Поликарпов А.А., Поддубный В.В., Кукушкина О.В., Кубарев А.И.,
Варламов А.А., Суровцева Е.В., Пирятинская Е.Ф.

Комплексная тексто-аналитическая система "СтилеАнализатор-2", основанная на Web-технологиях: разработка, наполнение данными и тестирование на прикладных задачах[1]

СОДЕРЖАНИЕ

[1] Данная система разрабатывалась в 2011-2013 гг. в МГУ и ТГУ на основе гранта РФФИ № 11-07-00776-а («Комплексная тексто-аналитическая система "СтилеАнализатор-2", основанная на Web-технологиях: разработка, наполнение данными и тестирование на прикладных задачах»).

2.2.3. Исследование потокового классификатора, основанного на С-мере
2.2.4. Арбитражный метод классификации на основе С-меры
Литература к разделу 2.2. ПОТОКОВАЯ КЛАССИФИКАЦИЯ ТЕКСТОВ НА ОСНОВЕ С-МЕРЫ

3. НОВЫЕ ИССЛЕДОВАТЕЛЬСКИЕ РЕЗУЛЬТАТЫ, ПОЛУЧЕННЫЕ В 2013 ГОДУ С ПОМОЩЬЮ ТЕКСТО-АНАЛИТИЧЕСКОЙ СИСТЕМЫ «СТИЛЕАНАЛИЗАТОР 2.0. WEB» НА ОСНОВЕ ПРИМЕНЕНИЯ НЕИСПОЛЬЗОВАВШИХСЯ РАНЕЕ ПРИЗНАКОВ

3.1. Анализ сочетания алгоритмов статистического и логического анализа при классификации текстов.
3.2. Оптимизация признакового пространства текстов.
3.3. Методика просеивания выборки на саму себя.
3.4. Исследование влияния диалогов на авторский стиль.
3. 5. Развитие экспериментов при использовании системы «Стилеанализатора2-web» по проверке новых, экспериментально ещё не исследованных признаков.
3.5.1. Использование в качестве признаков результатов морфемного членения и словообразовательного анализа.
3.5.2. Использование наборов лексических единиц для психо-лингвистического анализа особенностей личности автора.
3.5.3. Среднее число слогов в слове и средняя длина предложений.
3.5.4. Использование тезаурусных групп для распознавания индивидуального стиля.

==========

АННОТАЦИЯ

В настоящей статье представлены и кратко описаны основные методы и классы методов, реализующие программный инструментарий системы комплексной обработки текстов «СтилеАнализатор 2.0. WEB», а также некоторые важнейшие результаты, полученные в ходе её использования. Данная система разрабатывалась в 2011-2013 гг. на основе гранта РФФИ № 11-07-00776-а в ходе совместной работы творческого коллектива Лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ и творческого коллектива кафедры информатики факультета информатики Томского государственного университета. Руководитель проекта **— профессор МГУ** Поликарпов Анатолий Анатольевич. Соруководитель проекта – профессор Томского университета Поддубный Василий Васильевич.

ВВЕДЕНИЕ

Важнейшие результаты, полученные в ходе реализации данного проекта сгруппированы по трём основным тематическим разделам:

1. ОСНОВНЫЕ ПРИНЦИПЫ ОРГАНИЗАЦИИ СИСТЕМЫ «СТИЛЕАНАЛИЗАТОР 2.0. WEB».

Здесь представлены и кратко описаны все основные методы и классы методов, реализующие программный инструментарий комплексной обработки текстов «СтилеАнализатор 2.0. WEB». Приведено описание серверной и клиентской части, хранения текстов, преобразования текстов в таблицы и анализ таблиц различными методами.

Взаимодействие пользователя с корпусом текстов предоставляет возможность гибкой работы с корпусами текстов, позволяет автоматически и полуавтоматически выделять в текстах композиционные элементы, такие как заголовки, оглавления, предисловия, послесловия, диалоги, авторская речь и т.п.

