Программно-аппаратный комплекс интеллектуального поиска и анализа больших массивов текстов TextAppliance

И. В. Смирнов
ООО «Технологии системного анализа»,
Москва, Россия

TextAppliance представляет собой программно-аппаратный комплекс, состоящий из сервера (или группы серверов, объединенных в кластер) и интеллектуальных сервисов поиска и анализа больших коллекций текстовых документов. Основными сервисами являются:

1. Формирование и индексация больших массивов текстовых документов из интернета, баз данных, корпоративных хранилищ и т.д.
2. Семантический, фразовый и эксплоративный поиск.
3. Поиск тематически и семантически похожих документов.
4. Семантический поиск текстовых заимствований, в том числе скрытых и сильно перефразированных.
5. Быстрая кластеризация и классификация документов и коллекций.
6. Формирование, сопоставление и анализ пользовательских коллекций документов.
7. Тематический анализ коллекций документов – выявление динамики публикаций по заданным темам в разных коллекциях на временной шкале.
8. Автоматическое формирование ключевых слов для документов и коллекций.
9. Автоматическое реферирование документов.
10. Анализ качества научных текстов – проверка на соответствие формальным требованиям к научным публикациям.

В TextAppliance реализованы современные лингвистические и статистические методы, которые позволяют обрабатывать тексты с высоким качеством. Используются параллельные вычисления, за счет чего достигается высокая производительность и масштабируемость системы.

При помощи TextAppliance можно автоматизировать широкий спектр бизнес-процессов и решить ряд задач, которые в настоящее время решаются с применением большого числа аналитиков и различных инструментов. TextAppliance разработан для сегментов B2B/B2G и предназначен для клиентов, обладающих или имеющих доступ к большим массивам текстовых документов.

Потенциальными пользователями TextAppliance являются:

- Коллекторы электронных документов.
- Крупные издательства.
- Электронные библиотеки.
- Компании, специализирующиеся на защите интеллектуальной собственности.
- Любые организации, в которых существует потребность в интеллектуальных сервисах анализа большого количества электронных документов.

TextAppliance интегрируется в инфраструктуру организации и предоставляет различные сервисы по работе с коллекциями заказчика. TextAppliance имеет демонстрационный веб-интерфейс и API. Программные обращения к TextAppliance осуществляются по протоколу JSON/XML-RPC. TextAppliance поддерживает все распространенные форматы электронных документов, содержит средства распознавания PDF без текстового слоя, работает с документами на русском и английском языках, а также документами, написанными сразу на двух языках. TextAppliance имеет возможность прозрачного масштабирования с 1 сервера до нескольких сотен или тысяч серверов.

Основным конкурентным преимуществом TextAppliance является уникальный набор сервисов, который не имеет аналогов на рынке. Не требуются установка и настройка многих приложений по распознаванию, поиску, анализу текстовых заимствований и ряда других сервисов – все это интегрировано в TextAppliance и работает на одной информационной базе.

TextAppliance является интеллектуальным инструментом для создания систем анализа текстов и не предназначен для конечных пользователей. Через API возможна интеграция TextAppliance с практически любыми имеющимися программными системами и базами данных.

Подробнее о функциональных возможностях TextAppliance и условиях его распространения можно узнать на сайте http://textapp.ru/. Демонстрационная версия TextAppliance доступна по адресу http://demo.textapp.ru/. С ее помощью можно опробовать основные функции на тестовых коллекциях, в которые входят российские и зарубежные научные журналы, труды конференций, патенты и авторефераты диссертаций.

В докладе будут показаны примеры применения TextAppliance для решения задач в различных предметных областях.