

А.В. Палагин, Н.Г. Петренко

К ПРОЕКТИРОВАНИЮ ОНТОЛОГОУПРАВЛЯЕМОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ С ОБРАБОТКОЙ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ОБЪЕКТОВ

Abstract: *In this work the approach to the formalized projection ontology-driven an intelligence system with natural language processing objects is considered. Primal problems of the analysis and the synthesis, solved on all design stages are briefly considered. Also the example of solution of the task of classification of text documents is considered.*

Keywords: *domain ontology, natural language processing.*

Анотація: *У роботі розглянуто підхід до формалізованого проектування онтологокерованої інформаційної системи обробки природномовних об'єктів. Коротко розглянуто основні задачі аналізу та синтезу, вирішені на всіх етапах проектування. Також розглянуто приклад вирішення задачі класифікації текстових документів.*

Ключові слова: *онтологія прикладної області, комп'ютерна обробка природномовних об'єктів.*

Аннотация: *В работе рассмотрен подход к формализованному проектированию онтологоуправляемой информационной системы с обработкой естественно-языковых объектов. Кратко рассмотрены основные задачи анализа и синтеза, решаемые на всех этапах проектирования. Также рассмотрен пример решения задачи классификации текстовых документов.*

Ключевые слова: *онтология прикладной области, компьютерная обработка естественно-языковых объектов.*

1. Введение

Одной из важных ветвей современного развития интеллектуальных информационных систем являются онтологоуправляемые информационные системы (ОУИС). Построение последних тесно связано с разработкой теоретических основ и методологии проектирования, включающих формальный подход, фундаментальные принципы и механизмы, обобщённую архитектуру и структуру системы, формальную модель и методологию проектирования онтологии предметной области (ПрО), формальную модель представления знаний, обобщённые алгоритмы процедур обработки знаний и др. В свою очередь, каждая из перечисленных составляющих общей методологии проектирования представляет собой сложную информационно-алгоритмическую структуру, например разработка онтологии ПрО тесно связана с концептуализацией онтологических категорий, разработкой и усовершенствованием иерархических структур сущностей на всех уровнях, построением формальной системы аксиом и ограничений. Комплексное решение указанных задач проектирования должно повысить роль онтологических (концептуальных) знаний при решении конкретных задач в прикладных областях.

2. Постановка задачи

При проектировании знаниеориентированных информационных систем (какой является и ОУИС) существенным является выбор как формально-логического представления знаний, так и источников приобретения и пополнения знаний. В настоящее время признанным “де-факто”, наиболее обширным и общедоступным источником знаний является Интернет-пространство, с его естественным способом представления информации. Указанный способ предопределил появление многочисленных информационных технологий обработки знаний, содержащихся в естественно-языковых объектах (ЕЯО), в том числе и NLP–технологии (Natural Language Processing).

В известных методах формализованного проектирования информационных систем (ИС) [1] сам процесс проектирования представляется в виде последовательности этапов (в основном си-

стемного, алгоритмического и логического), на каждом из которых проект представлен совокупностью математических моделей, описывающих различные её части. Указанная совокупность математических моделей тесно связана с системой взаимосвязанных алгоритмов, которые, в свою очередь, описывают соответствующее множество решаемых задач и в своей совокупности представляют общий алгоритм проектирования ИС. Применительно к проектированию ОУИС обработки знаний, содержащихся в ЕЯО, обобщённая последовательность решаемых задач анализа и синтеза на всех этапах следующая.

3. Проектирование ОУИС

Известно, что проектируемые средства информатики в соответствии с их проблемной ориентацией базируются на некоторой совокупности фундаментальных принципов, методик и алгоритмов. Также известны многочисленные разработки в области обработки ЕЯО [2-6, 19, 20 и др.], в которой ввиду огромной сложности решаемых задач принимаются те или иные ограничения, снижающие эффективность средств обработки. Авторами были рассмотрены и проанализированы работы и проекты, в которых формальная онтология играет существенную роль. В результате были выявлены главные недостатки, снижающие эффективность реализации отдельных процедур и конечных результатов в целом.

