КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 001.103:002

ПРОГРАММНЫЕ СРЕДСТВА ВЫЯВЛЕНИЯ СЕМАНТИЧЕСКОГО ПОЛЯ СЛОВ

И. Е. Воронина, А. А. Кретов, О. С. Титова

Воронежский государственный университет
ЗАО НПП РЕЛЭКС

Поступила в редакцию 12.04.2008 г.

Аннотация. Рассматривается задача создания программного комплекса для получения семантически близких слов на базе алгоритма Е. Л. Гинзбурга. Представлены возможности и структура программного комплекса.

Ключевые слова: автоматическая обработка текста, семантический анализ, алгоритм Е. Л. Гинзбурга

Abstract. The article is devoted to the task of development of the bundled software for discovery of semantically related words based on Ginzburg's algorithm. The article describes the performance capabilities and the structure of the software.

Key words: semantic analisis, Ginzburg's algorithm.

ВВЕДЕНИЕ

Вся информация об окружающем мире может содержаться в текстах и извлекаться из текстов. В этом актуальность задачи их автоматизированного анализа. Задачи обработки текстов возникли практически сразу после появления вычислительной техники. Именно обработка текстовых данных позволила говорить не просто о вычислениях, а об обработке информации. Однако, несмотря на более чем полувековую историю исследований в области искусственного интеллекта, удовлетворительного решения большинства практических задач обработки текста пока нет. Это обусловлено проблемами формализации естественного языка. Естественный язык (ЕЯ) — сверхсложная семиотическая система, состоящая из неограниченного числа подсистем, каждая из которых конечна, а потому формализуема, при этом сам язык — незамкнутая система, которая не может быть формализована до конца [1].

Технологическая цепочка обработки ЕЯ включает в себя различные этапы: морфемный, морфологический, синтаксический, семантический анализ [2]. Семантический анализ — наиболее трудоемкая, тяжело поддающаяся формализации часть исследований по формализации ЕЯ: проблема кореферентности, снятия неоднозначности, формальная структура текста выходят за рамки предложения [1]. На этапе семантической обработки текста происходит выявление смысла предложения или его отдельных частей. Решение такого ряда задач базируются на результатах анализа отдельных слов, поэтому необходима какая-то общая для всех методов анализа база, позволяющая выявлять семантические отношения между словами. Такой основой может быть тезаурус языка.

Тезаурус задает набор бинарных отношений на множестве слов естественного языка (например, омонимия, синонимия, антонимия и т.д.). Для русского языка работа по созданию тезауруса еще не завершена, хотя имеются коммерческие продукты, включающие в себя тезаурус подмножества русского языка, а также отдельные словари синонимов и антонимов для подмножества русского языка.

В качестве базы также применима статистическая обработка данных, позволяющая решать практически важные задачи, среди которых выделение ключевых слов текста. Они не только отражают специфику данного текста, но и служат своеобразным показателем индивидуальности стиля писателя. Их выделение позволяет как выявить тематику текста, так и составить определенное мнение о идиолекте автора.

Для ключевых слов весьма значима нетривиальная задача получения семантического поля слова, то есть определения слов, близких по значению заданному слову (важность такой задачи, например, хорошо иллюстрируется проблемой эффективного поиска информации в различных хранилищах данных).

Все вышесказанное обусловливает актуальность создания программного комплекса для выявления семантически близких слов и построения шкал расстояний между словами. В качестве основы выбран алгоритм Е. Л. Гинзбурга [3].

Для решения задачи необходимо [4]:

1) разработать процедуры, осуществляющие обработку текстового файла, содержащего исходный текст, и построение частотного словаря;

2) разработать фильтр, основной функцией которого является отсеивание слов, составляющих специфику естественного языка в целом;

3) для выделения ключевых слов необходимо:

– автоматизировать методы выделения ключевых слов;

– создать базу данных, содержащую частотные словари текстов, списки ключевых слов для исходного текста, а также некоторую статистическую информацию, необходимую для работы алгоритмов;

– провести разбиение общего множества ключевых слов на группы, что позволяет сократить их количество и выявить слова, наиболее специфичные для данного текста;

– предоставить возможности повторной обработки текста и сравнения списков ключевых слов, полученных применением различных методов обработки;

1. АНАЛИЗ ЗАДАЧИ

Выделение ключевых слов из текстов на естественном языке с последующим обнаружением их контекстного окружения происходит на основе накопленной статистической информации о характеристиках слов в некотором наборе текстов. Формирование статистики осуществляется посредством анализа данных частотного словаря текста и некоторых параметров слов. Эти данные представляют собой основу фильтра, при прохождении которого отсеиваются слова, не составляющие специфики обрабатываемого текста.

Рассмотрим алгоритмы выделения ключевых слов текста.

При фильтрации слов текста могут быть использованы две группы методов (рис. 1). Первая группа включает алгоритмы, при работе которых используется информация о частотных характеристиках в некотором наборе текстов (межтекстовая фильтрация). Способы выделения ключевых слов, относящиеся к этой группе, отличаются друг от друга основанием статистики. Сюда относятся методы межтекстовой фильтрации по относительной частоте, рангам частот, количеству текстов и амплитуде колебания ранга.

Рис. 1. Схема алгоритмов выделения ключевых слов

Вторая группа методов фильтрации слов включает только алгоритм внутритекстовой фильтрации, при работе которого используется информация о частотных характеристиках слов в одном конкретном тексте.

Работа каждого из перечисленных алгоритмов начинается после построения частотного словаря исходного текста, содержащего список слов текста с указанием количества их употреблений (абсолютной частоты).