Блок анализа данных системы «СтилеАнализатор 2.0 WEB» включает в себя:

− статистический анализ текстов (подсчёт значений признаков, факторный анализ, в том числе метод главных компонент, дискриминантный анализ, кластерный анализ, методы байесовской классификации),
− информационный анализ и классификацию (деревья решений),
− логический анализ и тестовое распознавание,
− нейронные сети прямого распространения,
− самоорганизующиеся карты Кохонена (Self-Organizing Maps − SOM-сети),
− классификацию на основе суффиксных деревьев и др.
– построение гистограмм распределений признаков с наложением кривой гауссова распределения для визуального анализа

2. РАЗВИТИЕ ИССЛЕДОВАТЕЛЬСКИХ ФУНКЦИЙ СИСТЕМЫ «СТИЛЕАНАЛИЗАТОР 2.0. WEB». Здесь представлен новый статистический и логический аппарат, используемый в «СтилеАнализаторе 2.0. WEB» для классификации текстов русской классической литературы 19 века на основе байесовской классификации с обучением на основе использования копула-функций, а также на основе потоковой классификации текстов на основе С-меры.

3. НОВЫЕ ИССЛЕДОВАТЕЛЬСКИЕ РЕЗУЛЬТАТЫ, ПОЛУЧЕННЫЕ С ПОМОЩЬЮ ТЕКСТО-АНАЛИТИЧЕСКОЙ СИСТЕМЫ «СТИЛЕАНАЛИЗАТОР 2.0. WEB» В 2013 ГОДУ.

Здесь представлены результаты анализа сочетания статистических и логических алгоритмов классификации текстов, рассмотрены возможности оптимизации признакового пространства текстов, представлены эксперименты по использованию системы «Стилеанализатор2-WEB» для проверки новых, экспериментально ещё мало исследованных признаков текстов.

1. ОСНОВНЫЕ ПРИНЦИПЫ ОРГАНИЗАЦИИ СИСТЕМЫ «СТИЛЕАНАЛИЗАТОР 2.0. WEB»

1.1.Технологии используемые в «СтилеАнализатор 2.0 WEB» для развертывания приложения и авторизации пользователей.

Тексто-аналитическая система «СтилеАнализатор 2.0 WEB» является web-приложением, написанным на языке программирования C# с использованием технологий ASP.NET и ADO.NET, которые включены в состав платформы Microsoft.NET (распространяется Microsoft бесплатно).

Для развертывания приложения используются веб-сервер Microsoft IIS (Internet Information Services, распространяется совместно с операционными системами семейства Microsoft NT) и СУБД MySQL (распространяется свободно по General Public License).

Пользователь системы «СтилеАнализатор-2» при работе с системой связывается с ней, используя интернет (по определённому адресу через определённый порт), вводит свои логин и пароль доступа. Начальной страницей приложения является страница авторизации Login.aspx.

При обращении пользователя на сервере запускается отдельный экземпляр приложения, использующий основные настройки из файла SAWEB.ini, расположенном в папке с приложением. Данный файл содержит настройки авторизации пользователей, путь к базе данных (БД) приложения и путь к каталогу, хранящему файлы, созданные пользователями приложения. (Рис. 1.1)

```
[authorization]
;type of authorization (asp or win)
type = win
```

[database]
;path to application data base
datasourse = localhost
database = saweb_data

[datapath]
;path to user files
path = d:\user_data

Рис. 1.1 – Пример файла SAWEB.ini

Описанный подход позволил перенести функционал системы «СтилеАнализатор» на новую программную основу, без повторной реализации существующих алгоритмов.
В зависимости от типа авторизации, заданного в файле SAWEB.ini, могут использоваться учетные записи пользователей Windows или учетные записи пользователей, зарегистрировавшихся в системе (Рис. 1.2).

