УДК 004.912
DOI: 10.22213/2410-9304-2017-3-94-99

М. В. Втюрин, магистрант
А. И. Ястребов, магистрант
С. В. Моченов, кандидат технических наук, профессор
ИжГТУ имени М. Т. Калашникова

РАЗРАБОТКА ИНФОРМАЦИОННОЙ СИСТЕМЫ ДЛЯ УМЕНЬШЕНИЯ ОБЪЕМА ТЕКСТОВОЙ ИНФОРМАЦИИ В ПРОЦЕССЕ ИНФОРМАЦИОННОГО ПОИСКА

В статье рассматривается возможность применения пользователями специализированных алгоритмов для информационной системы, обеспечивающей сжатие анализируемой текстовой информации в процессе информационного поиска. Актуальность работы обосновывается сложностью информационного поиска, связанного с решением пользователем конкретной задачи и необходимостью переработки больших объемов текстовых данных. Целью является сокращение объема анализируемой текстовой информации русскоязычных текстов при сохранении их смысловой составляющей. Определены основные функциональные узлы разрабатываемой информационной системы. Модуль поиска совпадений формирует текст, состоящий из нескольких абзацев, содержащих заданные пользователем поисковые словосочетания. Данный текст по объему намного меньше исходного текста и отражает искомую пользователем информацию. Модуль сжатия представляет собой итерационную процедуру, позволяющую дополнительно уменьшить объем текста, выделенный пользователем для анализа. В предлагаемом подходе каждому слову предложения присваивается оценка, определяемая на основе ряда критериев. Разработан графический интерфейс пользователя, имеющий компактные размеры и удобную компоновку элементов. В результате применения описываемого подхода достигается существенное уменьшение объема текстовой информации, обрабатываемой пользователем в процессе информационного поиска. Для большего сокращения объема информации в дальнейшем предполагается проведение разработки модуля сжатия текста и его практическая реализация.

Ключевые слова: обработка текста, информационная система, поисковые слова, сжатие текста, информационный поиск.

С развитием информационных технологий высокими темпами растут объемы информации по различным направлениям науки и техники. В качестве информационных источников могут выступать: лекции, курсовые работы, диссертации, статьи, журналы и другие источники, представленные в текстовом виде. Возникает необходимость обработки пользователем значительных объемов текстовых данных.

Примерами видов обработки текста являются аннотирование и реферирование. Для уменьшения затрат времени на обработку текста используются современные компьютеры и процесс становится автоматизированным.

Таким образом, существует потребность в информационной системе, которая выполняла бы сокращение объема анализируемого текста, облегчая работу пользователя. Это особенно важно при решении различных сложных технических, информационных, интеллектуальных задач. Основной целью разработки описываемой информационной системы является ускорение работы пользователя в процессе информационного поиска при выполнении различных исследований.

В ходе работы были определены основные функциональные узлы разрабатываемой информационной системы. Структурная схема системы представлена на рис. 1.

