УДК 004.09

Ф.С. Кортиков

СЕМАНТИЧЕСКОЕ ОПИСАНИЕ ЭЛЕКТРОННОГО ДОКУМЕНТА С ИСПОЛЬЗОВАНИЕМ ОНТОЛОГИИ ПРЕДМЕТНОЙ ОБЛАСТИ

Кортиков Ф.С. **Семантическое описание электронного документа с использованием онтологии предметной области**

Аннотация. В статье рассмотрены и предложены пути решения проблемы анализа электронного документа и его семантического описания с использованием онтологии предметной области.

Ключевые слова: онтология, семантика, дескриптивная логика, предметная область, информационные ресурсы, открытые информационные системы.

Kortikov F.S. **Semantic description for electronic documents using the domain ontology.**

Abstract. Solutions for electronic document analysis are reviewed and suggested in the article. Approach is based on semantic markup and object domain and includes 3 steps.

Keywords: ontology, semantics, descriptive logic, data domain, information resources, public information system.

1. Введение. В настоящее время быстрое и эффективное извлечение и формализация знаний из электронного документа является актуальной задачей. Наличие различного рода электронных документов делает систему электронного документооборота высоко гетерогенной информационной системой. При этом электронные документы не обеспечиваются семантическим описанием, что резко затрудняет целенаправленное извлечение знаний для идентификации в предметной области. Процесс извлечения метаданных из электронного документа может быть как рунным, когда метаданные заносит пользователь, так и автоматизированным, когда метаданные автоматически выделяются из текста документа. В данной работе метаданные выделяются согласно специально разработанного тезауруса.

Автоматизированное извлечение метаданных состоит из нескольких этапов:

1. Составление онтологии предметной области
2. Разбиение электронного документа, написанного на естественном языке человека на отдельные фрагменты. Под фрагментом документа понимается часть текста, имеющая единую структуру, лежащая внутри ключевых слов (токенов) [3]. Выделение метаданных верхнего уровня.
3. Семантическая разметка выделенных фрагментов. Извлечение метаданных нижнего уровня. Построение таксономии в соответствии с

онтологией предметной области. Извлечение знаний из размеченных текстов.

2. Разбиение документа на отдельные фрагменты. Целью данного этапа является автоматизированное внесение в электронные тексты документов, составленных на естественном языке, формальных признаков отдельных понятий онтологии ПрО, характеризующих смысловое содержание документов. Семантическая разметка (СР) выполняется над ИР, которые пользователь отобрал как источники формирования и обновления БЗ ЭС. СР является подготовительным процессом для дальнейшего извлечения знаний и выполняется периодически по мере того, как возникает надобность актуализировать БЗ ЭС на основе новых ИР ПрО.

Можно выделить следующие шаги семантической разметки.

Шаг 1. Разбиение ИР на фрагменты.

Фрагментами ИР могут быть разделы документа, страницы и абзацы. Цель разбиения документов – облегчение ориентировки пользователя в массиве текстовых фрагментов, которые будут получены в результате извлечения знаний. Для фрагментации может быть использован набор символов XML, вставляемых в текст для фиксации информации о его структуре [4].

Шаг 2. Первичная семантическая разметка ИР.

Данный этап выполняется программно, согласно следующему алгоритму. Для каждого понятия c_i ($i=1,..,N$) построенной онтологии из тезауруса выбирается соответствующий ему класс терминов синонимов $W_i=\{w_{ij}\,|\,j=1,..,k_i;\, i=1,..,N\}$. Затем поочередно выполняется поиск этих терминов в размечаемом ИР. В случае если в некотором фрагменте текста обнаружен хотя бы один термин $wm_i \in W_i$, $m=1,..,k_i$, то данному фрагменту присваивается «ярлык» (ТЭГ), соответствующий понятию c_i, и поиск синонимов $w1_i, w2_i, .., wk_i$ продолжается в следующем фрагменте текста. После обработки всех фрагментов (поиска синонимов понятия c_i) процесс повторяется для очередного элемента онтологии (c_{i+1}).

В результате применения подобной процедуры ко всем понятиям онтологии, каждому j-му фрагменту размечаемого текста будет присвоено ij ТЭГов, $i, j \in (0, N)$, где N – количество понятий онтологии.

Шаг 3. Вторичная (дополнительная) разметка ИР.

На этом этапе выполняется дополнительная разметка, учитывающая онтологические отношения между понятиями.

Для создания и управления онтологией предметной области предлагается применить программный комплекс Protege-2000 [1], представляющего собой средство для создания и поддержки онтологий, использующее OKBC – совместимый интерфейс управления знаниями, что позволяет использовать единый интерфейс для работы с различными языками семантической разметки.

3. Определение семантики текста. Для работы с текстом используется программный инструмент, написанный на языке Python. Технологию извлечения знаний из различных текстов ПрО можно представить в виде следующих шагов:

Шаг 1: Формирование запроса для целевого извлечения знаний.

Для формирования запроса используются не ключевые слова, а понятия онтологии ПрО. При этом целесообразно использовать язык описания запросов SPARQL.

Шаг 2: Поиск по сформированному запросу в библиотеке размеченных текстов. Выбор искомых фрагментов выполняется по критерию соответствия запроса пользователя и совокупности ТЭГов, описывающих понятийное содержание фрагментов.