В качестве фундаментального механизма при разработке средств интерпретации ЕЯО в работе предлагается модификация известного логико-информационного подхода (ЛИП) [7],

(1)

где

- отображение i -ого уровня обработки ЕЯО;
- множество обрабатываемых на i -ом уровне объектов ЕЯО;
- множество слов, описывающих на i -ом уровне;
- множество синтаксических структур, связывающих на i -ом уровне;
- множество семантических структур, соответствующих на i -ом уровне;
- множество фрагментов онтологических структур, участвующих в формировании на i -ом уровне, композиция которых в идеале составляет ;
- множество информационно-кодовых представлений (индексов) на i -ом уровне.

В соответствии с (1) оптимальной считается такая структурная реализация модели ОУИС, для которой в соответствии с принятыми критериями найдены оптимальное количество уровней и оптимальные соотношения между обобщёнными характеристиками компонент на каждом уровне, а также соответствующими характеристиками компонент соседних уровней.

Компонента формальной модели ОИП в общем случае представляет интеграцию онтологических общелингвистических знаний и знаний предметных областей. При этом общелингвистические знания будем рассматривать в качестве отдельной предметной области. Тогда объём знаний W в предметных областях можно оценить через характеристики (параметры) их формально-онтологических представлений. В частности, при представлении онтологическим графом (ОГ) (без

учёта типов отношений и сложности функций интерпретации) величина W может характеризоваться числом вершин ОГ. Как показано в [10], в случае простой древовидной структуры это число может быть выражено формулой

(2)

где O_n - онтограф n -ой предметной области, n - степень вершины, равная числу исходящих из неё рёбер, R_n - количество уровней ОГ, i - номер вершины на соответствующем уровне ОГ.

Учёт типов отношений и сложность функций интерпретации приводит к ОГ с взвешенными вершинами и ребрами. Выражение (2) при этом сводится к виду

(3)

где w_i и w_r - значения весовых функций соответствующих отношений и функций интерпретации, приписанные вершинам (v_i) и ребрам (e_r) ОГ. Выражение (3) даёт полную оценку сложности ОГ, а отношение характеризует среднюю плотность взвешенного ОГ.

Актуальной для направления *онтологического инжиниринга* является проблема разработки методологии проектирования формальной онтологии ПрО. Несмотря на то, что обобщённая схема такой методологии была опубликована в [21] ещё в 1996 году, до сих пор не существует общепринятой методологии. Остаётся открытым вопрос, поставленный в [11] - "...появится ли конструктивная и работающая теория разработки онтологий, или по-прежнему каждый аналитик будет идти методом проб и ошибок, создавая сложнейшие и головоломные онтологические структуры, отражающие лабиринты профессиональных знаний?". Методология проектирования формальной онтологии ПрО представляет собой сложный инструментальный комплекс, состоящий из оригинальных инструментальных средств (ОИС) (возможно в сочетании с известными) и алгоритмов реализации этапов проектирования в ручном режиме, объединённых между собой программной оболочкой, которая обеспечивает разработчику онтологии автоматизированный и интерактивный процесс проектирования. Она предназначена, в первую очередь, для повышения уровня формализации и автоматизации проектирования онтологии ПрО, при этом обеспечивая выполнение условий целостности, непротиворечивости, интегрированности, операбельности и др. для создаваемых онтологий.

В качестве оригинальных инструментальных средств предлагаются прикладные программы "Concepts and Axioms Formation" и "Relations Formation". Оба средства выполняют обработку лингвистического корпуса текстовой информации из заданной предметной области и затем формируют соответственно базовые списки концептов и аксиом и семантических отношений в заданной ПрО.

В качестве известных автоматизированных инструментальных средств (АИС) предлагается использовать Protégé-2000 или OntoStudio как системы, формирующие описания проектируемой онтологии в двух, наиболее популярных подсистемах логики – соответственно фреймовом представлении и исчислении предикатов первого порядка.