Каждый из способов выделения ключевых слов более подробно рассмотрен в [5].

Алгоритм выявления контекста (на основе алгоритма Е. Л. Гинзбурга).

1. Находим в тексте Т для каждой словоформы ее частоту — абсолютную и относительную (частное от деления наблюдаемой, абсолютной частоты на количество словоформ в тексте Т).

2. Находим в Т те части этого текста, которые содержат заданное слово С — предложения с этим словом. Для совокупности всех этих пред-

ложений T* построим частотный словарь V(T*), содержащий абсолютную и относительную (частное от деления наблюдаемой, абсолютной частоты на количество слов в T*) частоты.

3. Сравниваем полученные относительные частоты в T и T*:

$$\textit{индекс значимости} = \frac{\textit{относительная частота в } T^*}{\textit{относительная частота в } T}$$

Если полученный индекс больше единицы, то словоформу с заданными характеристиками относим к семантическому полю слова C.

При обработке текстового файла возникает целый ряд сложностей, поэтому существуют определенные требования к исходным текстам. В большинстве случаев эти требования определяются нормами полиграфии и интуитивными соображениями:

– считается, что текст построен правильно (например, две запятые не могут идти подряд);

– дефис сразу после слова в конце строки говорит о переносе слова. После него в строке не могут идти никакие символы, в том числе пробел;

– тире после пробела является одиночным символом и должно отбиваться пробелом также справа;

– тире после буквенного символа при наличии последующего буквенного, расположенного на той же строке, говорит о словосочетании и никогда не отбивается пробелами.

Принятие таких правил облегчает обработку текстов, но усложняет их набор на компьютере, поскольку требует постоянного контроля вводимого материала.

Также для генерации всех словоформ слова и определение основы слова использовалось бинарное представление морфологического словаря, источником которого является грамматический словарь А. А. Зализняка [5] и словарь компании Диалинг, 162519 лемм [6]. Он был выбран по следующим причинам:

– при поиске в словаре используется конечный автомат, что позволяет находить слово за линейное от его длины время (очень быстро);

– удобный программный интерфейс;

– обладает развитой системой добавления новых слов;

– распространяется бесплатно под лицензией LGPL в исходных кодах.

2. РЕАЛИЗАЦИЯ

Схема обработки данных представлена на рис. 2.

Обработка текста начинается с определения параметров обработки (рис. 3).

В случае если задан режим интерактивного определения значимости слов, встретившихся впервые, пользователю предоставляется окно, где он может просмотреть все контексты употребления данного слова и решить является ли оно ключевым (рис. 4). После обработки текста существует возможность сохранить полученные списки ключевых слов. Количество файлов, отводящихся для сохранения результатов, ограничено. В случае если все файлы заняты, пользователю будет выдано сообщение о необходимости удалить некоторые из файлов. При этом удаление может быть осуществлено автоматически (будет удален файл с самой ранней датой обработки текста), а также в интерактивном режиме (удаляемый файл определяется пользователем).

Для просмотра результатов обработки текста, а также для осуществления его повторной обработки необходимо выбрать текст или его конкретную обработку в окне, содержащем список обработанных текстов. При работе с данными фильтра в окне отображается список текстов, включенных в фильтр, в противном случае — список текстов и их обработок, по которым происходило сохранение результатов обработки. Окно содержит наиболее важную информацию как о самом произведении (автор, название, длина в словоформах), так и о его конкретной обработке (способ, дата обработки, количество ключевых слов).

Окно просмотра результатов реализовано в виде записной книжки, состоящей из нескольких блоков. В верхней части окна содержится общая информация о тексте и текущей его обработке. Списки ключевых слов представляются в табличном виде с указанием частоты их встречаемости в тексте, определенных статистических характеристик, зависящих от способа обработки (относительной частоты, ранга частоты, амплитуды колебания ранга или количества текстов) и величины их отклонения от статистических значений. По желанию пользователя возможно отображение списков ключевых слов со следующими типами сортировок: по алфавиту, по частоте, по величине отклонения, по основанию статистики.