Шаг 3: Упорядочение найденных текстовых фрагментов. Цель данного этапа – подготовить пакет найденных текстовых фрагментов к виду, удобному для последующей фильтрации. Упорядочение выполняется автоматически по одному или нескольким ключевым признакам в зависимости от указания пользователя (инженера по знаниям). Такими признаками могут быть: понятия онтологии с учетом их важности в запросе; информационная содержательность фрагмента Il (3); дата происхождения информационного ресурса и др.

Шаг 4: Фильтрация пакета найденных фрагментов.

Цель фильтрации – удаление повторов, малозначимых фрагментов, ошибочно найденных фрагментов. Этап выполняется инженером по знаниям при сервисной программной поддержке.

Шаг 5: Первичная формализация знаний, представленных в отфильтрованном пакете фрагментов. Цель этапа – представить знания из каждого фрагмента в виде совокупности предложений на ограниченном естественном языке. По каждому фрагменту высвечиваются понятия онтологии в пределах одного абзаца текста. При этом учитываются категории понятий (объект, процесс, событие, свойство, значение и т.п.). Инженер по знаниям формирует предложение в соответствии с правилами ограниченного синтаксиса.

Шаг 6:Описание внутреннего представления знаний формализованных предложений и загрузка в БЗ. Описание данного этапа выходит за рамки данной статьи. Приведем лишь кратчайшее его содержание.

Инженер по знаниям последовательно в диалоговом режиме выводит предложения, сформированные в шаге 5, и преобразует их в форму, принятую в модели знаний ПрО, после чего производится загрузка извлеченных элементов знаний о ПрО в БЗ. По каждому элементу выполняется автоматическая проверка повторяемости знания его противоречивости с уже имеющимися знаниями. Результаты протоколируются, предоставляются инженеру по знаниям для дальнейшей интерпретации.

4. Заключение. Предлагается концепция автоматизированной технологии извлечения знаний (АТИЗ) из информационных ресурсов (ИР), не имеющих предварительного семантического описания. АТИЗ является одним из подходов снижения трудоемкости формирования базы знаний (БЗ) экспертной системы (ЭС), использующей ограниченную предметную область (ПрО). АТИЗ основана на Онтологии и Тезаурусе ПрО, которые позволяют связывать метаданные ПрО с их лексическими представлениями, что является основой для автоматизированной разметки ИР с использованием метаданных онтологии. Последнее обстоятельство, в свою очередь, создает возможность в дальнейшем выполнять целенаправленный поиск знаний в ИР не по ключевым словам, а с использованием понятий ПрО [2].

Литература

1. *Гаврилова Т.А.* Онтологический инжиниринг [Электронный ресурс]. URL: http://www.big. spb.ru/publications/bigspb/km/ontolog_engeneering.shtml.
2. *Кортиков Ф.С.* Методика оценки интероперабельности системы электронного документооборота как открытой информационной системы // Информационно-измерительные и управляющие системы. 2011. 1:11, С. 17–22.
3. *Нариньяни А.С.* ТЕОН-2: от Тезауруса к Онтологии и обратно // Труды Международного семинара «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2002. Т. 1. С. 199–154.
4. *Серебровский А.Н.* О технологии извлечения знаний // Искусственный интеллект, февраль 2010. Институт проблем математических машин и систем НАН Украины. С. 71–77.

Кортиков Федор Сергеевич — аспирант кафедры прикладной математики и информатики Санкт-Петербургского государственного архитектурно-строительного университета. Область научных интересов: создание онтологий, математический анализ, статистический анализ. feado@rambler.ru; СПбГАСУ, 2-я Красноармейская ул., д.4, г. Санкт-Петербург, 190005, РФ; р.т. +7(812) 575-05-34.

Kortikov Fedor Sergeevich — Ph.D. student, Applied mathematics department, St.Petersburg State University for architecture and civil engineering (SPbSUACE). Research interests: ontology, mathematical analysis, statistical analysis. feado@rambler.ru; SPbSUACE, 4, 2-nd Krasnoarmeiskaya St., St. Petersburg, 190005, Russia; office ph. +7(812) 575-05-34.

Рекомендовано лабораторией информационно-вычислительных систем СПИИРАН, заведующий лабораторией Воробьев В.И, д.т.н., проф.
Статья поступила в редакцию 20.03.2013.

РЕФЕРАТ

Кортиков Ф.С. **Семантическое описание электронного документа с использованием онтологии предметной области.**

Предлагается ввести автоматизированное внесение в электронные тексты документов, составленных на естественном языке, формальных признаков отдельных понятий онтологии предметной области, характеризующих смысловое содержание документов. Семантическая разметка выполняется над электронным документом, которые пользователь отобрал как источник формирования и обновления базы знаний экспертной системы. Семантическая разметка является подготовительным процессом для дальнейшего извлечения знаний и выполняется периодически по мере того, как возникает надобность актуализировать базу знаний экспертной системы на основе новых входящих документов.

SUMMARY

Kortikov F.S. **Semantic description for electronic documents using the domain ontology**.

It is proposed to introduce an automated entry in the electronic texts of documents written in natural language, formal features of individual concepts ontology describing the semantic content of the documents. Semantic markup is performed on an electronic document, which the user has selected as the sources and update the knowledge base of expert system. Semantic markup is the preparatory process for the further extraction of knowledge and executed periodically as the need arises to update the knowledge base of the expert system based on the new incoming documents.