ПРОБЛЕМА ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ТЕКСТОВ

Молнина Е.В., Картуков М.С.
Юргинский технологический институт (филиал) Томского политехнического университета
Юрга, Кемеровская область, Россия
E-mail: molnina@list.ru, maxkartukov@mail.ru

В России, как и во многих странах мира, образование перестает быть средством усвоения готовых общепризнанных знаний. Любой образовательный процесс, как для студента так и для преподавателя связан с переработкой больших объемов информации с целью выборки из нее действительно полезной, нужной, т.е уменьшающей степень неопределенности в той или иной области знаний.

Проблема быстрой и качественной обработки больших массивов текстовой информации актуальна для всех форм образования: дистанционного, заочного, очного и пр. Несомненно, в помощь каждому из нас различные поисковые системы, предлагаемые глобальной сетью Интернет. Современные библиотеки оснащены различными поисково-информационными системами, помогающими посетителям осуществить быстрый выбор необходимых книг, статей по определённой тематике. Область применения существующих систем анализа русских и естественно-языковых (ЕЯ) текстов достаточно разнообразна. Обобщённо можно выделить поисковые системы, вопросно-ответные системы, системы автоматизированного машинного перевода.

Авторами проведен анализ существующих технологий обработки ЕЯ текстов. Одна из устоявшихся - спиральная модель обработки ЕЯ текстов, программно реализованная и апробированная, с указанием математических, лингвистических, логических и других методов, а также компетенции для каждого этапа модели. Модель включает поэтапное применение, а в перспективе — циклическое повторение графематической, морфологической, фрагментационной, синтаксической, семантической, прагматической, логико-интуиционистской и диалоговой вех.

Задача поиска текстовой информации заключается в нахождении минимальных смысловых единиц текста, которые релевантные запросу. Найденные единицы должны отвечать требованиям полноты и точности. Под релевантностью понимается некая бинарная функция, входными параметрами которой являются запрос пользователя и очередная анализируемая единица текста. За информационную единицу текста может приниматься документ, абзац, предложение или другие фрагменты текста. Функция калькуляции релевантности выдает численное значение на отрезке от 0 до 1, которое вычисляется по особым алгоритмам. Проблема вычисления релевантности является центральной в задаче поиска текстовой информации. В идеальной поисковой системе релевантность должна вычисляться так же, как бы ее вычислил человек, если бы проводил поиск.

В начале поисковой сессии для каждой информационной единицы вычисляется его значение релевантности введенного пользователем запросу. По окончании поиска пользователю выдается список ссылок на информационные единицы, ранжированные по убыванию вычислительного значения релевантности.

Критерии, использующиеся для оценки и вычисления релевантности документа запросу, применяемые в современных поисковых системах, относятся к статистическим характеристикам анализа текстовых данных. Это: морфологическое расширение области поиска, близость и порядок слов в тексте, наибольшее совпадение очередности слов в запросе и в тексте, частота слов запроса, МЕТА-данные, индекс цитирования и др.

Поиск по ключевым словам является лишь первым приближением в решении задачи поиска текстовой информации, поскольку по наличию или отсутствию слов запроса в документе нельзя однозначно судить о релевантности последнего. Перечисленные критерии и методы позволяют проводить первоначальный отбор анализируемых текстовых данных, хотя получаемые результаты не отвечают требованиям точности и полноты. Очевидно, что наличие слов в документе в различных морфологических формах — далеко не свидетельство о том, что в данном документе можно найти ответ на заданный вопрос.

Проблема, которую бы хотели высветить авторы не может быть решена приведенными выше технологиями. Потребностям современного человека по обработке текстовой информации не помогут не ключевые слова, не простые шаблоны. Сегодня решение этой проблемы, как считают авторы, может быть связано только с применением систем искусственного интеллекта.

Таким образом в настоящее время одной из проблемных задач в области информационных технологий и искусственного интеллекта является задача по извлечению информации (смысла) из текста или — более широко — задача понимания текста.

Авторами были рассмотрены некоторые существующие технологии, предназначенные для решения вышеназванной задачи (такие как ТОМАТ, Абриаль, Alex).

ТОМАТ — Технология Объектно-ориентированного Многовариантного Анализа Текста. Данная технология использует объектно-ориентированный подход к построению системы шаблонов, а также использованию концепции недоопределенных вычислений для полей классов.

Следующая технология лексического анализа ALEX позволяет с помощью настраиваемых лексических шаблонов произвольной сложности решать следующие задачи:

- поиск в текстовых массивах различной степени структуризации определенных фрагментов, извлечение знаний;
- нормализация слабоструктурированных массивов данных, как с точки зрения структуры, так и с точки зрения качества их наполнения.

Технологию Абриаль можно рассматривать как принципиально новое хранилище данных, в котором пользователь работает с информацией через автоматически формируемый интерфейс гипертаблиц, позволяющий осуществлять навигацию в любых направлениях, отражающих ассоциативные связи данных. Абриаль предназначен для исследования семантических сетей, лингвистических баз данных. Эта технология лексического анализа помогает

быстро строить сложные сетевые базы знаний (или базы данных). Основная задача Абриаля состоит в предоставлении пользователю быстрого, удобного и гибкого доступа к сложной динамической структуре данных/знаний.

Перечисленные выше технологии используют разные методы (гипертекст, шаблон, объектно-ориентированный многовариантный анализ текста). Релевантность поиска в них повышается за счёт построениея некоторой информационной структуры текста, представляющей собой, например, список концептов (не слов и словосочетаний, а именно понятий) и сравнение этой структуры с некоторым эталонным образцом, который заведомо соответствует искомому типу текстов.

Несмотря на глубину анализа, общими, ключевыми моментами остаются понятия образца и структуры. От сюда следует вывод, что в целом решение проблемы анализа русских и ЕЯ текстов пока далеко от диктуемого практической потребностью. Каждая из этих технологий по обработке текста применима в определённом направлении. Как правило, в подобных технологиях, решение строится индивидуально под каждую конкретную задачу, где требуется извлечение смысла из текста; подходы, инновации и ноу-хау глубоко зашиты в программный код и не переносимы на другие схожие задачи.

Рассмотренные технологии обработки текстов, хотя и решают многие задачи, связанные с их обработкой, но не могут удовлетворить запросы участников образовательного процесса (как студентов, так и преподавателей), стать ежедневными помощниками в обработке больших объемов информации. Авторы ставят перед собой задачу по разработке технологии, облегчающей рутинный труд тех, кому необходимы знания.