных документов и методов их обработки наглядно показывает, что традиционные механизмы работы с электронными документами не удовлетворяют потребностям современного пользователя. Необходимы новые подходы к интеллектуальному поиску и анализу электронных документов, их интеграции в информационные системы. Наиболее актуальными на данный момент проблемами являются:

- Экспоненциальный рост числа документов является причиной все более и более возрастающей трудности поиска необходимых документов и организации их в виде структурированных по смыслу хранилищ [1]. При увеличении пространства поиска пропорционально возрастает и количество документов в отклике поисковой системы.
- Отсутствие стандартизированных механизмов семантического индексирования также пагубно влияет на эффективность работы с электронными документами. Большинство современных технологий подготовки и работы с документами (текстовые редакторы, язык HTML) ориентированы на организацию удобной работы с информацией для человека.
- Неструктурированный характер информации большинства электронных документов не позволяющий применить традиционные механизмы ее обработки и анализа. Неструктурированная информация составляет значительную часть современных электронных документов, основные знания располагаются именно в таких документах.

Для решения перечисленных выше проблем необходимо расширить понятие традиционного документа: с документом необходимо связать метаданные, позволяющие интерпретировать и обрабатывать хранящуюся в этом документе информацию, т.е. включить в документ информацию, описывающую структуру и семантику его содержания.

2 Постановка задачи

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 эксплуатации, 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 экспертов – 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 [3].

CASE-технология METAS, разработанная сотрудниками АНО «Институт компьютеринга», обеспечивает возможность эффективного использования предложенных в данной статье средств как при создании информационной системы, так и в ходе ее эксплуатации пользователями, а также при выполнении анализа документов для осуществления динамической настройки системы. Расширение CASE-системы средствами анализа слабоструктурированных документов, классифицированных и каталогизированных на основе созданных разработчиками информационной системы и ее пользователями онтологий, может существенно снизить трудоемкость сопровождения системы, ее настройки на меняющиеся в ходе эксплуатации условия и информационные потребности пользователей.

3 Описание документа с помощью онтологии

Методы искусственного интеллекта, как правило, используются для решения трудно формализуемых задач, постановка которых проста и понятна для человека, но при разработке алгоритмов их решения возникают трудности. Одна из таких задач – работа с документами в информационных системах: их поиск и каталогизация, анализ и извлечение информации.

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 [4]. 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.

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.

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 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.

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.

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 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. 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.

4 Агентный подход

Для решения проблемы выделения общих понятий на основе формальных описаний предлагается агентный подход [4]. 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 *социальность или способность к взаимодействию* [4]. 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. 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 удастся найти точное соответствие. Мы можем найти обобщающее или конкретизирующее понятие и т.п.

Для обеспечения работы с различными типами документов в систему необходимо включить компоненты, обеспечивающие унифицированный доступ к документам в различных форматах. Такую функциональность можно реализовать за счет «устанавливаемых драйверов форматов» – компонентов, реализующих заранее определенный интерфейс, обеспечивающих доступ к документам определенного формата. Реализация таких «драйверов» может быть основана на использовании шаблонов, образцов, позволяющих распознать структуру документа. 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 использованием *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. Данные особенности делают подход расширяемым и масштабируемым.

5 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.

6 Заключение

Предложенный подход решает следующие задачи:

- семантическое индексирование документов и интеллектуальный поиск данных, соответствующих запросам пользователей и специфике предметной области;
- извлечение информации из неструктурированных документов;
- интеллектуальная классификация и каталогизация и автоматическое реферирование найденных документов;
- ведение хронологии электронных документов.

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 «обратной связи» и интеллектуального анализа документов.

7 Благодарности

Работа выполнена при поддержке грантов РФФИ № 08-07-90006-Бел_а и РГНФ № 09-02-00373В/И.

Литература

- [1] Ефремов В. Search 2.0: огонь по «хвостам» // Открытые системы. СУБД №08 (134), 2007.
- [2] Ландэ Д. Поиск знаний в Internet. Профессиональная работа. М.: Издательский дом «Вильямс», 2005.
- [3] Ланин В.В. Интеллектуальное управление документами как основа технологии создания адаптируемых информационных систем // Труды международной научно-технической конференции «Интеллектуальные системы» (AIS'07). Т. 2 / М.: Физматлит, 2007. С. 334-339.
- [4] Тарасов В.Б. От многоагентных систем к интеллектуальным организациям: философия, психология, информатика. М.: Эдиториал, УРСС, 2002.
- [5] Федотов А.М., Барахнин В.Б. Ресурсы интернета как объект научного исследования [Электронный ресурс]. – 2007. – Режим доступа: <http://www.rfbr.ru/pics/28320ref/file.pdf>.
- [6] Хорошевский В.Ф., Гаврилова Т.А. Базы знаний интеллектуальных систем. СПб.: Питер, 2001.
- [7] Черняк Л. Корпоративный поиск 2.0 // СУБД. – 2007. – №07 (133).
- [8] 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.