Работа с большими массивами текстовой информации предполагает их поиск, первичную обработку и хранение. В качестве исполнителя указанных процедур могут быть привлечены известные популярные поисковые машины. Альтернативным вариантом может быть ориентация на знаниеориентированную поисковую систему (ЗОПС), описанную в [12]. Одним из основных пре-

имущества её использования можно отметить построение онтологии найденных в Интернет текстовых документов и их первичную обработку.

Ручные и интерактивные процедуры проектирования онтологии ПрО поддерживаются программной оболочкой, которая обеспечивает “дружественный” интерфейс с разработчиком.

Обобщённая блок-схема инструментального комплекса, реализующего алгоритм формализованного проектирования онтологии ПрО, представлена на рис.1.

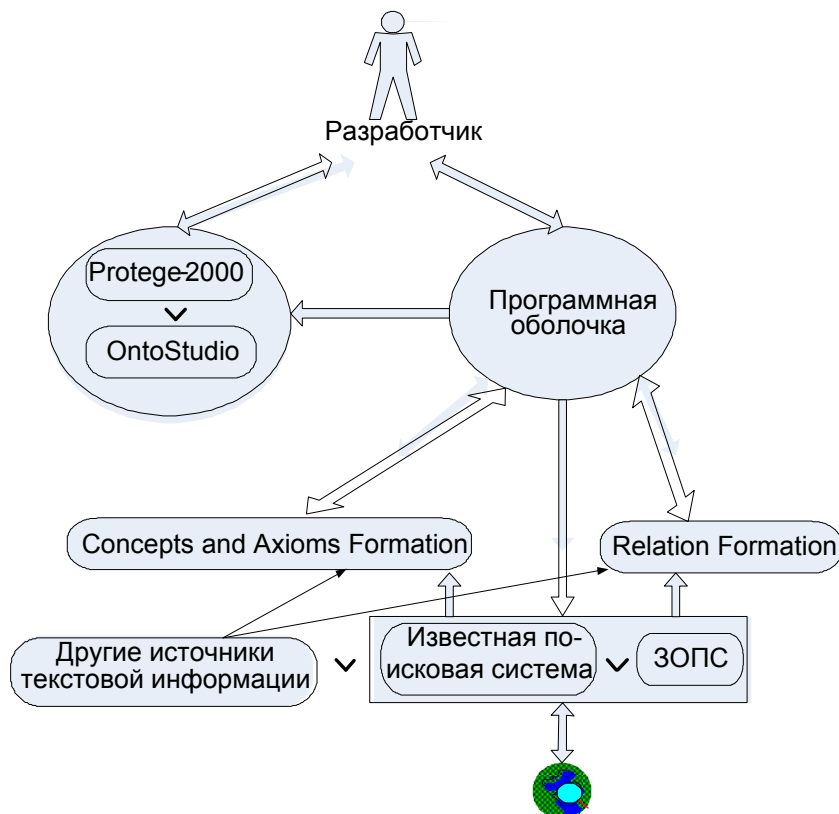


Рис.1. Обобщённая блок-схема инструментального комплекса

По своей функциональной полноте различают три вида онтологий ПрО [17]: полная (или строгая), простая и множество промежуточных (или неполных) онтологий.

Строгая или полная онтология ПрО (в терминах работы [11] - “весомая” онтология) – это такая онтология, в которой множества концептов и концептуальных отношений максимально полные, а к аксиомам и определениям добавляются ограничения. При этом аксиомы, определения и ограничения представлены на некотором формальном языке, доступном для их интерпретации компьютером. Схема формальной модели полной онтологии описывается формулой:

(4)

Простая онтология ПрО (в терминах работы [11] - “легкая” онтология) – это такая онтология, в которой множество R или пусто, или определено одним отношением, а множество F - пусто. При этом аксиоматизация может быть представлена определениями терминов словаря ПрО на естественном языке (ЕЯ).

