Секция «Програмирование»

Правдоподобный вывод в задаче формирования тезауруса

Мартынов Роман Андреевич
E-mail: roman24082011@mail.ru

Информационные технологии постоянно совершенствуются. В частности, это связано с внедрением систем, способных работать со знаниями в конкретной предметной области.

Знания представляются в определенной форме, удобной для дальнейшей обработки. Тезаурусы в предметных областях есть именно такая форма отображения знаний [1]. Под тезаурусом предметной области понимается словарь, в котором лексические единицы (слова или словосочетания) связаны между собой по смыслу.

Сейчас существует много тезаурусов предметных областей, но разработка тезауруса новой предметной области является непростой задачей. Стандартным способом создания тезауруса является ручной способ. Однако такой способ требует значительных затрат как времени, так и денег, что, иногда, неприемлемо для малых проектов, в процессе реализации которых требуются специализированные тезаурусы для решения своих задач. К тому же, создание тезауруса вручную сопряжено с фактором субъективности, когда качество сформированного тезауруса зависит от опыта специалиста. Поэтому разрабатываются методы автоматизированного формирования тезауруса.

К сожалению, существующие методы создания словарей обладают значительными недостатками, из-за чего следует предпочесть интегрированный способ формирования тезауруса с использованием комбинации теоретических и практических операций.

Библиографический предметный анализ показал, что в литературе нет упоминаний о методах, использующих правдоподобные рассуждения. Правдоподобные рассуждения, по заявлению отечественных исследователей [2,3], во многом схожи с рассуждениями людей, и являются неотъемлемой частью познания, поэтому их можно применить для формирования словаря.

В рамках работы рассматривался ДСМ-метод, основанный на правдоподобных рассуждениях. Выбор ДСМ-метода в качестве основного для задачи формирования тезауруса обусловлен простотой реализации и наличием обобщенного математического описания на языке наивной теории множеств (Аншаков О.М. [2] и Липкин А.А. [3]).

Базисными сущностями, с которыми работает ДСМ-метод, являются «объект» и «свойства» объекта. Предполагается, что объект состоит из «фрагментов»; обнаружение определенного фрагмента у объекта может указать на присутствие тех или иных свойств. Существенным преимуществом ДСМ-метода служит то, что можно находить связь между структурой объектами и свойствами, формировать предположения о наличии свойств у объектов, для которых ранее эта информация отсутствовала, но сведения о структуре объекта известны.

Для переквалификации ДСМ-метода в метод, применимый к задаче формирования тезауруса, была использована гипотеза распределения в лингвистике. Согласно этой гипотезе лексические единицы, близкие по смыслу, имеют общих соседей при достаточном объеме исходных данных.

На основе адаптированного ДСМ-метода реализована система на языке Python, автоматически формирующая тезаурус. Коррекция её результатов проводилась с использованием различных предложенных в работе способов, включая методы машинного обучения, а оценка эффективности системы осуществлялась на основе уже имеющихся таблиц семантически близких слов тезауруса РуТез-Lite [4].

Источники и литература

1) Гладун, А.Я. Онтологии в корпоративных системах // Корпоративные системы – М.: Комиздат, 2006. – С. 13-26.

2) Аншаков О.М. ДСМ-метод: теоретико-множественное объяснение. // Корпоративные системы – М.: Комиздат, 2012. – С. 1-19

3) Липкин А.А. ДСМ-метод порождения гипотез для объектов, описываемых атрибутами с весами. // Автореферат диссертации на соискание ученой степени – 2008. – 21 с.

4) Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Издательство Московского университета, 2011. – 512 с.