УДК 608.01 519.765

Ю. Е. Котельникова

ОБРАБОТКА ТЕКСТОВЫХ ДОКУМЕНТОВ И ЭВОЛЮЦИЯ АВТОМАТИЗИРОВАННЫХ СИСТЕМ ПРОЕКТИРОВАНИЯ

Исследованы системы обработки текстовой информации, рассмотрены проблемы анализа текстов в производственных задачах.

Ключевые слова: *неструктурированные данные, Text Mining, текстовые данные, автоматизированные системы.*

Введение. Из всей существующей информации не менее 90 % составляют неструктурированные данные, т.е. совокупность документов, представляющих собой логически объединенный текст без каких-либо ограничений на его структуру. Такая информация хранится в текстовых полях, дальнейшая обработка которых невозможна без потери семантики текста и отношений между его элементами. Для анализа неструктурированных данных на стыке нескольких областей (DataMining, обработка естественных языков, поиск информации, извлечение информации и управление знаниями) разрабатывается особая группа методов — Text Mining.

Программное обеспечение, реализующее методы Text Mining. На настоящий момент существует программное обеспечение, реализующее методы Text Mining, это — масштабируемые системы, имеющие развитые графические интерфейсы, богатые возможности

визуализации и манипулирования данными, которые предоставляют доступ к различным источникам данных, функционируют в архитектуре клиент—сервер. Рассмотрим их подробнее.

SemioMap — это продукт компании Entrieva, созданный в 1996 г. ученым-семиотиком Клодом Фогелем. Центральным блоком SemioMap является лексический экстрактор — программа, которая выявляет в текстовой совокупности фразы, объединенные общей семантикой [1].

Autonomy Knowledge Server. Основное преимущество системы — мощные интеллектуальные алгоритмы, основанные на статистической обработке. Эти алгоритмы базируются на информационной теории Клода Шаннона, байесовых вероятностях и нейронных сетях.

Galaktika-ZOOM — продукт российской корпорации „Галактика“. Основное назначение системы — интеллектуальный поиск по ключевым словам с учетом морфологии русского и английского языков, а также формирование информационных массивов по конкретным аспектам [1].

InfoStream. Ядром механизма обработки содержания InfoStream является полнотекстовая информационно-поисковая система InfoReS. Технология позволяет создавать полнотекстовые базы данных и осуществлять поиск информации, формировать тематические информационные каналы, автоматически „рубрицировать“ информацию, формировать таблицы взаимосвязей понятий, гистограммы распределения весовых значений отдельных понятий.

Средства Oracle — Oracle Text, InterMedia Text. В Oracle9i средства текстового анализа развились и получили новое название — Oracle Text — программный комплекс, интегрированный в СУБД, обеспечивающий решение следующих задач анализа текстовой информации: поиск документов по их содержанию, классификацию документов, кластеризацию документов, извлечение ключевых понятий, автоматическое аннотирование, поиск в документах ассоциативных связей.

Intelligent Miner for Text. Этот продукт фирмы IBM представляет собой набор отдельных утилит, запускаемых из командной строки или из скриптов независимо друг от друга. Система включает ряд базовых компонентов, которые имеют самостоятельное значение вне пределов технологии Text Mining.

Text Miner. Американская компания SAS Institute выпустила систему SAS Text Miner для сравнения определенных грамматических рядов в письменной речи. Text Miner обеспечивает логическую обработку текста в среде пакета SAS Enterprise Miner. Это позволяет пользователям обогащать процесс анализа данных, интегрируя неструктурированную текстовую информацию с существующими структурированными данными.

TextAnalyst компании Мегапьютер Интеллидженс решает следующие задачи методов Text Mining: создание семантической сети большого текста, автоматическое аннотирование текста, поиск по тексту, классификацию документов, кластеризацию текстов. Система TextAnalyst рассматривает технологию TextMining в качестве отдельного математического аппарата, который разработчики программного обеспечения могут встраивать в свои продукты, не опираясь на платформы информационно-поисковых систем или СУБД.

WebAnalyst — также продукт компании Мегапьютер Интеллидженс — представляет собой интеллектуальное масштабируемое клиент-серверное решение для компаний, желающих усовершенствовать результат анализа данных в web-среде. Сервер WebAnalyst функционирует как экспертная система сбора информации и управления контентом web-сайта.