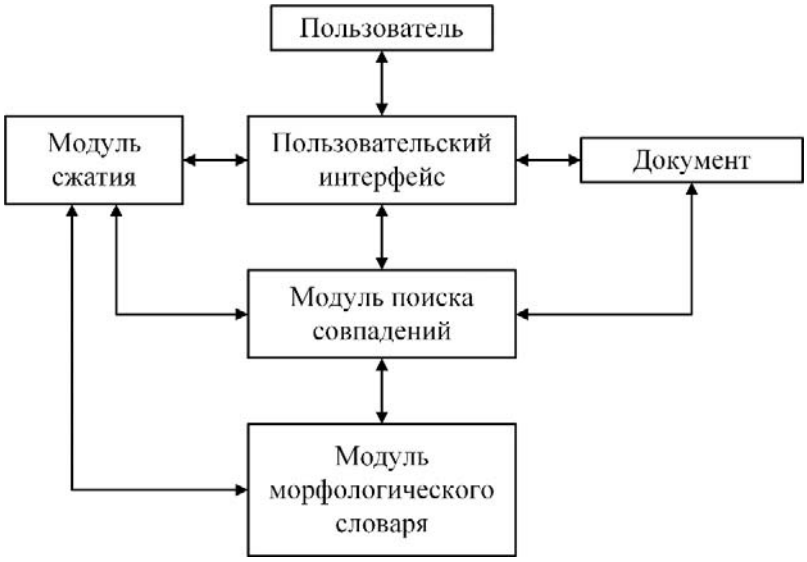

Рис. 1. Структурная схема системы

Взаимодействие пользователя с системой происходит через пользовательский интерфейс, с помощью которого осуществляется ввод поисковых слов или словосочетаний, настраивается работа модулей поиска и сжатия.

Разработка модуля поиска совпадений

При проведении поиска происходит деление текста на предложения, а затем деление предложений на слова. Каждому слову соответствует свой уникальный номер, полученный с помощью модуля морфологического анализа *mcr.dll* [1].

Важно отметить, что различные словоформы одного слова имеют один и тот же номер. Таким образом, на основе сравнения номеров слов текста с номерами поисковых слов оценивается важность предложений, а затем абзацев. Оценка зависит от установленного пользователем порога совпадений: при 100%-м пороге в одном предложении должны присутствовать все поисковые слова. В то же время при пороге, равном 66 %, и трех поисковых словах достаточно присутствия двух слов. Блок-схема общего алгоритма работы модуля поиска приведена на рис. 2. В результате работы модуля поиска совпадений формируется текст, состоящий из нескольких абзацев, в которых есть заданные пользователем поисковые словосочетания. Данный текст по объему намного меньше исходного текста и отражает искомую пользователем информацию. На рис. 3 представлена гистограмма, отражающая число совпадений слов, соответствующих поисковому запросу в абзацах исходного текста.

Модуль сжатия представляет собой итерационную процедуру, позволяющую дополнительно уменьшить объем текста, выделенный пользователем для анализа. В качестве основы для построения модуля сжатия была использована общая схема обработки текстов, предложенная А. М. Бледновым [2]. Эта схема включает в себя определение морфологических характеристик всех слов текста, определение взаимосвязей между отдельными словосочетаниями.

В предлагаемом нами подходе, в отличие от этой схемы, дополнительно каждому слову предложения присваивается оценка, определяемая на основе ряда критериев. В качестве критериев выбираются задаваемый тип части речи, принадлежность слова к поисковому запросу, статистические характеристики всех слов анализируемого текста [3] и ряд других.