Множество промежуточных или неполных онтологий ПрО описываются различными сочетаниями мощностей множеств R и F , а аксиоматизация выполнена на ЕЯ.

Такая схема классификации по функциональному признаку согласуется с описанием [22]: “Онтология или концептуальная модель предметной области состоит из иерархии понятий предметной области, связей между ними и законов, которые действуют в рамках этой модели”, а также с работами [11, 13, 14].

4. Архитектура ОУИС

Архитектура ОУИС, разработанная в соответствии с ОИП-моделью для проблемной области обработки ЕЯТ, представлена на рис.2. На нём приняты следующие обозначения:

ЛБД представляет набор таблиц,

ЯОКМ представляет лингвистическую онтологию,

Грамматический анализатор.

Семантический анализатор.

СИИ

Цепочки преобразования информации и , по сути, представляют (соответственно) базовые процедуры анализа и понимания ЕЯТ, средствами интерпретации которых являются грамматический и семантический анализаторы и семантико-информационный интерпретатор.

Модули базы знаний предметной области и прикладного процессинга предназначены для решения конкретных задач пользователя. Причём для простых задач обработки ЕЯО реализующие их алгоритмы могут войти в состав ОУИС с обработкой ЕЯО.

Рассмотрим работу системы (рис.2) на примере решения задачи классификации найденных в сети Интернет текстовых документов как таковой, алгоритмы которой лежат в основе более сложных задач обработки текстовой информации (в том числе поиска релевантной информации, построения персонифицированных баз знаний и др.). При этом блок прикладного процессинга представлен упомянутой выше ЗОПС.

Формальную постановку задачи классификации представим следующим образом:

- предполагается, что ЗОПС находит в сети некоторое множество текстовых документов, представленных для классификации в заданной ПрО;

- всё множество D разбивается на непересекающиеся подмножества классов

Задачей классификации является определение класса, к которому принадлежит данный документ.

Исходя из архитектурно-структурной организации ОУИС, представленной на рис.2, задано базу знаний ПрО (в нашем случае достаточно и онтологии ПрО) и базу знаний лексики ЕЯ. При этом заметим, что:

- если множество лексем, описывающих заданную ПрО, не является подмножеством множества лексем ЯОКМ, то необходимо создать лингвистическую онтологию заданной ПрО;

- мы сознательно опускаем описание работы той части алгоритма, в которой реализуется снятие неоднозначностей при анализе текстовых документов;

- в блоке прикладного процессинга имеются алгоритмы построения онтологий текстового документа, для классов С и семантическая память большого объема для хранения классифицированных документов.

Последовательность шагов решения задачи классификации:

- 1.
- 2.
- 3.
- 4.
- 5.

4. Выводы

В работе рассмотрен подход к формализованному проектированию онтологоуправляемой информационной системы с обработкой естественно-языковых объектов. Кратко рассмотрены основные задачи анализа и синтеза, решаемые на всех этапах проектирования: разработка онтолого-инфологического подхода к проектированию, предложена формальная модель онтологии ПрО и интегрированная среда для её проектирования, с учётом результатов, полученных в предыдущих работах [9, 15-18], синтезирована структура проектируемой ОУИС. В заключение рассмотрен пример решения задачи классификации текстовых документов, при этом вычислительные процедуры реализуются на синтезированной структуре ОУИС и с применением в качестве блока прикладного процессинга знаниеориентированной поисковой машины.

Основные результаты проведенного исследования предполагают выполнение дальнейших работ по детализации схемы общего алгоритма проектирования ОУИС с обработкой ЕЯО и формализованной методики в целом. Кроме того, предстоит разработать формализованный переход от NLP-технологии к технологии Knowledge processing для решения прикладных задач работы со знаниями.