Рис. 2. Принцип обработки данных

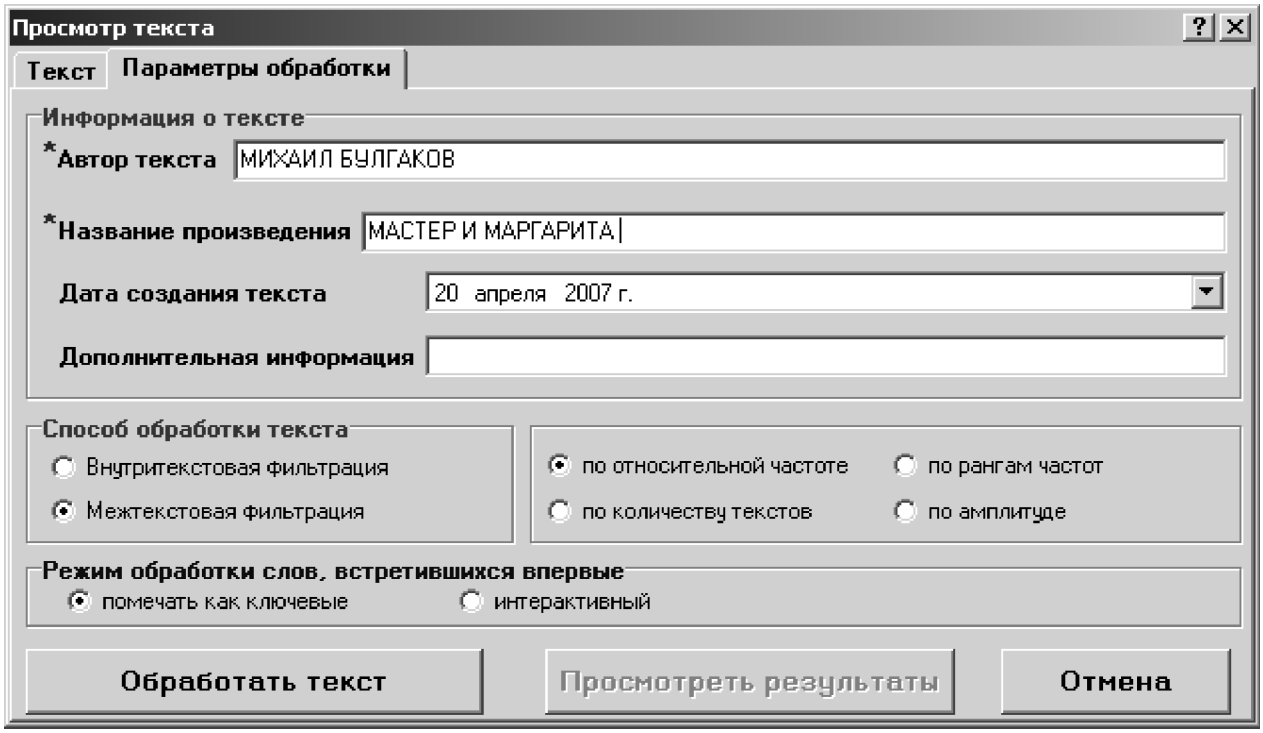


Рис. 3. Задание параметров обработки текста

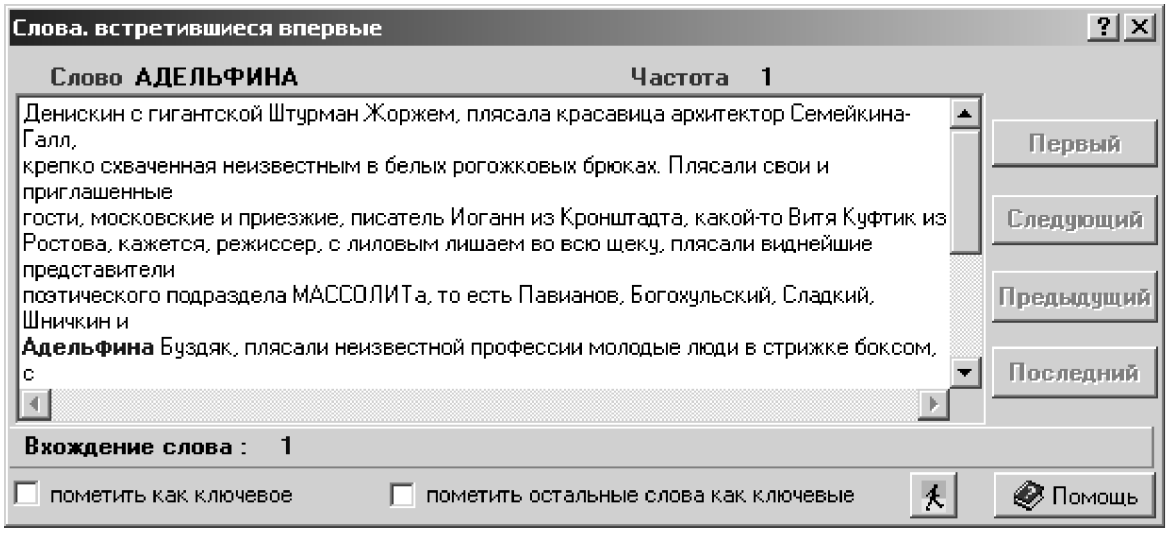


Рис. 4. Окно просмотра контекстов слов

Существует возможность построения гистограммы распределения значимости ключевых слов (рис. 5). Рядом с гистограммой приводится список слов, поскольку слова на графике не указываются явно, а нумеруются. Построение гистограммы возможно как для каждой длины слова отдельно, так и для всей совокупности ключевых слов.

При просмотре списка ключевых слов можно инициировать выявление семантического окружения выбранного слова.

Окно просмотра результатов состоит из 2-х блоков. В верхней части окна содержится общая информация о тексте и текущей его обработке. Списки контекстных слов представляются в табличном виде с указанием частоты их встречаемости в тексте, в подтексте и индекса значимости. Возможна настройка диапазонов частот и индекса значимости для отображения выборочного числа слов контекста (рис. 6).

По желанию пользователя возможно отображение списков контекстных слов со следующими типами сортировок: по алфавиту, по частоте в тексте, по индексу значимости, по частоте в подмножестве.

Предусмотрен режим просмотра ядер. На рис. 7 представлена диаграмма, отражающая зависимость между длиной слова и количеством ядер. Эта диаграмма поможет пользователю получить полную и наглядную информацию о том, слова с какой длиной наиболее часто встречаются в тексте, а с какой отсутствуют вовсе. Можно просмотреть группы ключевых слов в порядке зависимости от длины слова и номера ядра.

Также реализовано три способа просмотра ключевых слов по ядрам: метод слоев, метод пиков и метод расслоения пиков. Рассмотрим каждый из этих способов на примере.

Пусть есть список ключевых слов с длиной от одной до десяти букв. На рис. 8 показана диаграмма зависимости количества ядер от длины слова (по оси абсцисс откладывается длина слова, по оси ординат — номер ядра).