Как видим, эти системы пытаются обрабатывать текст, учитывая определенные нормы языка (так как разработки в основном зарубежные — английского). Следовательно, их использование для русских текстов сильно ограничено. Кроме того, нет ни одной системы, последовательно обеспечивающей весь процесс обработки неструктурированного текста (рис. 1). Каждый из этапов процесса подразумевает использование набора шаблонов, с которыми сравнивается найденная информация. Для разных областей знаний необходимы специфиче-

ские базы шаблонов. Данная база должна постоянно обновляться. Таким образом, определяется структура, обеспечивающая использование имеющихся шаблонов для новых запросов (рис. 2).

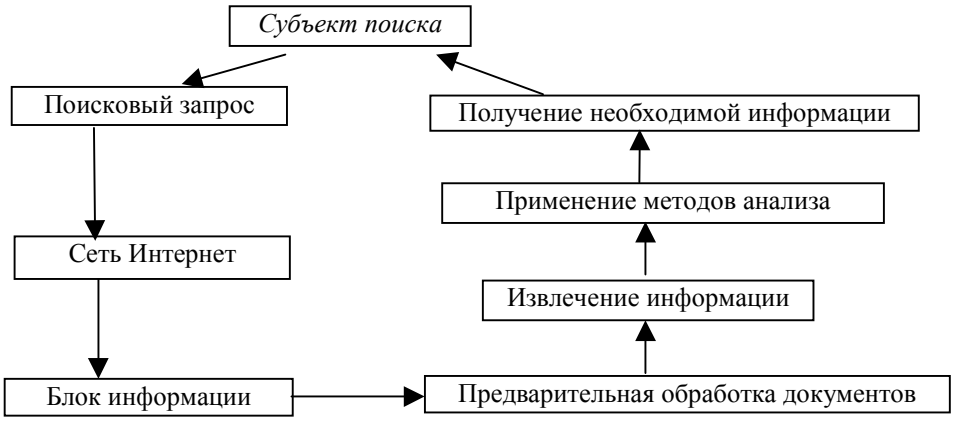

Рис. 1

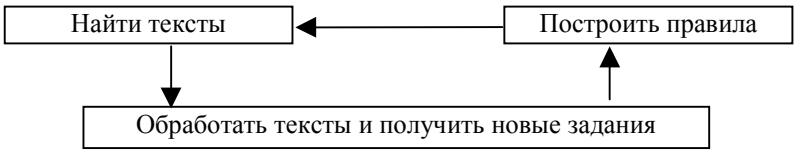

Рис. 2

Проблема анализа текстов в производственных задачах. Реализация указанной схемы, по нашим представлениям, происходит в три этапа.

1. Создание системы поиска. Цель этапа — помочь человеку (специалисту, автору) в поиске текстов, в которых могут содержаться данные, необходимые для решения поставленной задачи.

2. Обработка текста и извлечение знаний. Цель этапа — формирование новых понятий и отношений между понятиями с участием человека и без его участия.

3. Интерпретация новых знаний. Цель этапа — формирование новых или корректировка старых текстов, создание новых документов, проектирование новых правил или корректировка старых [2, 3].

На первом этапе создается автоматизированная система поиска текстов по шаблонам, которые формируются специалистами. Шаблоны могут храниться в базе или создаваться оперативно в процессе обработки текста. На этом этапе осуществляется:

1) подбор текстов, которые необходимо прочитать и проанализировать специалисту перед началом или в процессе

— выполнения конкретного проекта;

— постановки задачи и написания пояснительной записки;

— решения изобретательской задачи;

— решения конкретной инженерной задачи;

— написания книги, учебника, учебного пособия и т.п.;

2) формирование подсказки и подбор материала для

— специалиста, решающего конкретную инженерную задачу, чтобы оперативно пополнить его знания новыми материалами и знаниями в данной проблемной области;

— специалиста, решающего изобретательскую задачу, чтобы оперативно позволить „подсмотреть“ подходы к решению похожих задач в других проблемных областях;

— автора текста (книги, учебника, учебного пособия и т.п.), который был сформирован ранее, чтобы дополнить и откорректировать его.

На этом этапе следует особое внимание обратить на достижения в области *библиографии*, *перевода текстов* и *математической лингвистики*.

