Исследование методов извлечения данных при помощи СПО и материалов энциклопедии “Википедия”

Пожидаев М. С., Жарков Д. С., Кирюшкина В. Е.

7-я конференция “СПО в высшей школе” (Переславль-Залесский, 28 января 2012 г.)

В крупнейших репозиториях свободного программного обеспечения (СПО) наблюдается недостаток развитых инструментов, традиционно называемых “интеллектуальными системами” (ИС). К ним относится программное обеспечение (ПО), способное к анализу и обработке информации в форме, сравнимой с восприятием информации человеком. Научные сотрудники, желающие вести исследования в области ИС, сталкиваются с рядом трудностей, связанных с большим количеством сопутствующих прикладных задач и недостатком материалов, пригодных для использования в качестве входных данных для исследуемых алгоритмов.

С конца 90-х годов прошлого века получили развитие такие научное направления, как *Data Mining* и *Text Mining*. Они занимаются исследованием различных подходов к выделению значимой для человека информации из больших массивов слабоструктурированных данных. Обычно алгоритмы *Data Mining* носят эвристический характер и основаны на вычислении различных статистических параметров, но дополнительно предпринимаются попытки разработки методов, основанных на строгих и формализованных правилах логического вывода, с исследованием моделей хранения знаний, как, например, семантические сети.

Ряд зарубежных университетов ведёт разработку инструментов, способных значительно упростить исследования в области *DATA Mining*, и распространяемых на условиях свободных лицензий. Отметим среди них проект *CoreNLP* Стэнфордского университета и проект *WEKA* университета Уайкато (Новая Зеландия).

Пакет *CoreNLP* предоставляет реализацию инструментов для решения некоторых задач по обработке естественного языка (*Natural Language Processing*), среди которых получение нормальной формы слов, обработка членов предложения и пр. Решение подобных задач часто необходимо при исследовании алгоритмов *Text Mining*.

Проект *WEKA* предоставляет реализацию распространённых типовых подходов к *Data Mining*, включая извлечение правил ассоциаций и выборку атрибутов на основе заранее подготовленной базы данных (БД), которая должна представлять из себя множество объектов с заданными значениями атрибутов. Применение пакета *WEKA* требует наличия достаточно обширной БД для получения значимых полезных результатов. Оба указанных проекта реализованы на языке *Java*.

В Томском государственном университете (ТГУ) ведутся исследования возможности подготовки БД для запуска алгоритмов *WEKA* на основе обработки статей свободной энциклопедии “Википедия”. Изучают-

ся подходы к выделению множества объектов и их атрибутов на основе текста статей. При таком рассмотрении алгоритмы *Text Mining* выступают как предварительный этап для алгоритмов *Data Mining*.

В ходе исследований планируется постановка множества вычислительных экспериментов. В них главные члены предложений выступают как обозначения объектов БД в двух случаях: если они являются ссылками на некоторые статьи или если они не являются ссылками, но их нормальная форма совпадает с названием статьи, где они упоминаются. Набор второстепенных членов предложений рассматривается как потенциальные значения атрибутов объектов. При этом ряд сопутствующих задач по обработке естественного языка решается при помощи инструментов *CoreNLP*. Множество статей Википедии доступны для загрузки в виде единого файла в формате *XML*. Большой объём исходных материалов значительно усложняет задачу, и в случае серьёзных трудностей рассматривается возможность привлечения мощностей вычислительного кластера ТГУ "СКИФ *Cyberia*".

Получение значимых результатов позволит подготовить новые свободные инструменты для решения сложных задач. К ним относится крайне актуальная задача фильтрации нежелательной почты. Сообщения рекламного характера содержат множество имён собственных, для анализа которых требуется подготовленная БД. При помощи материалов Википедии возможно назначение различным именам собственным оценки их близости к наиболее распространённым темам, встречающимся среди нежелательных сообщений. Другим примером потенциального применения БД на основе статей Википедии может быть система, воспринимающая для обработки утверждения на ограниченном естественном языке и оценивающая степень их достоверности. Задача фильтрации нежелательной почты во многом сводится к хорошо изученному методу классификации, и поэтому является более простой для исследований, чем система оценки достоверности утверждений на ограниченном естественном языке. Помимо этого, в последнем случае требуется дополнительное исследование набора хранимых атрибутов объектов и их назначения, поскольку они должны охватывать предельно большое количество характеристик объектов.