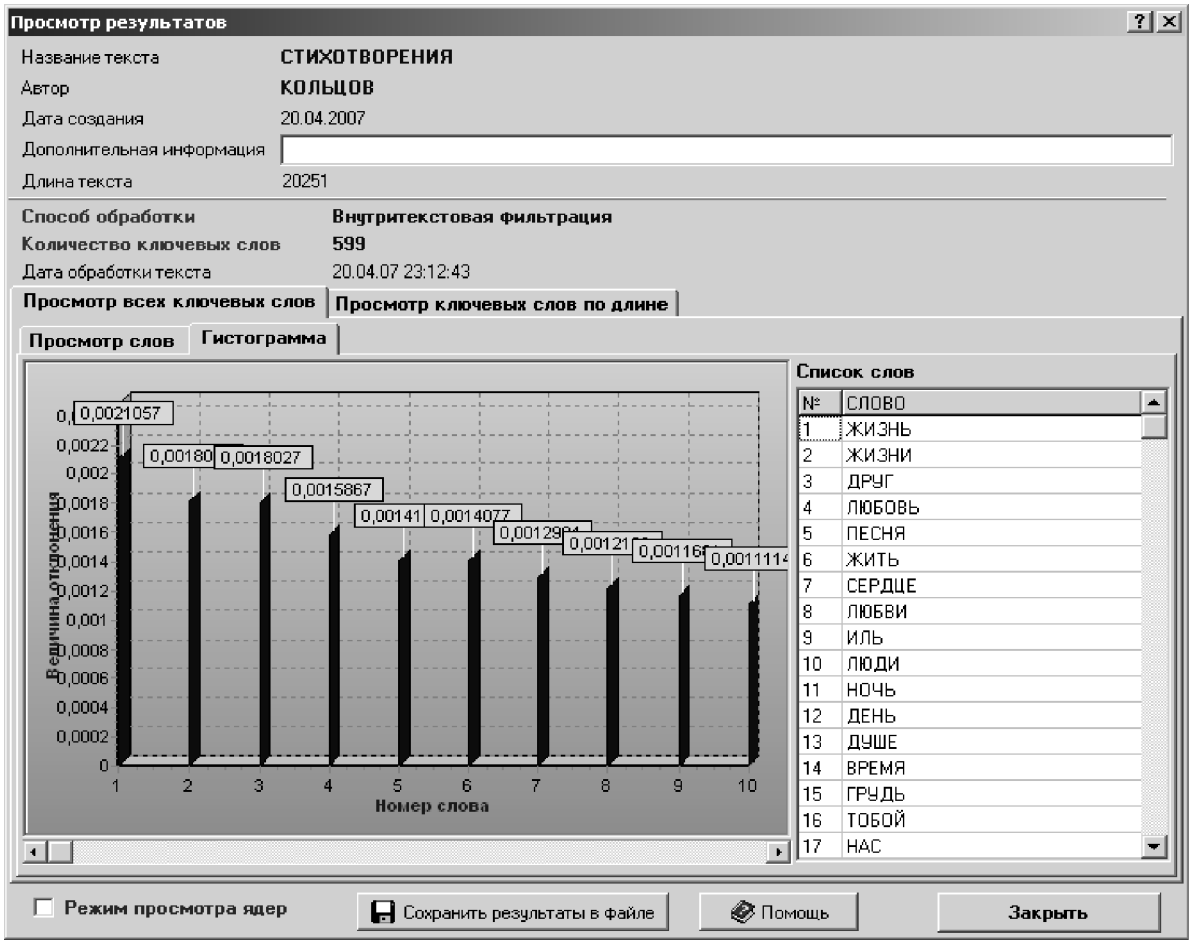


Рис. 5. Гистограмма распределения ключевых слов

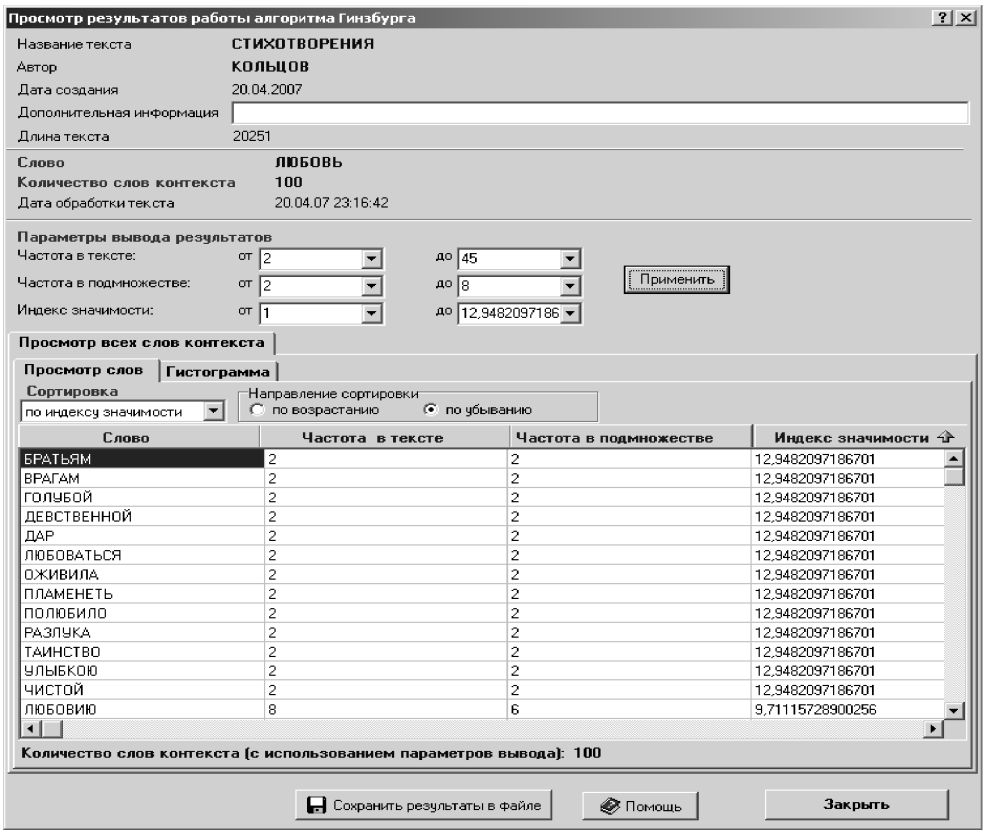


Рис. 6. Окно просмотра результатов работы алгоритма Е. Л. Гинзбурга

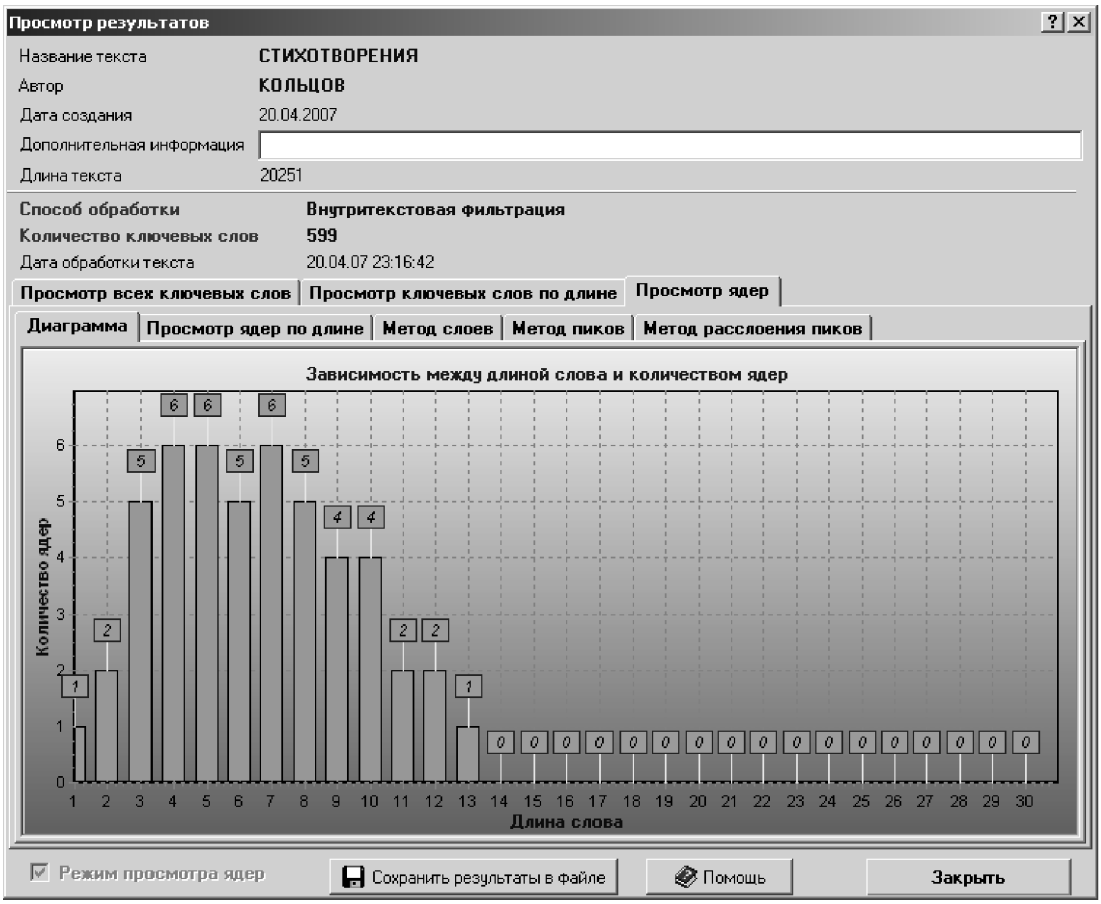


Рис. 7. Диаграмма зависимости между длиной слова и количеством ядер ключевых слов

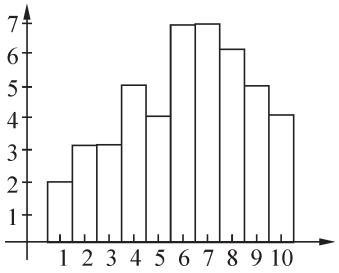


Рис. 8. Зависимость количества ядер от длины слова

При просмотре ключевых слов методом слоев выбирается номер ядра, после чего отобразятся те слова, максимальный номер ядра которых не ниже выбранного. На рис. 9 показаны слова, которые отобразятся при выборе номера ядра 2 и 5.

При просмотре ключевых слов методом пиков выбирается не номер ядра, а отклонение номера ядра от максимального номера в каждой длине. В список ключевых слов войдут слова, у которых отклонение номера ядра от максимального не меньше выбранного. На рис. 10 показаны слова, которые отобразятся при выборе отклонения номера ядра 0 и 3.

Третий способ просмотра ядер — метод расслоения пиков — несколько схож с методом пиков. Также выбирается отклонение номера ядра от максимального номера в каждой длине, при этом в список ключевых слов попадут слова, у которых отклонение номера ядра от максимального совпадает с выбранным. Для рассмотренного выше примера на рис. 11 показано отображение слов при выборе отклонения 1 и 3.

