Р. Р. Мавлютов (Уфа, ГОУ ВПО УГНТУ)**. Особенности концептуальной модели автоматического анализа текста.**

В данной работе предлагается новая технология, обеспечивающая более эффективную обработку массивов технических текстов на основе достижения технической, программной, информационной и методологической совместимости. Данная технология предусматривает обработку данных в модели клиент/сервер, хранение данных, формирование базы знаний, выполнение приложений, т. е. выборка и обработка данных для нужд прикладной задачи, а также представление данных и результатов обработки конечному пользователю.

Автоматический анализ текста начинается с его первичной обработки. В задачи этого этапа входит фрагментация и графематический анализ потока символов, которым изначально является текст.

Первичный анализ текста содержит много подводных камней и, как всякий лингвистический алгоритм, представляет собой довольно сложный и громоздкий механизм. Для решения подобных задач традиционные методы алгоритмизации и программирования оказываются малоэффективными. Результаты работы программы сильно зависят от особенностей текста. При попытке усложнения поставленных задач число условий (различных возможных ситуаций) начинает возрастать в геометрической прогрессии, алгоритм становится перегруженным, сложным для восприятия и отладки. Растет число ошибок. В такой ситуации целесообразно применение методов экспертных систем.

Имеющиеся знания об особенностях организации текста представляются в виде набора продукционных правил, а процесс первичной обработки осуществляется методом построения дерева решений. Число продукционных правил можно изменять в широких масштабах в зависимости от поставленной задачи и особенностей текста.

Применяется объектная модель, где объект имеет свое название, описание и набор характерных черт, по которым его можно идентифицировать в тексте. Последовательности символов, подошедшие под описание объекта, помечаются его именем. Фрагменты текста, не подошедшие по описанию ни к одному объекту, не включаются в результаты анализа. Таким образом, объектная модель представляет собой совокупность понятий, объединенных иерархическими связями. Понятие или объект представляет собой ярлык, которым могут быть помечены отдельные участки текста. Эти ярлыки характеризуют ту информацию, которая была получена в результате анализа. Объектная модель является чем-то вроде спецификации, в которой описывается, какие элементы будут присутствовать в результатах анализа, что они должны означать для пользователя, и как они будут структурированы.

Данная методика не была бы достаточно эффективной, если бы не две ее особенности: допущение многозначности и иерархические обобщения. Это обусловлено следующими причинами. Одной из главных проблем при анализе технических текстов является неоднозначность. Определенная последовательность символов может подходить под разные шаблоны, которые в итоге определяют разные единицы текста вышележащих уровней.

На каждом уровне анализа единицы текста классифицированы по их смысловой нагрузке и роли в тексте. Классификация представляет собой иерархическую систему каталогов (групп), по которым распределены единицы текста. Корневой группой является понятие единица текста. Она разбита на подгруппы, элементы которых объединены по общим характерным признакам. Единица текста, находящаяся в определенном месте иерархии обладает признаками всех каталогов, в которые она вложена. Анализатор представляет собой программу, которая, двигаясь по последовательности единиц текста определенного уровня, формирует следующий уровень. Принятие решения осуществляется исходя из имеющейся совокупности шаблонов данного уровня в рабочей объектной модели. На каждом этапе своего движения анализатор рассматривает некоторую область уровня на соответствие имеющимся шаблонам. Область

исследования постоянно меняет свое расположение и размеры. После того как анализатор сформировал все уровни текстовых единиц, в работу вступает парсер. Эта программа, которая, двигаясь по верхнему уровню, составляет результирующий отчет согласно требованиям, установленным в объектной модели результатов анализа.

Разработанная экспертная система позволяет извлекать максимум информации из массивов технических текстов. Архитектурные особенности системы позволяют быстро изменять глубину обработки текста и тип извлекаемой информации путем смены набора продукционных правил.