Целесообразно организовать поиск текстов с соблюдением их структуры. В общем случае текст имеет следующую структуру: название, шифр по классификатору, аннотация ко всему тексту, оглавление, введение, аннотация к разделам текста, текст раздела, выводы по разделу текста, заключение ко всему тексту, тезаурус (глоссарий).

Каждая структурная часть позволяет найти ответ на вопрос поиска: „Может быть полезен данный текст?“. Как нам кажется, для этого следует привлечь небольшой объем уже каким-то образом структурированных данных, а именно: название, шифр по классификатору, аннотацию ко всему тексту, оглавление, тезаурус (глоссарий) [4].

На втором этапе создается автоматизированная система извлечения знаний из текстов и формирования шаблонов для поиска новых текстов. Сформированные шаблоны заносятся в базу и могут уточняться оперативно специалистом.

Процесс проектирования автоматизированной системы извлечения знаний из текстов и формирования шаблонов для поиска новых текстов состоит из следующих этапов:

— автоматизация функций фильтрации, агрегации данных (обобщение данных должно выполняться с участием специалиста, при автоматическом выполнении только отдельных операций обобщения);

— автоматизация процесса обобщения данных и формирования новых понятий и отношений между понятиями. Процесс выполняется автоматически. Специалист либо контролирует результаты анализа, либо проводит анализ и синтез новых понятий и отношений на паритетных началах с автоматизированной системой;

— автоматизация всех операций анализа и синтеза извлечения знаний. Результаты могут контролироваться специалистом, но могут выполняться автоматически.

На втором этапе проводятся следующие виды автоматизированных и автоматических работ:

— аннотирование новых материалов,

— формирование понятий, отношений и шаблонов,

— структурирование процесса изучения нового материала (речь идет о последовательности изучения материала),

— анализ и обобщение нового материала.

Построение новых документов, правил и текстов может выполняться в автоматизированном режиме, при котором основная роль отводится специалисту, система реализует только вспомогательные функции.

На этом этапе также необходимо особое внимание обратить на достижения *в переводе текстов* и *математической лингвистики*.

При решении задач на втором этапе требуется обработать большой объем данных, в лучшем случае слабо структурированных, а именно: введение, аннотацию к разделам текста, текст раздела, выводы по разделу текста, заключение ко всему тексту.

На третьем этапе создается автоматизированная система интерпретации результатов анализа данных и синтеза новых знаний. С помощью данной автоматизированной системы возможно проектировать и корректировать алгоритмы, формировать документы, формировать новые и корректировать старые тексты.

Заключение. В настоящий момент отсутствует программное обеспечение, осуществляющее полный и последовательный анализ неструктурированного текста, а также в полном объеме работающее с русскоязычными текстами. Решено создать автоматизированную систему нового типа, специализирующуюся на технологической базе знаний, в которой будут реализованы все эти требования. Вынесено предложение о структуре системы, целях и содержании каждого этапа.

Работа проводилась в рамках инновационной образовательной программы „Инновационная система подготовки специалистов нового поколения в области информационных и оптических технологий“ при создании образовательного модуля „Поиск научных и технических решений“.

СПИСОК ЛИТЕРАТУРЫ

1. Технологии анализа данных: Data Mining, Text Mining, OLAP / *А. А. Берсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод*. 2-е изд., перераб. и доп. СПб: БХВ-Петербург, 2008. 384 с.
2. Применение ЭВМ в технологической подготовке серийного производства / *С. П. Митрофанов, Ю. А. Гульнов, Д. Д. Куликов, Б. С. Падун.* М.: Машиностроение, 1981. 287 с.
3. Технологическая подготовка гибких производственных систем / *С. П. Митрофанов, Д. Д. Куликов, О. Н. Миляев, Б. С. Падун*. М.: Машиностроение, 1987. 352 с.
4. Автоматизированные системы технологической подготовки производства в машиностроении / Под ред. *Г. К. Горанского*. М.: Машиностроение, 1976. 240 с.

Сведения об авторе

Юлия Евгеньевна Котельникова — Санкт-Петербургский государственный университет информационных технологий, механики и оптики, кафедра технологии приборостроения; ассистент; E-mail: jkt1977@mail.ru

Рекомендована кафедрой технологии приборостроения

Поступила в редакцию 14.12.09 г.