ЧАСТОТНЫЙ АНАЛИЗ КАК ПЕРВЫЙ ШАГ В ПОСТРОЕНИИ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ИССЛЕДОВАНИЯ ТЕКСТА

О. В. Головань
Алтайский государственный технический университет им. И. И. Ползунова
г. Барнаул

Интеллектуальные системы исследования текста позволяют, анализируя смысловое и словарное содержание текста, представленность в нем отдельных слов и словосочетаний, частоту их употребления и пр., устанавливать основную идею или скрытый смысл, вложенный автором в произведение. Эксперименты такого рода часто применяются в прикладной социологии, психологии, педагогике [1].

Для оптимизации работы такой системы слова и словосочетания исходного текста можно разбивать на группы по частоте их употребления, признакам соответствия одному гнезду, однокоренности и др., выстраивая соответствующие зависимости. Даже простой сопоставительный анализ таких частотных диаграмм позволит решить ряд задач по идентификации авторства, нацеленности текста на определенный императив и пр. Таким образом, частотный анализ является первой ступенью интеллектуальной обработки и исследования текста.

На основе возможностей сервера Fire Bird v. 1.0 был разработан комплекс компьютерных программ и база данных (**БД**), предназначенные для исследования смыслового компонента различных концептов, выявления уровня понимания содержания текста респондентом и программа-опросник для тестирования. В комплекс входят: программы для ЭВМ «Фрактальная размерность языка (Language Fractal Dimension)» и «Концепт-анализ (C-Analysis)», а также специализированная база данных «БД Фрактальная размерность языка (DB Language Fractal Dimension)» [2, 3].

Программа «Фрактальная размерность языка» организована по типу анализатора текста, а программа «Концепт-анализ» - опросника. Исходными объектами являются текстовые файлы, слова из которых, после обработки программой, заносятся в базу данных (БД) через используемый сервер. Остановимся подробнее на работе каждой программы в отдельности и использовании комплекса в целом для информационного обеспечения исследований.

Так программа «Фрактальная размерность языка (LangFracDim)» предназначена для определения квантитативных характеристик языка и корреляций между ними. Определение производится путем составления частотного словаря языка и определения ранга и частоты слов. Программа позволяет проводить группировку слов в базе по определенным признакам (темам, алфавиту и т.п.).

Определение корреляций между квантитативными (частотными) характеристиками основывается на установлении зависимости между частотой и рангом слова, параметры которой, после линеаризации, определяются методом наименьших квадратов.

Из сервисных возможностей в программе «LangFracDim» предусмотрены: экспорт базы данных в виде частотного словаря или списка слов, упорядоченных по другому признаку, в текстовый файл, вычисление частоты и ранга слов, представление зависимости между квантитативными характеристиками в виде графика в двойных логарифмических координатах, установление параметров искомой зависимости. Работа программы организована по блочному принципу путем взаимосвязанного выполнения четырех основных алгоритмов: загрузки текста из файла *.txt, разбора текста, переноса слова или списка слов в базу данных (БД), координируемых основным алгоритмом анализа исходного текстового файла (рисунок 1).

Рисунок 1 – Схема алгоритма анализа исходного текстового файла

Загрузка текста в буфер программы осуществляется стандартными средствами, предоставляемыми операционной системой Windows. Так как для хранения данных в БД используется кодировка шрифта WIN1251, исходный текст перед работой программы должен быть в текстовом (*.txt) формате. Поэтому, если используется не исходный электронный текст, а текст на бумажном носителе, то при переводе последнего в электронную форму и сохранению необходимо пользоваться предоставляемыми программами-распознавателями или текстовыми редакторами форматами текстового или rtf-файла.

После заполнения буфера, программой запускается алгоритм разбора текста, позволяющий из общего набора символов кодировки WIN1251 выделить отдельные слова и сформировать на выходе список всех слов текста. Полученный список слов или отдельное слово затем переносится в БД с помощью соответствующего алгоритма переноса.

Блок-схема алгоритма разбора текста приведена на рисунке 2., его работа организована следующим образом. Сначала происходит подсчет общего количества знаков (с учетом пробелов и пустых строк) – *n* в тексте, для чего производится поячеечное сканирование исходного массива символов с помощью встроенных возможностей, предоставляемых библиотеками языка Pascal.

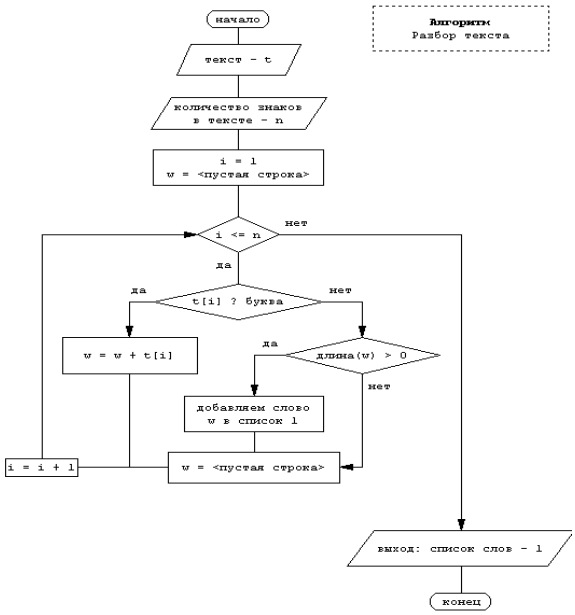

Рисунок 2 – Блок-схема алгоритма разбора текста в список слов

Затем происходит инициализация переменных *i*, *l* и *w*, обозначающих слово, список слов и пустую строку, соответственно, и запускается циклическая процедура проверки каждой ячейки массива текста на условия соответствия элемента букве, слову, пустой строке или пробелу.