Рис. 2. Общий алгоритм работы модуля поиска совпадений

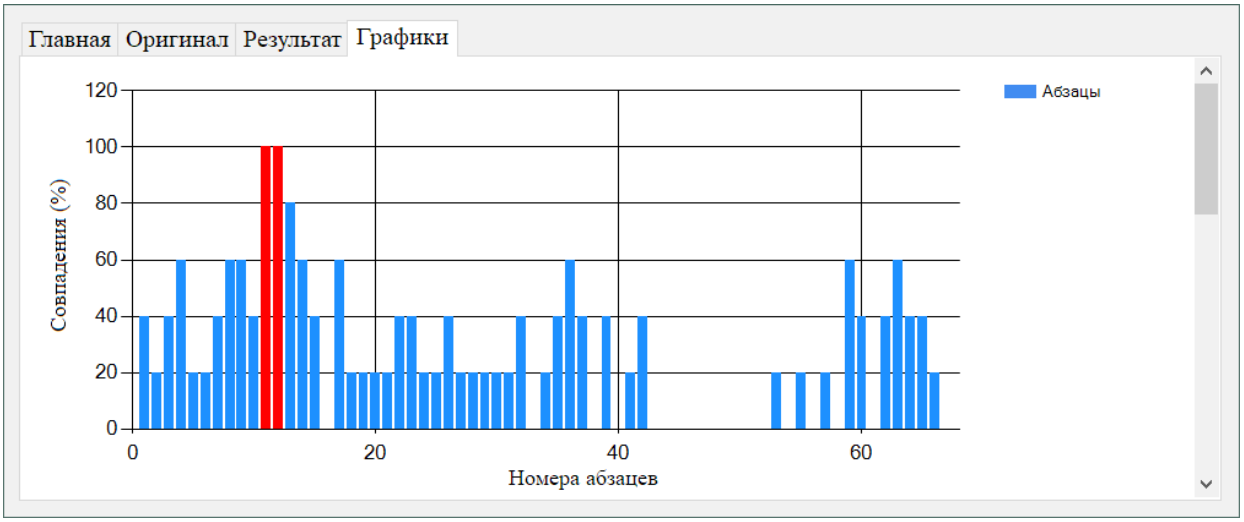

Рис. 3. Гистограмма совпадений слов в исходном тексте

Разработка графического интерфейса

Для удобства отображения информации в разработанном графическом интерфейсе были использованы следующие вкладки: главная страница, оригинальный текст, результат, графики. При этом окно программы, представленное на рис. 4, имеет компактные размеры и удобную компоновку.

Главная страница. Данная вкладка отображает кнопки «Открыть», «Добавить», «Найти», «График», два поля для ввода и отображения поисковых слов и поле для изменения порога совпадений.

При нажатии на кнопку «Открыть» появляется диалоговое окно открытия файла, в котором пользователь должен указать нужный текстовый файл с расширением *.txt*.

При нажатии на кнопку «Добавить» введенные поисковые слова заносятся в отдельное поле ниже. У пользователя есть возможность задать порог совпадений, который по умолчанию равен 100 %.

При нажатии на кнопку «Найти» производится поиск введенных пользователем слов по всему тексту. Абзацы, в которых нашлись совпадения, заносятся во вкладку «Результат».

При нажатии на кнопку «График» во вкладке «Графики» отображаются гистограммы, полученные на основе результатов поиска.

Оригинал. Вкладка содержит исходный текст, выбранный пользователем для анализа.

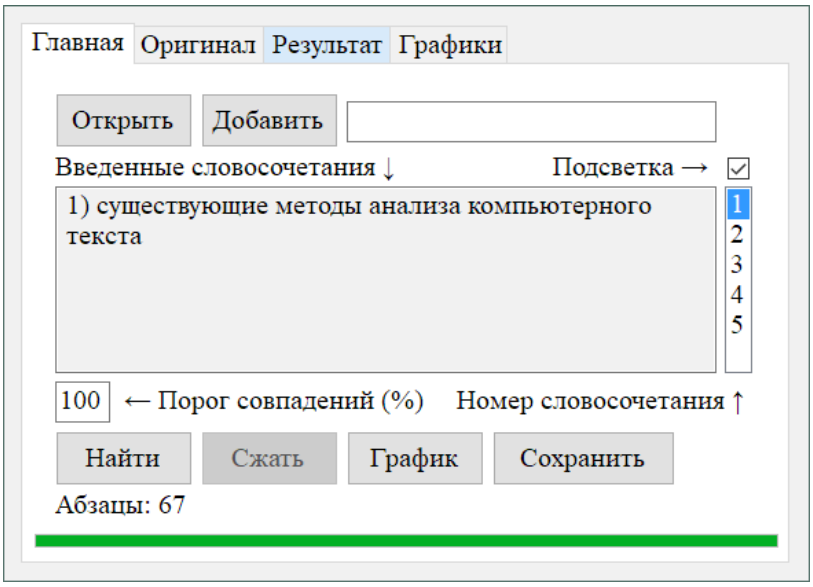

Рис. 4. Интерфейс программы

Результат. Вкладка отображает абзацы оригинального текста, в которых были найдены совпадения с введенными поисковыми словами и установленным пользователем порогом совпадений. На рис. 5 представлен результат информационного анализа выбранного текста.

Графики. Вкладка содержит гистограммы, отображающие совпадения в абзацах текста и разницу между объемами оригинального и полученного текстов.

Программная часть системы реализована в среде разработки *Microsoft Visual Studio* на языке программирования *C#* с использованием дополнительного модуля: *mcr.dll*.