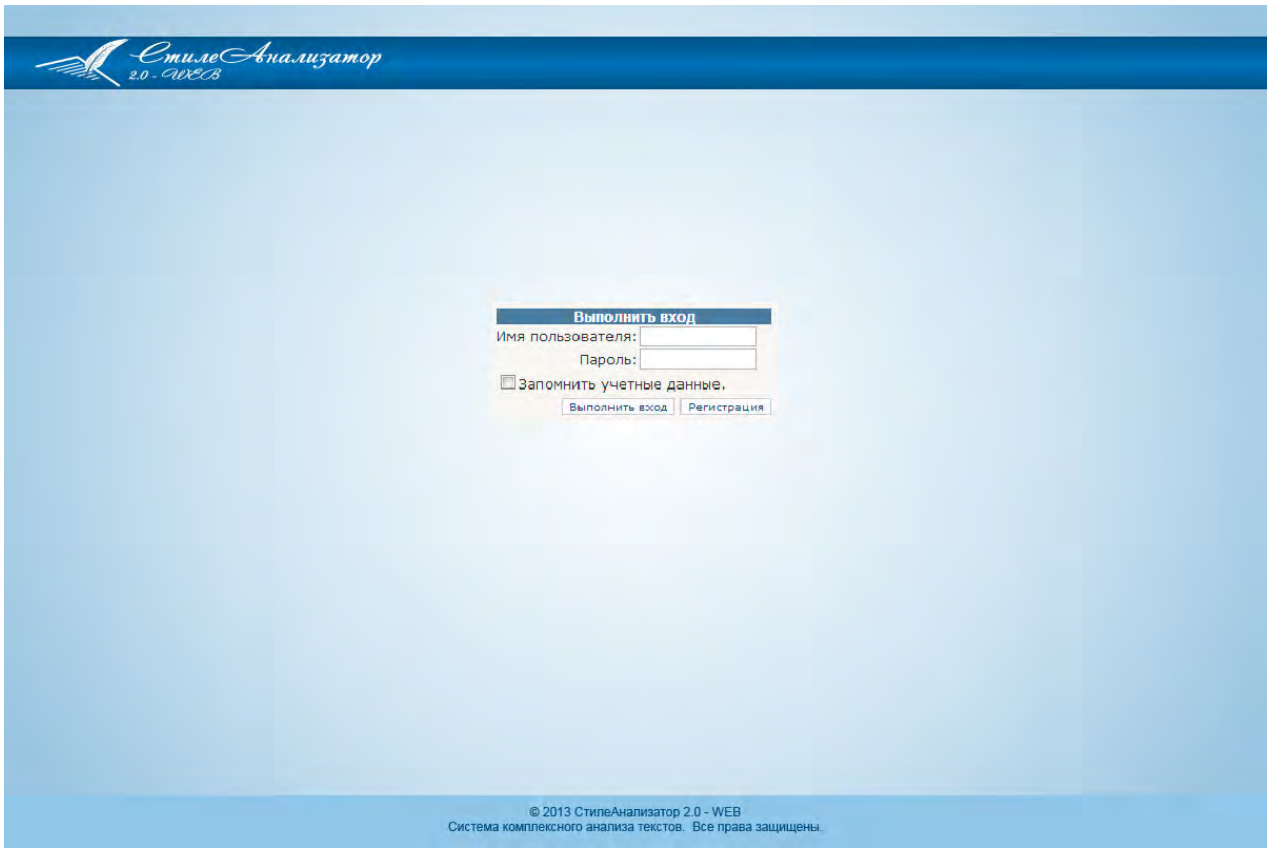


Рис. 1.2 – Страница авторизации пользователя (в случае авторизации учетными записями Windows кнопка «Регистрация» отсутствует).

В первом случае добавление пользователей производится администратором сервера приложения путём созданием учетной записи на сервере системы или в домене, в который входит сервер (учетная запись может обладать любыми правами). Во втором случае пользователи проходят самостоятельную регистрацию в системе, заполнив форму регистрации (Рис. 1.3), переход на которую осуществляется нажатием на кнопку «Регистрация» (Рис. 1.4).

