TextAppliance – новое решение для интеллектуального поиска и анализа больших массивов текстов

TextAppliance – a new solution for intelligent search and analysis of large-scale text collections

Г. С. Осипов, И. В. Смирнов, И. А. Тихомиров, И. В. Соченков
Институт системного анализа Федерального исследовательского центра «Информатика и управление» РАН,
Москва, Россия

Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov and Ilya Sochenkov
Institute for Systems Analysis of FRC CSC, Russian Academy of Sciences,
Moscow, Russia

В докладе представлен TextAppliance – программно-аппаратный комплекс интеллектуального поиска и анализа больших массивов текстов, который предназначен для автоматизации деятельности компаний и различных учреждений, которые работают с большими коллекциями электронных документов. Рассмотрены основные функции и архитектура программно-аппаратного комплекса.

The authors introduce TextAppliance – a hardware-software solution for intelligent search and analysis of large-scale text collections. The main purpose of TextAppliance is automation of companies and organizations that need intelligent services for large-scale text collections analysis. Functions and system architecture are also under consideration.

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 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.

Указанные сервисы используют ситуационно-реляционную модель текста [1], специализированные структуры данных, индексы и ряд оригинальных авторских методов [2].

В TextAppliance реализованы современные лингвистические и статистические методы, которые позволяют обрабатывать тексты с высоким качеством. Все модули используют 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.

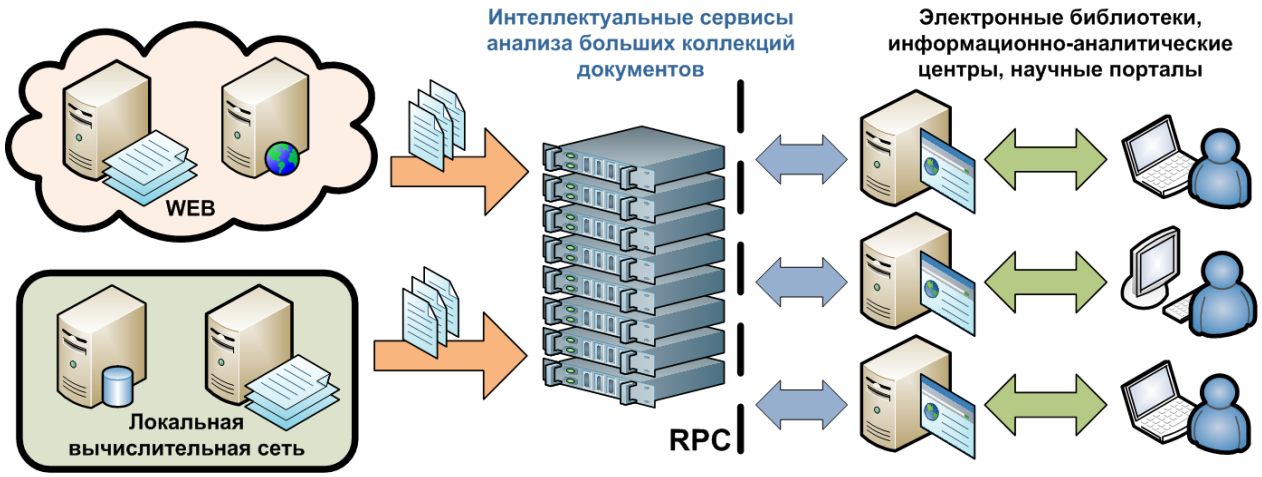

Рис. 1 – Схема интеграции TextAppliance

TextAppliance интегрируется в инфраструктуру организации и предоставляет различные сервисы по работе с коллекциями заказчика. Схема интеграции TextAppliance представлена на рис. 1. 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. На одном сервере TextAppliance может быть проиндексировано до 2 млн. документов, при этом TextAppliance имеет возможность прозрачного масштабирования с 1 сервера до нескольких сотен или тысяч серверов [3].

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.

Литература

1. Осипов Г.С., Смирнов И.В., Тихомиров И.А. Реляционно-ситуационный метод поиска и анализа текстов и его приложения //Журнал "Искусственный интеллект и принятие решений". Номер 2-2008. – С. 3-10.
2. Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Artem Shelmanov Relational-Situational Method for Intelligent Search and Analysis of Scientific Publications // Proceedings of the Integrating IR technologies for Professional Search Workshop, Moscow, Russia, 24-March-2013, pp. 57-64.
3. Зубарев Д.В. Тихомиров И.А. Платформы межкомпонентного взаимодействии в поисково-аналитических системах: состояние и перспективы// Журнал "Информационные технологии и вычислительные системы". Номер 1-2013. – С. 11-20.