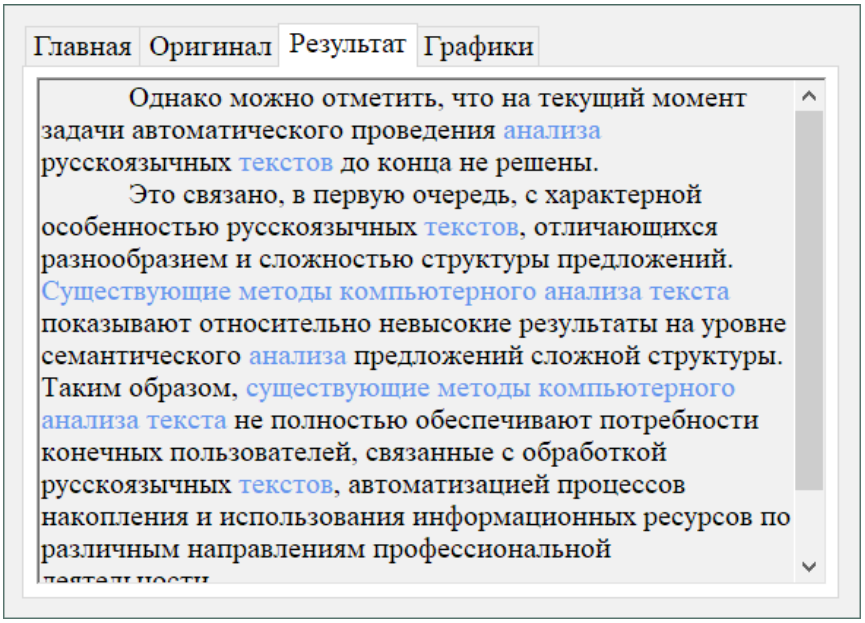

Рис. 5. Результат анализа

В результате применения модуля поиска совпадений обеспечивается значительное уменьшение объема текстовой информации, обрабатываемой пользователем в процессе информационного анализа. Для большего сокращения объема информации в дальнейшем предполагается проведение разработки модуля сжатия текста.

Библиографические ссылки

1. MCR.DLL // Морфоанализ русского языка. – URL: http://macrocosm.narod.ru/madown.html (дата обращения: 12.04.2017).

2. *Бледнов А. М.* Разработка и исследование моделей и информационной технологии семантико-синтаксического анализа русскоязычного текста : дис. … канд. техн. наук. – Ижевск, 2007. – 120 с.

3. *Моченов, С. В.* Применение статистических методов для семантического анализа текста / С. В. Моченов, А. М. Бледнов, Ю. А. Луговских. – Ижевск : НИЦ «Регулярная и хаотическая динамика», 2005.

***

M. V. Vtyurin, Master's Degree Student, Kalashnikov ISTU
A. I. Yastrebov, Master's Degree Student, Kalashnikov ISTU
S. V. Mochenov, PhD in Engineering, Professor, Kalashnikov ISTU

Development of an Information System for Reducing the Volume of Text Information in the Process of Information Search

The paper considers the possibility of applying specialized algorithms for an information system by the users that provides compression of the analyzed text information in the process of information retrieval. The relevance of the work is justified by the complexity of information retrieval associated with the user's solution of a particular task and the need to process large amounts of text data. The goal is to reduce the volume of the analyzed text information of Russian-language texts, while preserving their semantic component. The main functional nodes of the developed information system are determined. The coincidence search engine generates the text consisting of several paragraphs

containing user-defined search phrases. This text is much smaller by volume than the original text and reflects the information that the user wants. The compression module is an iterative procedure that further reduces the amount of the text allocated by the user for analysis. In the proposed approach, each word of the sentence is assigned an estimate, determined on the basis of a number of criteria. A graphical user interface has been developed that has compact dimensions and a convenient layout of elements. As a result of the described approach, a significant reduction in the amount of text information processed by the user in the process of information retrieval is achieved. To further reduce the amount of information in the future, it is proposed to develop a text compression module and its practical implementation.

Keywords: word processing; information system; search words; text compression; information search.

Получено: 08.06.2017