Е. Годлевская

ПРИНЦИПЫ СОЗДАНИЯ ПОИСКОВОГО ОБРАЗА АНГЛОЯЗЫЧНОГО ПУБЛИЦИСТИЧЕСКОГО ТЕКСТА

В настоящее время системы автоматической обработки текста, определяющие его идею и структуру, малочисленны. Данные системы определяют только смысл отдельных слов или словосочетаний. Системы, анализирующие семантическую композицию текста на уровне выше предложения, до сих пор не созданы.

Процесс представления семантического содержания текста с помощью средств информационно-поискового языка (ИПЯ) относится к индексированию. Конечный результат процесса находится в поисковом образе документа.

Поисковый образ документа – поисковый образ, отражающий основное семантическое содержание документа.

В Интернете используются два вида ИПЯ: координатные и булевые. Применяя координатный поиск, запрос рассматривается как список терминов, которые называются ключевыми словами. Данный список соотносится со списками терминов документов.

Используя булевый поиск, слова и словосочетания связываются операторами булевой алгебры, которые называют логическими коннекторами.

В системах поиска имеются списки «stop»-слов и общих слов. «Stop»-слова обозначают общеупотребительные слова языка. Общие слова – слова, которые характеризуются высокой частотой встречаемости в документах. «Stop»-слова и общие слова не включаются в поисковый образ документа.

Анализируя англоязычные публицистические тексты, мы отметили, что процесс индексирования охватывает несколько основных и второстепенных операций. В число главных операций входят: 1) анализ содержания документа и выделение из текста номинативных лексических единиц, важных по отношению к его содержанию; 2) создание списка ключевых слов; 3) избыточное индексирование.

Индексирование является одним из главных понятий информационного поиска. Назначение этого процесса в информационной системе заключается в том, чтобы присвоить каждому документу и запросу определенный ряд «индексов». Данные «индексы» отображают содержание документа и регулируют поиск.

Процесс индексирования и создания поискового образа документа является действительно важным в сфере компьютерной лингвистики. Подобные системы требуются для обработки публицистических текстов, поскольку в настоящее время насчитывается огромное число текстов публицистической тематики. Обработка таких текстов вручную является довольно длительным и трудоемким процессом для человека. В связи с этим определение основного содержания и формирование современных систем создания поискового образа документа компьютером находится в процессе разработки.