УДК 004.89

Арбузова О.В.,Егошина А.А., Линкин В.О.
Донецкий Национальный Технический Университет, г. Донецк
кафедра систем искусственного интеллекта

СИСТЕМА ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ОБ ОДНОТИПНЫХ ОБЪЕКТАХ ИЗ МАССИВОВ ТЕКСТОВЫХ ДОКУМЕНТОВ

Аннотация:

Егошина А.А., Линкин В.О., Арбузова О.В. Система извлечения данных об однотипных объектах из массивов текстовых документов. *Рассмотрен подход к решению задачи Text Mining по извлечению информации об объектах из текстовых документов и переносу в базу данных для последующего извлечения знаний на основе стандартных методов Data Mining. Сформулированы требования к системе извлечения и структурирования данных. Предложена структурная организация системы и выбраны алгоритмы извлечения.*

Ключевые слова: *Text Mining, текстовый документ, объект, база данных, организация системы, алгоритмы извлечения*

Общая постановка проблемы. Развитие индустрии систем электронного документооборота сопровождается накоплением и хранением на серверах баз данных огромных массивов документов и приводит к ряду проблем поиска конкретных документов и извлечения необходимой информации и знаний. Для решения этих проблем актуальны задачи создания систем автоматизированной классификации документов, которые будут учитывать особенности документооборота и технологий Text Mining (TM) - автоматической добычи знаний из больших объемов текстового материала. TM часто называют текстовым Data Mining. TM добавляет к технологии DM дополнительный этап - перевод неструктурированных текстовых массивов в структурированные. После чего данные могут обрабатываться с помощью стандартных методов DM.

Представление текстовых документов в структурированном виде, с целью извлечения знаний, является в общем случае сложной задачей из-за проблем семантической неоднозначности и неопределенности естественных языков. Решение задачи упрощается при анализе естественно языковых специализированных текстовых документов связанных с конкретной предметной областью. В этом случае в текстах используется ограниченная иерархия понятий и терминов и с использованием знаний экспертов в предметной области такие документы могут быть структурированы, поэтому актуальной является задача автоматизации процесса структуризации текстовых документов с целью исключения больших трудозатрат экспертов.

Анализ существующих средств Text Mining. На сегодняшний день существует множество систем добычи текстовых данных и глубинного анализа текстов. Разработкой таких систем занимаются как небольшие частные компании, группы ученых и программистов, так и гиганты компьютерной индустрии.

Так, IBM предлагает свою разработку Intelligent Miner for Text [1], которая, по сути, является набором отдельных утилит:

Language Identification Tool – утилита определения языка – для автоматического определения языка, на котором составлен документ; Categorisation Tool – утилита классификации – автоматического отнесения текста к некоторой категории (входной информацией на обучающей фазе работы этого инструмента может служить результат работы следующей утилиты – Clusterisation Tool);

Clusterisation Tool – утилита кластеризации – разбиения большого множества документов на группы по близости стиля, формы, различных частотных характеристик выявляемых ключевых слов;

Feature Extraction Tool – утилита определения нового – выявление в документе новых ключевых слов (собственные имена, названия, сокращения) на основе анализа заданного заранее словаря;

Annotation Tool – утилита «выявления смысла» текстов и составления рефератов – аннотаций к исходным текстам.

Компания SAS Institute предложила систему, способную сравнивать определенные грамматические и словесные ряды в письменной речи человека, с которым вы общаетесь посредством электронной почты, с тем, что было написано им ранее, и выявлять подозрительные несовпадения. Система получила название Text Miner [2].

Среди разработок на постсоветском пространстве стоит выделить системы GALAKTIKA-ZOOM и «Медиалогия». Программный комплекс GALAKTIKA-ZOOM предназначен для аналитической обработки динамично пополняющихся больших массивов (до десятков миллионов) текстовых документов, находящихся в подключаемых неструктурированных и структурированных электронных базах данных [2].

Система «Медиалогия» не предусматривает передачи программы заказчикам, производя обслуживание клиентов в онлайновом режиме. «Медиалогия» — это web-приложение, представляющее собой мощное решение со сложной архитектурой и обеспечивающее непрерывную обработку поступающей информации, структурированное хранение данных, расчет аналитических параметров, проведение анализа по запросам пользователя и хранение настроек и отчетов.