После повторной обработки текста существует возможность сравнения списков ключевых слов, полученных применением различных способов обработки (рис. 12).

После процедуры сравнения обработок пользователь имеет возможность просмотреть общий набор ключевых слов, полученных применением каждого из методов, а также списки слов, встречающихся только в одной из обработок. Предоставлены возможности сортировки указанных списков по алфавиту и частоте встречаемости в возрастающем и убывающем порядке.

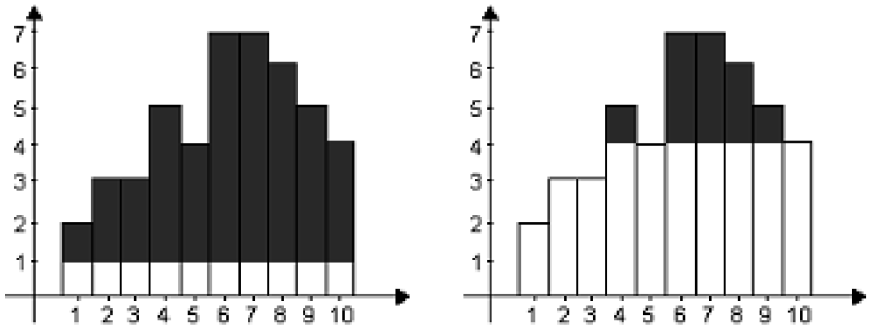


Рис. 9. Отображение слов по методу слоев: а) при выборе ядра 2; б) при выборе ядра 5

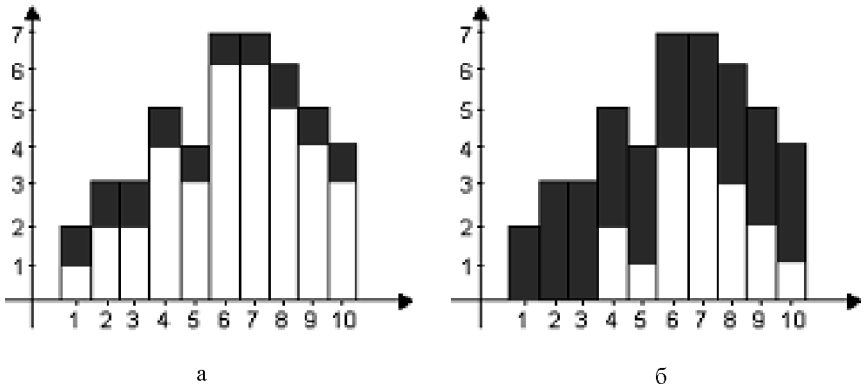
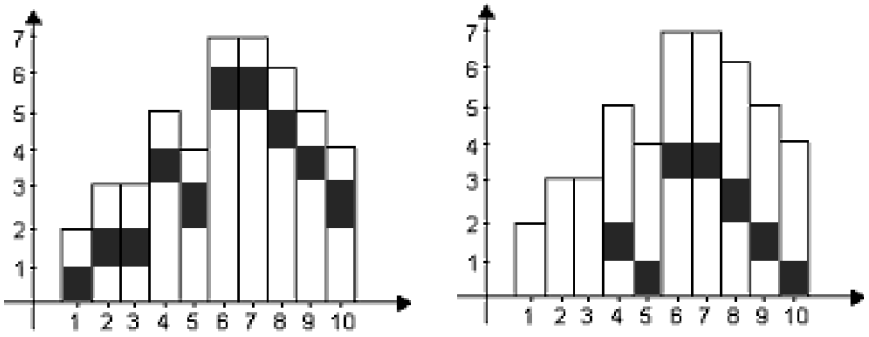


Рис. 10. Отображение слов по методу пиков: а) при выборе отклонения ядра 0; б) при выборе отклонения ядра 3



а б

Рис. 11. Отображение слов по методу расслоения пиков: а) при выборе отклонения ядра 1; б) при выборе отклонения ядра 3

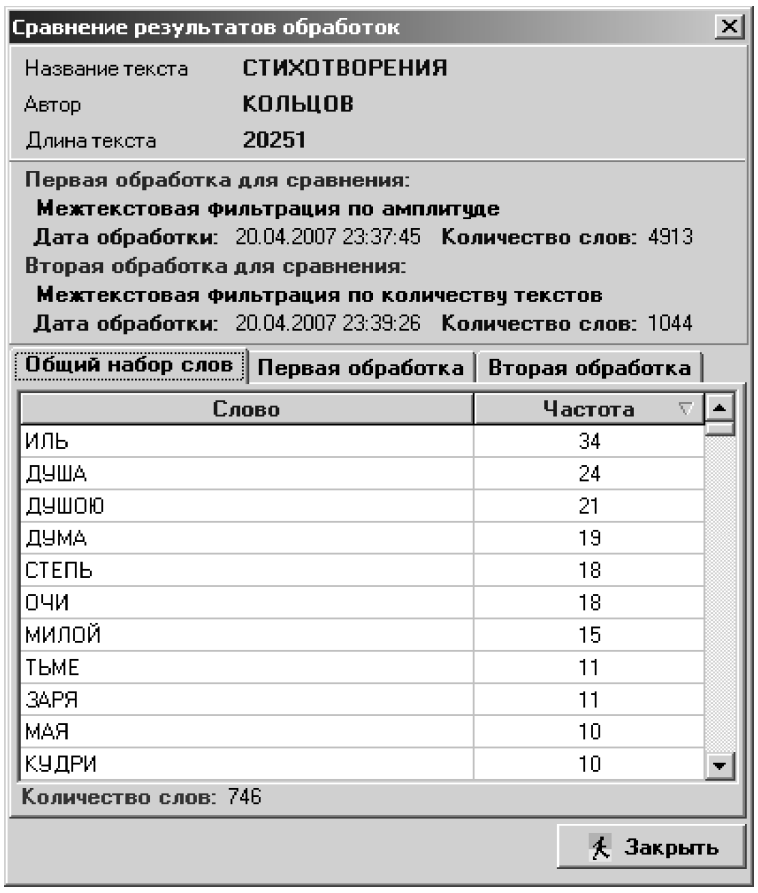


Рис. 12. Сравнение результатов обработок

3. СТРУКТУРА ПРОГРАММНОГО КОМПЛЕКСА

Программный комплекс состоит из ряда подсистем.

