Программно-аппаратный комплекс интеллектуального поиска и анализа больших массивов текстов

System for intelligent search and analysis of large text collections

Г. С. Осипов, И. В. Смирнов, И. А. Тихомиров, И. В. Соченков
Институт системного анализа РАН,
Москва, Россия

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

В докладе представлена информация о программно-аппаратном комплексе 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 paper presents the system for intelligent search and analysis of large text collections. The system is designed to computerize companies and organizations processing vast arrays of digital documents. Functions and system architecture are also under consideration.

Программно-аппаратный комплекс интеллектуального поиска и анализа больших массивов текстов (ПАК) позволяет решать ряд задач, связанных с текстовой аналитикой и построен на базе технологии Exactus Expert [1]. ПАК состоит из сервера (или группы серверов, объединенных в кластер) и интеллектуальных сервисов анализа больших коллекций текстовых документов. Основными сервисами являются:

1. Семантический и эксплоративный поиск.
2. Поиск тематически похожих документов.
3. Семантический поиск текстовых заимствований.
4. Формирование, сопоставление и анализ пользовательских коллекций документов.
5. Тематический анализ коллекций документов.
6. Автоматическое формирование ключевых слов для документов и коллекций.
7. Автоматическое реферирование документов.
8. Анализ качества научных текстов.
9. Подсветка слов запроса в документах.

Указанные сервисы используют ситуационно-релялионную модель текста[2], специализированные структуры данных и индексы и ряд оригинальных авторских методов [3].

При помощи ПАК можно автоматизировать широкий спектр бизнес-процессов и решить ряд задач, которые в настоящее время решаются с применением большого числа аналитиков и различных инструментов. Архитектура ПАК представлена на рис. 1.

Рис. 1 – Архитектура программно-аппаратного комплекса интеллектуального поиска и анализа больших массивов текстов

ПАК интегрируется в инфраструктуру организации и предоставляет различные сервисы по работе с коллекциями заказчика. ПАК имеет демонстрационный веб-интерфейс, все обращения к ПАК осуществляются по протоколу JSON/XML-RPC. ПАК 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 ПАК может быть проиндексировано до 2 млн. документов, при этом ПАК имеет возможность прозрачного масштабирования с 1 сервера до нескольких сотен или тысяч серверов [4].

Основным конкурентным преимуществом ПАК является уникальный набор сервисов, который не имеет аналогов на рынке. Не требуется установка и настройка целого ряда приложений по распознаванию, поиску, анализу плагиата и ряда других сервисов – все это интегрировано в ПАК и работает на одной информационной базе.

Литература

1. Тихомиров И.А., Смирнов И.В., Соченков И.В., Девяткин Д.А., Шелманов А.О., Зубарев Д.В., Швец А.В., Лешкин А.В., Суворов Р.Е. Exactus Expert: Поисково-аналитическая система поддержки научно-технической деятельности // Труды тринадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2012. Б.: БГТУ, 2012. т. 4. – С. 100–108.
2. 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.
3. 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.
4. 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.