Естественно, что в стороне не могли остаться разработчики СУБД. Так, средства Text Mining можно увидеть в продуктах Oracle начиная с версии 7.3.3. В версии Oracle9i эти средства развились и получили название Oracle

Text – программный комплекс, интегрированный в СУБД, позволяющий эффективно работать с запросами, относящимися к неструктурированным текстам. При этом обработка текста сочетается с возможностями, которые предоставлены пользователю для работы с реляционными базами данных. В частности, при написании приложений для обработки текста стало возможным использовать SQL [3].

Это только некоторые разработки в области интеллектуального анализа информации, иллюстрирующие возможности и широту применения средств TextMining.

Цель статьи – анализ подходов к извлечению информации об объектах из текстовых документов систем документооборота предприятий и разработка структуры системы извлечения информации из текстов и ее структурирования с целью дальнейшей автоматической обработки.

Постановка задачи исследования. Существующие программные средства для решения задач ТМ позволяют решать широкий спектр задач по извлечению знаний из текстовых документов, однако все они являются коммерческими и для большинства организаций малодоступны. Для решения актуальных задач, решаемых на предприятиях с использованием системы документооборота, необходимы средства извлечения информации об однотипных объектах из массивов текстовых документов. Такая система должна удовлетворять следующим требованиям.

1. Извлеченная информация должна храниться в базе данных, которая пополняется автоматически.
2. На этапе обучения системы в базу заносятся тестовые данные, на которых происходит обучение.
3. Система должна производить поиск по значениям атрибутов однотипных объектов. Перечень объектов и атрибутов может расширяться.

Организация системы извлечения объектов из специализированных текстов. В системах документооборота предприятий имеется большое количество текстовых документов с ограниченной лексикой (характеристики продукции, комплектующих и сырья, отчеты о поломках, результаты опросов и т.д.).

Для таких текстов упрощается создание алгоритмов преобразования их к стандартизованному структурированному виду. Стандартизованную информацию проще обрабатывать и можно использовать в различных производственных целях.

Например, описание манипулятора мышь: компьютерная мышка Genius, модель DX-7010, цвет Black, датчик оптический, радио интерфейс, 2 стандартные клавиши, колесо прокрутки с функцией нажатия, Windows-совместимая. Данный текст требуется преобразовать к виду, приведенному в таблице 1.

Такие тексты объединяет упрощенный язык, правила построения проще предложений естественного языка, описания практически всегда схожи друг с другом по структуре, чаще всего используются один и тот же относительно небольшой, если сравнивать с естественной речью, набор слов, очень часто встречаются аббревиатуры и сокращения. Поэтому существенно упрощается система разбора таких текстов.

В предлагаемой системе будем использовать шаблонный метод с обучением. Состав системы показан на рисунке 1.

Таблица 1 – Структурированный вид текста

Поле	Значение
Тип устройства	манипулятор мышь
Торговая марка	Genius
Модель	DX-7010
Тип датчика	оптический
Интерфейс	радио
Количество кнопок	2 стандартные клавиши
Цвет	Black
Совместимость	Windows

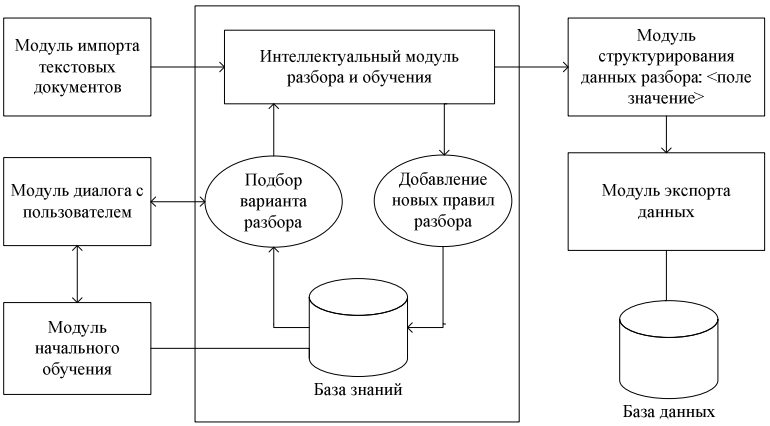

Рисунок 1 -Структура системы извлечения объектов из специализированных текстов

Перед началом работы пользователем и модулем начального обучения формируются шаблоны разбора для предметных областей системы документооборота и правила разбора текстов с помощью шаблонов. Шаблоны, например, для периферийных устройств ПК представляются фреймом имеющим вид, приведенный в табл.2.