Работа цикла завершается, как только все элементы текста, воспринимаемые машиной как отдельные слова (собственно слова как части речи, междометия, предлоги и пр.) не будут сформированы в новый список. На этом этапе работы программы «LangFracDim», путем сравнения кодировок элементов массива с кодами символов (, - запятая), (. - точка), (– - дефис), других знаков препинания (!, «, ‘, “, ?, (, :, ..., и др.), происходит их исключение из списка слов. Объединение символов в слова осуществляется после проверки на наличие внутри цепочки пустой строки или пробела. Таким образом, словами в списке будут являться не только самостоятельные части речи, в той форме, в которой они использованы в тексте, но и предлоги, и другие служебные части речи. После разбора текста, осуществляется перенос полученного списка слов в БД с помощью соответствующего алгоритма (рисунок 3).

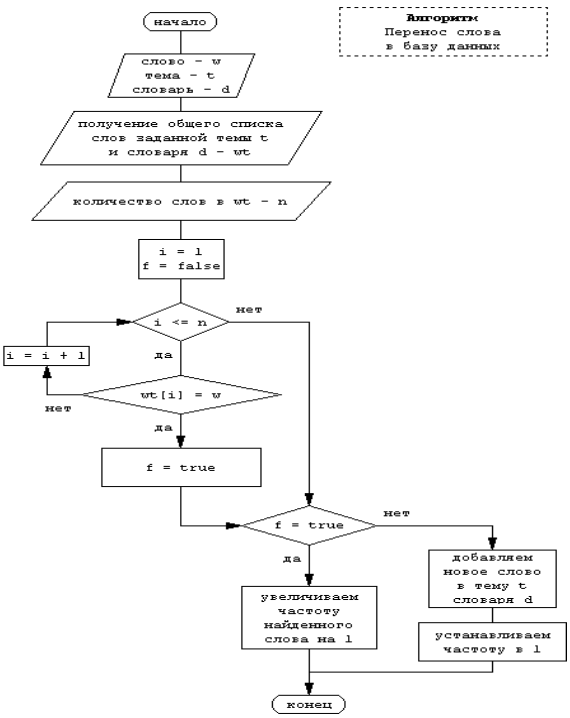

Рисунок 3 – Блок-схема алгоритма переноса слова в БД

Такая организация работы программы и ее взаимодействия с базой данных позволяет ставить в соответствие каждому слову его уникальный номер, связанный с темой и словарем, что позволяет избегать повтора одинаковых слов и ускорять извлечение слова из БД SQL-сервером.

Рассмотрим использование программы на практике для анализа составляющих и восстановления компонентов социокультурного контекста на примере языка.

В качестве одного из индикаторов смыслового компонента концепта, представленного в тексте, нами была выбрана функциональная зависимость между частотой и рангом слова, описываемая законом Ципфа [4]:

$$Ч = k?P^{\alpha}, \qquad (1)$$

где *Ч*– частота встречаемости слова в тексте; *k*– коэффициент пропорциональности; *P*– ранг слова; α– степень развитости и наполняемости текста различными лексическими единицами (для современных языков близок к 1).

Показатель степени в искомой зависимости отражает, как следует из (1), насколько сильно данная категория включена в тезаурус респондента, а значит и в его внутренний мир. Значения же коэффициента пропорциональности показывают, как много слов из предложенных текстов опрашиваемый связывает с конкретной темой (рисунок 4).

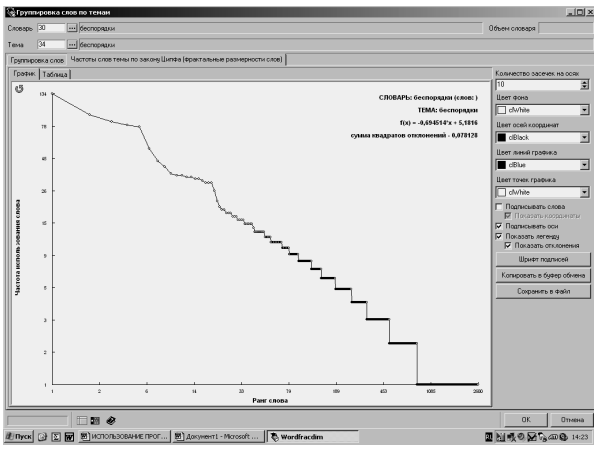

Рисунок 4 – Интерфейс программы «Фрактальная размерность языка»

СПИСОК ЛИТЕРАТУРЫ

1. Zipf G.K. The psycho-biology of language. - Boston, 1935.
2. Св-во рег. пр. для ЭВМ № 2005610982 (RU) // Опубл. Бюлл. № 2. 2005.
3. Св-во рег. пр. для ЭВМ № 2005611226 (RU) // Опубл. Бюлл. № 2. 2005.
4. Св-во рег. пр. для ЭВМ № 2005620308 (RU) // Опубл. Бюлл. № 4. 2005.
5. Головань О.В. Частотный словарь современного языка средств массовой информации. – Барнаул: Изд-во АлтГТУ, 2006.