СРЕДСТВА ОБРАБОТКИ ИНФОРМАЦИИ ПРИ ИСПОЛЬЗОВАНИИ ТЕХНОЛОГИЙ DATA MINING

Петрова Д.Н.

Минский государственный лингвистический университет

Аннотация. В данной статье описываются инструменты Data Mining, основные элементы методов интеллектуального анализа. Кроме того, приводятся программные продукты, реализующие методы анализа текста.

Ключевые слова: Data Mining, интеллектуальный анализ данных, текстовый анализ, Image Mining.

Инструменты Data Mining предполагают извлечение ("добычу", "mining") данных и направлены на определение взаимосвязей между информацией, хранящейся в цифровой базе данных предприятия, которые аналитики могут использовать для построения моделей и количественной оценки степени влияния соответствующих факторов. Кроме того, такие инструменты полезны для построения гипотез о вероятном характере взаимосвязей информации в цифровой базе данных предприятия.

Текстовый анализ (TM) – это набор инструментов, позволяющих анализировать большие объемы информации для поиска тенденций, закономерностей и взаимосвязей, которые могут помочь в принятии стратегических решений.

Методы Image Mining (IM) содержат инструменты для идентификации и классификации различных визуальных изображений, хранящихся в корпоративных базах данных или в результате оперативного поиска, полученного из внешних источников информации.

Для решения всех проблем обработки и хранения данных используются следующие методы.

– Создание нескольких систем резервного копирования или распределенной системы управления файлами, которая позволяет хранить данные, но медленно получать доступ к хранимой информации в соответствии с требованиями пользователя.

– Создание системы на базе Интернета, которая обладает высокой гибкостью, но не подходит для реализации поиска и хранения текстовых файлов.

– Внедрение интернет-порталов, которые хорошо приспособлены к требованиям пользователя, но не имеют описательной информации о текстовых данных, загруженных на них.

Системы обработки текстовой информации, свободные от вышеперечисленных проблем, можно разделить на две категории: системы лингвистического анализа и системы анализа текстовых данных.

Основными элементами методов интеллектуального анализа текста являются:

– Резюме (подведение итогов).

– Поиск темы (извлечение признаков).

– Кластеризация.

– Классификация.

– Ответы на вопросы.

– Предметное индексирование.

– Поиск по ключевым словам.

– Создание и поддержка таксономий и тезаурусов.

Программные продукты, реализующие методы анализа текста, включают:

IBM Intelligent Miner for Text – набор отдельных утилит, запускаемых из командной строки или пропускаемых, независимых друг от друга (основной упор делается на механику добычи данных – информационный поиск).

Oracle InterMedia Text – комплекс систем, интегрированных в СУБД, позволяющий наиболее эффективно работать с запросами пользователей (позволяет работать с современными реляционными СУБД в контексте сложного многоцелевого поиска и анализа текстовых данных).

Megaputer Text Analyst – набор COM-объектов, встроенных в программу для решения задач текстового анализа.

Сегодня в области автоматизации управления анализ информации доминирует на предварительном этапе подготовки решений – обрабатывая основную информацию, разбивая проблемные ситуации и позволяя понять только фрагменты и детали процесса, а не общую картину. Чтобы преодолеть этот недостаток, мы должны научиться использовать опыт лучших экспертов для создания базы знаний, а также генерировать недостающие знания.

Использование информационных технологий во всех областях человеческой деятельности, экспоненциальный рост объема информации и необходимость быстро реагировать в любой ситуации требуют поиска соответствующих методов для решения возникающих проблем.