Значения слотов накапливаются при обучении. Составленный вручную шаблон, обладает высокой точностью и низкой полнотой для извлечения именованных сущностей из текста. Применяя этот шаблон, на начальном этапе в режиме диалога с системой, получаем обучающее множество для машинного обучения.

Если рассмотреть кластер, объединяющий несколько тематически близких документов, то в нем часто оказывается достаточное количество близких по смыслу предложений, включающих как предложения, в которых некоторый фрейм распознан вполне успешно, так и предложения, в которых этот же фрейм не распознан совсем или распознан частично. Эту вторую группу предложений можно использовать для наращивания шаблонов для распознавания данного фрейма.

Таблица 2 – Структура фрейма шаблона

Периферийные устройства ПК
Тип устройства
Торговая марка
Изготовитель
Совместимость
Интерфейс
Цвет

Для обеспечения дополнительного обучающего множества, необходимо выделить предложение, в котором найден некоторый фрейм объекта, а затем найти множество предложений кластера, похожих на исходное, но в которых нужный фрейм не обнаружен. Предложения, содержащие шаблоны извлекаемой информации, служат центрами для кластеров схожих предложений, в которых такие шаблоны не обнаружены.

При своей работе система извлечения данных определяет наличие в текстовом фрагменте некоторого объекта (фрейма) и выделяет связанную с этим объектом информацию (слоты фрейма и их значения).

В алгоритмах изначально производится обработка текстов кластера при помощи имеющегося шаблона и извлекается информация об объекте из текста. Далее производится поиск извлеченной информации в каждом предложении кластера, в котором не удалось установить наличие извлекаемого объекта и подсчитывается количество найденных слотов.

При работе с предметной областью отличной от ранее обработанной, но достаточно близкой к ней, система может использовать найденные ранее правила разбора, однако при этом могут извлекаться не все объекты или слоты.

Для повышения точности отбора предложений предполагается использовать алгоритмы, основанные на вычислении меры близости предложений по количеству выделенных слотов, по косинусу угла между предложениями и по значению меры TF*IDF слова [4].

Установление порога по количеству найденных слотов не всегда позволяет выявить все объекты.

Косинус угла между вектором слов данного предложения x и вектором слов предложения y, в котором анализатор обнаружил наличие некоторого объекта, вычисляется по формуле (1).

$$C(x,y)=\frac{\sum_{i=1}^{n}x_i y_i}{\sqrt{\sum_{i=1}^{n}x_i^2}\sqrt{\sum_{i=1}^{n}y_i^2}} \qquad (1)$$

В алгоритме TF*IDF также для оценки близости двух предложений используется мера косинуса, но вместо слов в векторе предложения используется значение меры TF*IDF слова, вычисляемой по формуле (2)

$$TFIDF=\frac{tf}{2(0.25+0.75dl/dlavg)+tf}\log\left(\frac{N-df+0.5}{df+0.5}\right), \qquad (2)$$

где N — количество кластеров;

tf — частота слова в кластере;

df — количество кластеров, содержащих данное слово;

dl — количество слов в кластере

$dlavg$ — средняя число слов в кластере.

Выводы

Предложенная система может использоваться в системах документооборота различных организаций. Использование данной системы позволит исключить ручную обработку и значительно увеличить скорость обработки текстов, повысить качество благодаря тому, что одни и те же термины всегда разбираются идентично, после "обучения системы" использовать менее квалифицированные кадры для обработки значительной части текстов, т. е. повысить эффективной работы.

Список литературы

1. Intelligent Miner for Text (IBM). Интернет-ресурс. - Режим доступа: www/ URL: http://www.ibm.com/software/data/iminer/fortext/

2. Products & Solutions / *Text Mining*. Интернет-ресурс. - Режим доступа: www/ URL: http://www.sas.com/

3. «Галактика-Zoom». Интернет-ресурс. - Режим доступа: www/ URL: http://zoom3.galaktika.ru

4. Колесов А.Н. Извлекая знания из хаоса информации // PC Week. – 2003. – № 43. – С. 18-20.

5. Hatzivassiloglou V., Klavans J., Holcombe L., Barzilay R., Min-Yen Kan, McKeown R. SIMFINDER: A Flexible Clustering Tool for Summarization. Proceedings of the NAACL Workshop on Automatic Summarization, 2001.