Подсистема обработки текста включает модули анализа текстового файла, построения частотного словаря текста и контекстов употребления слов, встретившихся впервые, автоматизации методов фильтрации слов, а также модули статистической обработки данных и повторной обработки текстов.

Подсистема работы с файлами результатов содержит модули, отвечающие за сохранение результатов обработки текстов, удаление файлов с результатами, а также модули, реализующие диалог с пользователем при сохранении результатов обработки и удалении файлов (в интерактивном режиме).

Подсистема работы со списками ключевых слов отвечает за реализацию функций просмотра списков слов с различными видами сортировок, расчета ядер ключевых слов с возможностью просмотра их по методу слоев, пиков, расслоения пиков, построения гистограмм распределения значимости слов и диаграммы зависимости количества ядер от длины слова.

Подсистема работы со списками слов контекста отвечает за реализацию функций просмотра списков слов с различными видами сортировок, построения гистограмм распределения значимости слов.

Подсистема сравнения обработок текста представлена модулями, реализующими средства выбора пользователем текста и его обработок

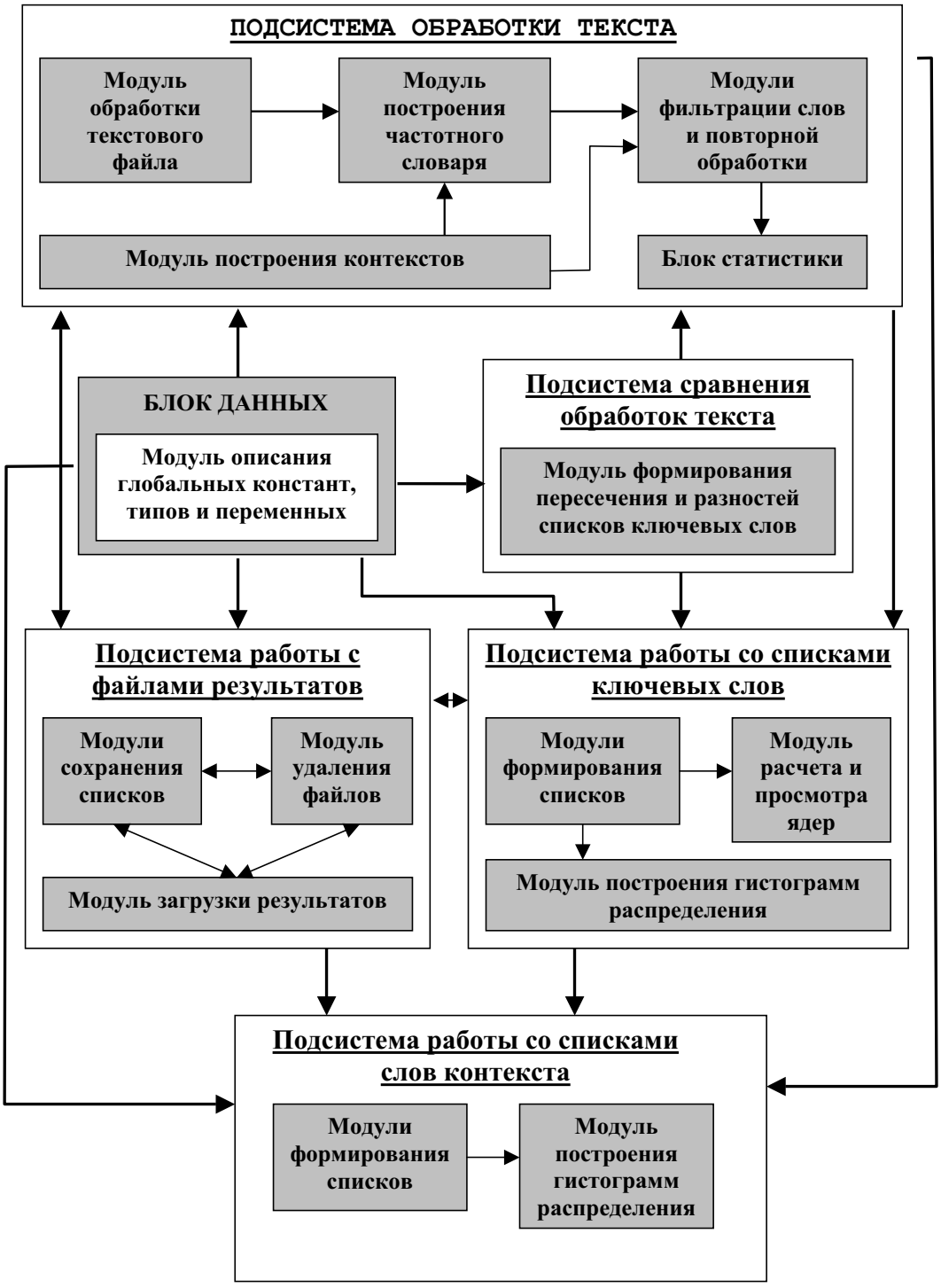


Рис. 13. Взаимодействие функциональных блоков

для сравнения, а также процедуры загрузки данных для сравнения и непосредственно формирования списков ключевых слов. Модули также реализуют возможности удобного представления результатов сравнения и отображения списков с различными видами сортировок.

