Обзор архитектуры UIMA современных вопрос-ответных систем

Золотин Игорь Андреевич
Магистр МТУ (МГУПИ), г. Москва
goldin7777@gmail.com

Аннотация. Данная статья посвящена обзору технологии управления неструктурированной информацией и соответствующая архитектура UIMA.

Введение. Вопрос-ответные системы предназначены для поиска точных ответов на вопросы, поставленные на естественном языке (Natural Language Processing, NLP). Важно подчеркнуть, что речь идет о точных ответах, человек-пользователь должен иметь возможность для однозначной интерпретации ответа, поэтому ответ может сопровождаться какой-то детализирующей или конкретизирующей информацией. Источником сведений могут быть неструктурированные данные (книги, журналы, Web-страницы, блоги), квазиструктрированные (справочники, словари, энциклопедии, вики и ее аналоги) и базы данных.

Обзор. Технология управления неструктурированной информацией (Unstructured Information Management, UIM) и соответствующая архитектура UIMA (Unstructured Information Management Architecture) разрабатывалась в IBM Research еще с 90-х годов. Деятельность была сосредоточена на средствах для работы с NLP и включала поддержку диалога на естественном языке, выделение полезной информации, анализ текстов, классификацию документов, машинный перевод и вопрос-ответные системы. Итогом стало создание связующего ПО, получившего название UIMA, которое может служить ядром для создания и внедрения распределенных аналитических машин (analysis engine), или UIM-приложений, позволяющих извлекать знания из неструктурированной информации, в том числе из текстов, аудио, видео и изображений.

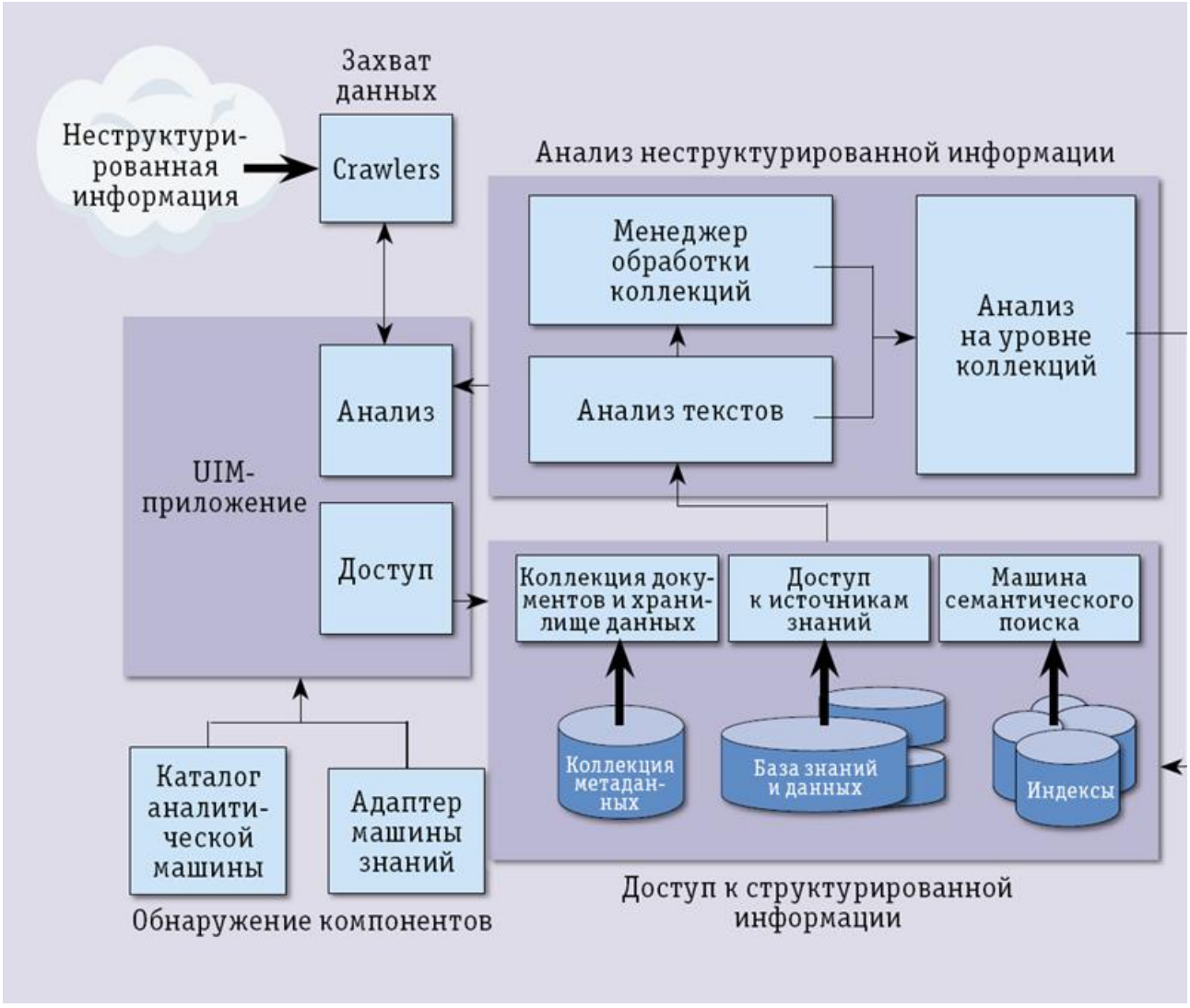

Рис. 1. Архитектура UIMA.

Структура UIMA (рис. 1.) состоит из нескольких компонентов:

1. Захват данных (Acquisition) обеспечивает сбор документов из разных источников и формирование необходимых коллекций (collection), предназначенных для определенных приложений. Функцию захвата могут, например, осуществлять Web-пауки (web crawler), а также иные средства, какие именно, для приложений не важно, поскольку имеется специальный уровень интерфейса Collection Reader, связывающий приложения с коллекциями данных и метаданных.
2. Анализ неструктурированной информации (Unstructured Information Analysis) делится на два последовательных этапа — сначала выполняется анализ документов, а затем анализ коллекций документов. Входные документы обрабатываются текстовыми аналитическими машинами (Text Analysis Engine), в том числе трансляторами и модулями, выполняющими грамматический разбор, классификацию, обобщение. Используя входные документы, текстовые аналитические машины вырабатывают обобщенные аналитические структуры (Common Analysis Structure). На этап анализа коллекций документы могут поступать напрямую или через промежуточный этап, на котором выполняется необходимая фильтрация и переформатирование для последующей параллельной обработки. Анализ на уровне коллекций (Collection Level Analysis) позволяет обобщить сведения, содержащиеся в коллекции документов.
3. Анализ структурированной информации (Structured Information Analysis) используется как для

входных данных, поступающих в структурированной форме, так и для данных, появляющихся после анализа неструктурированной информации, где их значительная часть структурируется, с тем чтобы к ним можно было применить известные методы анализа. В результате аналитические механизмы, предназначенные для двух типов данных, оказываются охваченными общей петлей обратной связи.

4. **Заключение.** Сейчас суперкомпьютер IBM Watson способен находить ответы на 85% вопросов в течение 5 секунд, причем все его основные компоненты — это стандартные серверы под управлением открытой операционной системы, использующие технологии Hadoop и UIMA.