1. Отчет за 2016 год

к.т.н. Черкашин Е.А.

Заявлено в планы на 2016 год. *Разработка методов извлечения онтологии, как грубой модели информационной системы, из неформализованных описаний системы.*

Предложена методика автоматизированного построения концептуальной модели предметной области. В основу методики входит полисистемное онтологическое представление предметной области, где составляющие онтологии представляют собой систему слоев (категорий), отображаемых друг в друга при помощи функторов. Полисистемное представление позволяет создавать концептуальный базис, поддерживающий тождественность элементов онтологий, предназначенных для различных аспектов моделирования предметной области и исследуемого объекта.

На *первом шаге* методики в текстах входных документов автоматически выделяется набор ключевых слов, производится построение иерархической классификаций документов согласно выделенному набору ключевых слов и словосочетаний. На *втором шаге* осуществляется поименование узлов иерархической классификации наиболее часто встречающимися ключевыми словами в соответствующем под поддереве классификации. Первый и второй шаги выполняются согласно существующим алгоритмам, адаптированным для обработки текстов предметной области. Результатом второго этапа является тезаурус, онтология, где задан словарь концептов и отношения вида «is–a» («является») между этими концептами.

На *третьем шаге* производится сопоставление сгенерированного тезауруса терминологическому базису используемой полисистемы онтологий. В результате такого сопоставления тезаурус обогащается как новыми «горизонтальными» отношениями, так и производится его оценка на полноту относительно отношений, заданных в полисистеме онтологий. После анализа полученной онтологии производится ее интегрирование в полисистему в виде нового слоя. При необходимости более детальной декомпозиции предметной области предусматривается повторение методики над частями текста более мелкого масштаба (переход от документам к их разделам, от разделов к отдельным таблицам и строкам).

Основная цель исследований – разработка методики рекурсивной декомпозиции автоматизируемой предметной области и ее привязка к существующим слоям полисистемы онтологий. Такая привязка позволяет переносить существующие программные реализации теорий слоев на новые слои по их образу и подобию.

В 2016 году в рамках исследования разработана программная инфраструктура для хранения слоев полисистемы с использованием существующих стандартных онтологий. Слой представляет собой часть онтологии (ядро), которая подвергается отображению в другой слой через функтор или является результатом такого отображения. Слои хранятся в онтологической базе в виде графов. Хранилище онтологий реализовано при помощи системы ClioPatria, использующей язык логического программирования Prolog в качестве языка реализации. Это позволяет создавать логические теории слоев непосредственно на сервере. Кроме того, декларативная сущность Prolog-а значительно упрощает задачу переноса реализаций теорий между слоями, т.е. порождение программ-аналогов.

В хранилище онтологий загружен ряд онтологий, предназначенных для описания организационной структуры предприятий, сообществ и документов. Разработаны программные модули для представления аннотаций документов в стандарте Open Annotation и NEPOMUK (Network Environment for Personalized, Ontology-based Management of Unified Knowledge), а также данные физических и юридических лиц – FOAF, в онтологическом

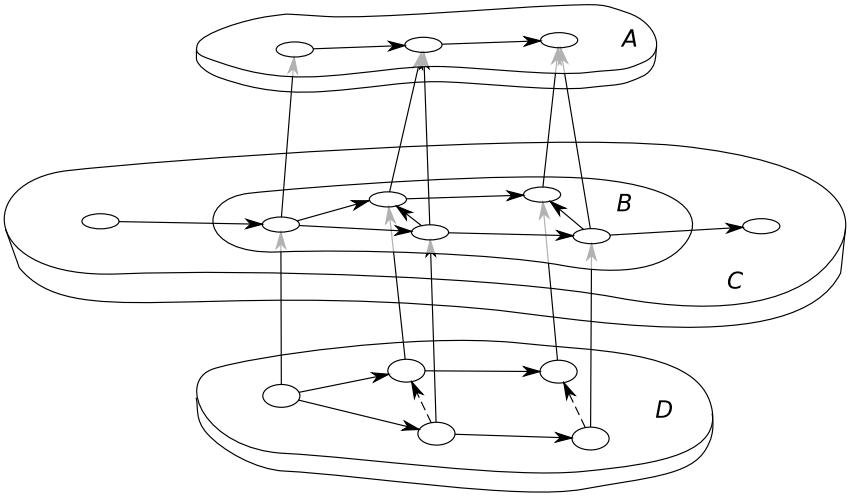

Рис. 1: Представление предметной области в виде полисистемы онтологий.

Слои A, C – стандартные онтологии; A, B – слои, категории; B – ядро (категория) стандартной онтологии C; $B \to A$ – интерпретация объектов и стрелок B в онтологии A (морфизм); D – онтология предметной области; $D \to B$ – интерпретация предметной области D (онтологии, категории, слоя) в слой B. Пунктирные стрелки восстановлены в D по их аналогам из B. Морфизмы стрелок между слоями на рисунке не отображены.

хранилище, а также модули для порождения аннотаций из результатов анализа текстов документов.

Разработанная распределенная система применена в решении задачи автоматизации анализа структуры текстов учебных программ технического вуза. Для этого произведена разметка абзацев учебных программ обозначениями концептов, задающих семантическое значение материалу, представленному в абзаце или таблице. Затем выполнено машинное обучение, направленное на распознавание концепта в зависимости от свойств, выделенных в абзаце (наличие ключевых слов, определенных словосочетаний и стилей представления текста). Результатом применения синтезированной системы знаний является семантическая разметка документа, которая отображается на слои полисистемы, что в дальнейшем позволяет применять к элементам документов те или иные алгоритмы обработки данных, а также стили и форматы отображения.