Список литературы

1. Капитонова Ю.В., Летичевский А.А. Математическая теория проектирования вычислительных систем. – М.: Наука. Гл.ред.физ.-мат.лит., 1988. – 296с..
2. Апресян Ю.Д. и др. Лингвистический процессор для сложных информационных систем. - М.: Наука, 1992. – 287с.
3. Широков В.А., Бугайов О.В., Грязнухина Т.О. та ін. Корпусна лінгвістика. – К.: Довіра, 2005. – 471с.
4. Гладун В.П. Процессы формирования новых знаний. - София: СД "Педагог 6", 1994. -192с.
5. Замаруева И.В. Об одном подходе к компьютерному моделированию процесса понимания естественно-языковых текстов. – Труды VI Межд. конф. "ЗНАНИЕ-Диалог-Решение", KDS-97. – Ялта: 15-20 сентября. - 1997г. - С.241-248.
6. Нариньяни А.С. Кентавр по имени ТЕОН: Тезаурус + Онтология, Российский НИИ искусственного интеллекта - /Available at <http://www.artint.ru/articles/narin/teon.htm>.
7. Палагин А.В. К решению основной задачи эмуляции // УСиМ. – 1980. – №3. – С.24-28.
8. Реконфигурируемые вычислительные системы: Основы и приложения. / А.В. Палагин, В.Н. Опанасенко. – К.: Просвита, - 2006. – 280с.
9. Палагин А.В., Петренко Н.Г. Архитектура онтологоуправляемой ИС с обработкой ЕЯТ//Матеріали XIV міжнародної конференції з автоматичного управління "Автоматика - 2007" – Севастополь, 10-14 вересня. – 2007. – Ч.2 - С.166.
10. Палагин А.В., Петренко Н.Г. К вопросу системно-онтологической интеграции знаний предметной области //Математические машины и системы. – 2007. - №3, 4. - С.63-75.
11. Гаврилова Т.А. – Онтологический инжиниринг. - /Available at http://www.kmtec.ru/publications/library/authors/ontolog_engeneering.shtml.
12. Севрук О.О., Петренко М.Г. Знання-орієнтована пошукова система на основі мовно-онтологічної картини світу //Тези доповідей XIII міжнародної конференції з автоматичного управління "Автоматика-2006". – Вінниця. – 2006. - 25-28 вересня. – С.413.
13. Палагин А.В., Яковлев Ю.С. Системная интеграция средств компьютерной техники. – Винница: «УНІВЕРСУМ-Вінниця», 2005. – 680 с.
14. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2001. – 384с.
15. Палагин О.В., Петренко М.Г. Модель категоріального рівня мовно-онтологічної картини світу //Математичні машини і системи. – 2006. - №3. - С.91-104.

16. Палагин А.В., Петренко Н.Г., Севрук А.О., Михайлюк А.В. Интеллектуальные технологии обработки естественно-языковых объектов и системно-онтологическая интеграция знаний//Тези доповідей міжнародної наукової конференції MegaLing'2007 – Україна, Крим, Партеніт, 24-28 вересня 2006 року. - Сімферополь: Вид-во "ДиАйПи", 2007. – С.280.
17. Палагин О.В., Петренко М.Г. Архітектурно-онтологічні принципи розбудови інтелектуальних інформаційних систем //Математичні машини і системи. – 2006. - №4. - С.15-20.
18. Палагин О.В., Петренко М.Г. Розбудова абстрактної моделі мовно-онтологічної інформаційної системи //Математичні машини і системи. – 2007. - №1. - С.42-50.
19. Sowa, John F. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA, ©2000. - 594pp.
20. Fellbaum, Christiane, ed. (1998) *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA.
21. Uschold, M. and Gruninger, M. (1996). Ontologies: Principles, Methods and Applications. *KnowledgeEngineering Review* 11(2). - pp.93-136.
22. Guarino N. Formal Ontology and Information Systems. In N. Guarino (ed.) *Formal Ontology and Information Systems*. //Proceedings of FOIS'98. - Trento, Italy. - 1998. - 6-8 June. – IOS Press, Amsterdam. – pp.3-15.