



НАУЧНЫЙ ЖУРНАЛ НАУКА И МИРОВОЗЗРЕНИЕ

ТЕХНОЛОГИИ TEXT MINING И WEB MINING

Гершевич Екатерина Константиновна

студент Сибирского государственного аэрокосмического университета им. ак. М.Ф. Решетнева, РФ, г. Красноярск

Кукарцев Владислав Викторович

научный руководитель, канд. техн. наук, доц., инженерно-экономический факультет, Сибирский государственный аэрокосмический университет им. ак. М.Ф. Решетнева, РФ, г. Красноярск

На сегодняшний день все большую популярность приобретает термин “Big Data”, объединяющий в большинстве случаев информацию, которая находится как в структурированном, так и неструктурированном виде. И если структурированные данные поддаются автоматической обработке, то данные, которые относятся к неструктурированному виду, лишены такой возможности.

В большинстве случаев, около 80% корпоративной, управляющей и регулирующей информации хранится в неструктурированном виде. Данное обстоятельство в значительной мере затрудняет поиск необходимых сведений, что и обуславливает спрос на различные специализированные программные решения для обработки такого рода данных.

В настоящее время существует множество программных продуктов, целью которых является автоматизация информационно-аналитической деятельности. Все они различаются между собой по количеству и качеству выполняемых функций, производительности и другим показателям эффективности. Принимая за основу функциональные возможности данных программных решений, их можно разделить на следующие классы:

- Информационно-поисковые системы;
- Системы анализа текстовой информации;
- Лингвистические процессоры;
- Системы визуализации структурированной информации;
- Системы анализа структурированной информации.

Несмотря на все разнообразие проектных решений, наибольшая эффективность в обработке неструктурированных данных достигается за счет интегрированного применения аналитических систем различных классов и технологий Text Mining и Web Mining. Суть технологии Text Mining лежит на пересечении таких областей, как поиск информации, Data Mining, машинное самообучение, статистика и компьютерная лингвистика. Данная технология глубинного анализа текстов дает возможность обрабатывать большие объемы неструктурированной информации и извлекать знания и высококачественную информацию

из текстовых массивов. Данное обстоятельство в значительной степени сокращает время- и трудозатраты человека на поиск необходимых данных.

Результаты, полученные благодаря использованию технологии Text Mining, могут быть использованы для математического прогнозирования, анализа уровня напряженности социальной обстановки и анализа рынков. Широкие функциональные возможности данной технологии открывают целый спектр отраслей, в которых она может быть применена, начиная от корпоративной бизнес-аналитики и мониторинга социальных медиа до национальной безопасности и разведки.

Основными стадиями и задачами технологии Text Mining являются:

1. Поиск информации и определение исходных данных. Этот этап является подготовительным, суть которого сводится к сбору текстовых материалов для анализа. Такие материалы могут содержаться в базах данных, файловых системах и различных Web-ресурсах.
2. Использование четко формализованных статистических и лингвистических методов анализа.
3. Выявление смыслов. На данном этапе выявляются именованные сущности (наименования организаций, мест, имена людей и т.д.) с помощью различных техник статистического анализа. Использование контекста позволяет определить обозначение того или иного слова в конкретном вхождении.
4. Выявление шаблонов.
5. Выявление перекрестных ссылок.
6. Выявление взаимосвязей, фактов и событий.
7. Смысловой анализ. Данный этап предназначен для выявления различных форм оценочной информации, таких как мнения, настроения и эмоции.
8. Количественный анализ.

Этот этап является заключительным и предназначен для извлечения семантических или грамматических связей между отдельными словами для понимания смысла стилистических шаблонов, проведения психологического профилирования и т.д. Это достигается за счет использования комплекса техник, заимствованных из социальных наук. Современный Интернет изобилует различного рода знаниями и информацией, что осложняет процесс поиска необходимых данных. Именно поэтому возникает такая острая необходимость в использовании специализированных программных решений для извлечения полезных знаний и высококачественной информации из сети Интернет. Технология Web Mining может успешно служить данным целям и окончательно решить эту проблему.