ПРИМЕНЕНИЕ МЕТОДОВ TEXT MINING ДЛЯ РЕШЕНИЯ ЗАДАЧ ОНТОЛОГИЧЕСКОГО ИНЖИНИРИНГА

Рябова Н.В., Волошина Н.А., Гринев С.А.
Харьковский национальный университет радиоэлектроники

Онтологический подход к разработке Web-систем, ориентированных на структурированное представление и обработку данных, информации и знаний, в настоящее время признан наиболее эффективным и получил широкое распространение в области современных IT-технологий. Онтологический инжиниринг (Ontological Engineering - OE) активно развивается как отдельное направление научно-практических исследований, корни которого лежат в инженерии знаний, изучающей методы, модели и алгоритмы извлечения, структурирования, представления и обработки знаний с целью построения баз знаний интеллектуальных систем. В рамках OE рассматривается онтологическая парадигма представления знаний в гетерогенных распределенных средах типа Интернет-пространства, с использованием базовых технологий Semantic Web. В связи с этим OE включает в себя решение основных задач, связанных с различными видами деятельности по разработке, управлению жизненным циклом онтологии, методами и методологиями для построения онтологий, а также разработке инструментально-программных средств для их поддержки.

Онтологическая структура формально специфицирует модель предметной области (ПрО), экстенсиональная же часть, определяющая объем модели, обеспечивается базой знаний (БЗ), которая содержит утверждения об экземплярах концептов и отношениях, определенных в онтологии. (Полу)автоматическую поддержку в построении онтологии обычно относят к онтологическому обучению (Ontology Learning - OntoL), которое выделилось в отдельное направление исследований в рамках OE. OntoL может быть охарактеризовано как построение модели ПрО на основе знаний, извлеченных из исходных данных [1].

Входные данные, репрезентативные для ПрО, могут быть представлены в виде схем, таких как XML-DTD, UML-диаграмм, схем БД. Такой вид OntoL относят к так называемому лифтингу (lifting), поскольку он в основном состоит в «подтягивании» или отображении определений из схем в соответствующие онтологические определения. OntoL также может осуществляться на основе полуструктурированных источников, таких как XML, HTML-документы или табличные структуры. В том случае, когда OntoL осуществляется на основе неструктурированных текстовых источников, говорят об онтологическом обучении «из текстов».

Несмотря на то, что в последние годы было предложено довольно много методов для решения отдельных задач OntoL на основе знаний, извлеченных из текстов, до сих пор нет общепризнанной методологии такого типа онтологического обучения, что, в свою очередь осложняет возможность сравнения предлагаемых подходов. В данной работе выделяется и анализируется последовательность задач онтологического обучения на основе

текстовых документов (ТД), которые вместе составляют комплексную задачу разработки онтологии с основным акцентом специфицирования семантики ее сущностей [2]. При этом каждая последующая задача является этапом построения онтологии, опирающимся на результаты предыдущего этапа. Перечислим эти задачи:

1) извлечение релевантной терминологии из ТД;

2)идентификация синонимичных терминов (лингвистических вариантов, в том числе, возможно, межъязыковых);

3) формирование множества концептов;

4) иерархическая организация концептов;

5) обучение отношениям и атрибутам (свойствам концептов) вместе с отнесением их к соответствующим области и диапазону действия;

6) иерархическая организация отношений;

7) означивание схем аксиом примерами для возможных ограничений интерпретаций концептов и отношений;

8) определение общих аксиом.

Систематическая организация и формализация взаимосвязанных задач OntoL обеспечивается единой онтологической моделью, которая включает в себя непересекающиеся множества концептов, отношений, атрибутов и типов данных. Теоретическим базисом для построения такой единой онтологической модели являются методы интеллектуального анализа текстов, Text Mining [3,4,5], последовательно применяемые для кластеризации, классификации и извлечения из ТД релевантной терминологии, а также метод анализа формальных концептов (Formal Concept Analysis - FCA), используемый для автоматизации получения таксономий из текстовых коллекций. Модель контекста формируется в результате применения парсера (лингвистического анализатора) и построения векторного представления синтаксических зависимостей определенных терминов. Затем с помощью FCA строится решетка, которая конвертируется в иерархию концептов.

1. Cimiano Ph. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications [Текст]. – Springer Science+Business Media, LLC. – 2006. – 347 p. 2. Nirenburg S. Ontological Semantics [Текст] / S. Nirenburg, V. Raskin. – The MIT Press, 2004. – 420 p. 3. Feldman R. The Text Mining Handbook: Advanced Approaches in Analizing Unstructered Data [Текст] / R. Feldman, J. Sanger. – Cambridge University Press, 2007. – 410 p. 4. Бодянский Е.В. Классификация текстовых документов с помощью нечеткой вероятностной нейронной сети / Е.В. Бодянский, Н.В. Рябова, О.В. Золотухин // Восточно-Европейский журнал передовых технологий – 2011. – №6/2 (54). – С.16-18. 5. Рябова Н.В. Методы согласования онтологий в задачах семантической интеграции информационных Web-систем / Н.В. Рябова, Н.А. Волошина, И.В. Тесленко // Материалы международной научно-технической конференции «Информационные системы и технологии»: тезисы докл. – Харьков, 2013. – С. 61-62.