Вспомогательная подсистема состоит из модулей, реализующих интерфейс программы, реакцию на действия пользователя.

Структура подсистем программного комплекса и взаимосвязь между ними отражена на рис. 13.

ЗАКЛЮЧЕНИЕ

В процессе разработки программного комплекса было сделано следующее:

– реализованы алгоритмы межтекстовой и внутритекстовой фильтрации слов исходного текста;

– разработаны процедуры разбиения списков ключевых слов на ядра, что позволяет существенно сузить круг специфичных слов для проведения исследований;

– проведена статистическая обработка набора текстов с целью формирования фильтра, на основе которого происходит определение ключевых слов произвольного текста;

– предоставлены возможности сохранения результатов обработки текста и работы с полученными списками ключевых слов (построение гистограмм распределения значимости слов);

– реализованы возможности повторной обработки текстов, а также сравнения результатов, полученных применением различных способов выделения ключевых слов;

– реализован метод построения семантического поля для заданного ключевого слова;

– предоставлены возможности работы со списками слов полученного семантического поля (построение гистограмм распределения значимости слов);

– разработан фильтр, основной функцией которого является выделение слов, семантически близких к заданному слову;

– разработан фильтр, основной функцией которого является отсеивание слов, составляющих специфику естественного языка в целом.

Программа может применяться при проведении широкого круга научно-исследовательских работ, позволяет повысить эффективность их проведения за счет использования средств вычислительной техники, устранения наиболее трудоемкой части исследовательской деятельности по анализу текста.

Расширение программного комплекса предполагает:

– добавление блока семантического наполнения дискурсивных слов. Этот модуль должен подключаться перед построением частотного словаря исходного текста. В результате работы блока получится переработанный текст, в котором каждое местоимение исходного текста заменено соответствующим ему именем собственным или нарицательным. Такая процедура значительно повысит частоту встречаемости отдельных единиц текста за счет исключения дискурсивных слов;

– проведение лемматизации слов исходного текста на этапе построения частотного словаря, т.е. приведение словоформ текста к начальной форме, при этом частота встречаемости соответствующих слов суммируется. Объединение слов позволит, во-первых, значительно сократить количество ключевых слов, а во-вторых, повысить качество получаемых результатов.

СПИСОК ЛИТЕРАТУРЫ

1. *Воронина И.Е.* Компьютерное моделирование лингвистических объектов: монография / И. Е. Воронина. — Воронеж: Издательско- полиграфический центр ВГУ, 2007. — 177 с.

2. *Воронина И.Е.* Метод последовательной фильтрации при разработке лингвистического обеспечения информационных процессов / И. Е. Воронина, А. А. Кретов // Межвуз. сб. научных трудов “Математическое обеспечение ЭВМ”, Вып. 1, Воронеж, 1999. — С. 17—21.

3. *Гинзбург Е.Л.* Идиоглоссы: проблемы выявления и изучения контекста / Е. Л. Гинзбург // Семантика языковых единиц: Доклады VI Международной конференции. Т. I, М., 1998. —С. 26—28.

4. *Титова О.С.* Программа для выявления контекста (алгоритм Е. Л. Гинзбурга) / О. С. Титова, А. А. Кретов., И. Е. Воронина // Информатика: проблемы, методология, технологии : материалы 7-й регион. науч. — метод. конференции, Воронеж, 8—9 февр. 2007 г. — Воронеж. : Изд-во ВГУ, 2007 . — С. 420—423.

5. *Крештель Е.В.* Алгоритмы выделения ключевых слов в текстах на естественных языках / Е. В. Крештель, А. А. Кретов., И. Е. Воронина // Информатика: проблемы, методология, технологии : матер. 3-й регион. науч.-метод. конфер., Воронеж, 2001 г. — Воронеж. : Изд-во ВГУ, 2001 . — С. 35.

6. *Зализняк А.А.* Грамматический словарь русского языка. Словоизменение / А. А. Зализняк. — М.: Русский язык, 1987. — 3-е изд., испр . —878 с.

Воронина Ирина Евгеньевна — к.т.н., доцент кафедры ПОиАИС, факультет прикладной математики, информатики и механики, Воронежский государственный университет. Тел. 208-337. E-mail: voronina@amm.vsu.ru

Кретов Алексей Александрович — д. фил. н., профессор, зав. каф. структурной и прикладной лингвистики факультета РГФ, Воронежский государственный университет. Тел. 204-149. E-mail: a_a_kretov@rambler.ru

Титова Олеся Сергеевна — инженер-программист ЗАО НПП РЕЛЭКС, E-mail: oleti85@yandex.ru

Voronina Irina Ye. — Candidate of Technical Sciences, Associate Professor, the dept. of the Software and Information System Administration, Voronezh State University. Tel. 208-337. E-mail: voronina@amm.vsu.ru.

Kretov Aleksey Aleksandrovich — Doctor of Philology Sciences, Professor, the Head of the dept. of Theoretical and Applied Lingvistics , Voronezh State University. Tel. 204-149. E-mail: a_a_kretov@rambler.ru

Titova Olesy Sergeevna — ingeneer "ЗАО НПП РЕЛЭКС", E-mail: oleti85@yandex.ru.