Для каждого пользователя, зарегистрированного в системе и выполнившего первый вход, на сервере создается директория {Path}\{UserName}\ (параметр Path находится в SAWEB.ini), хранящая в себе результаты различных методов анализа данных, полученные пользователем, такие как таблицы и диаграммы, а также временные файлы, которые создает приложение во время работы. Информация о всех данных пользователя, хранящихся на сервере, содержится в БД (Рис. 1.3 – 1.5).

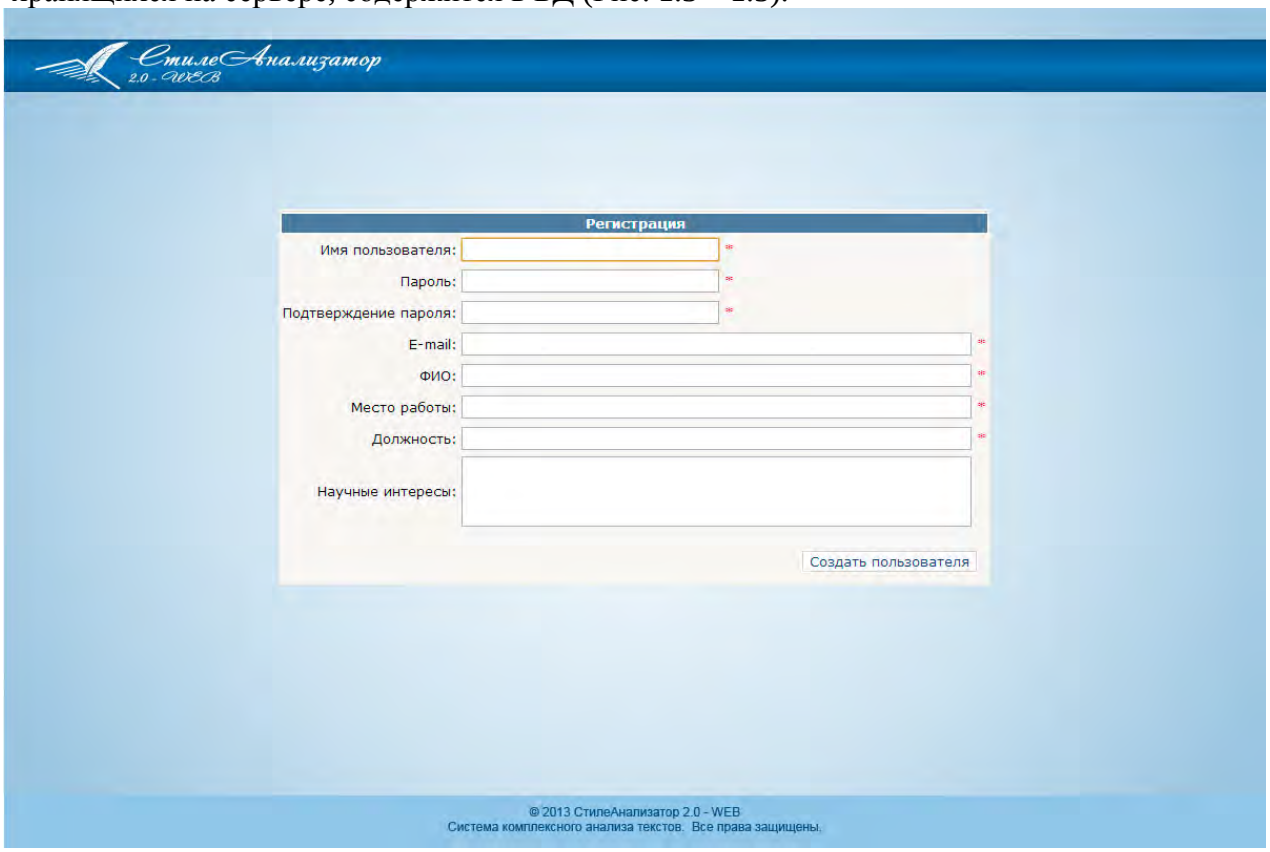


Рис. 1.3 – Страница регистрации нового пользователя.

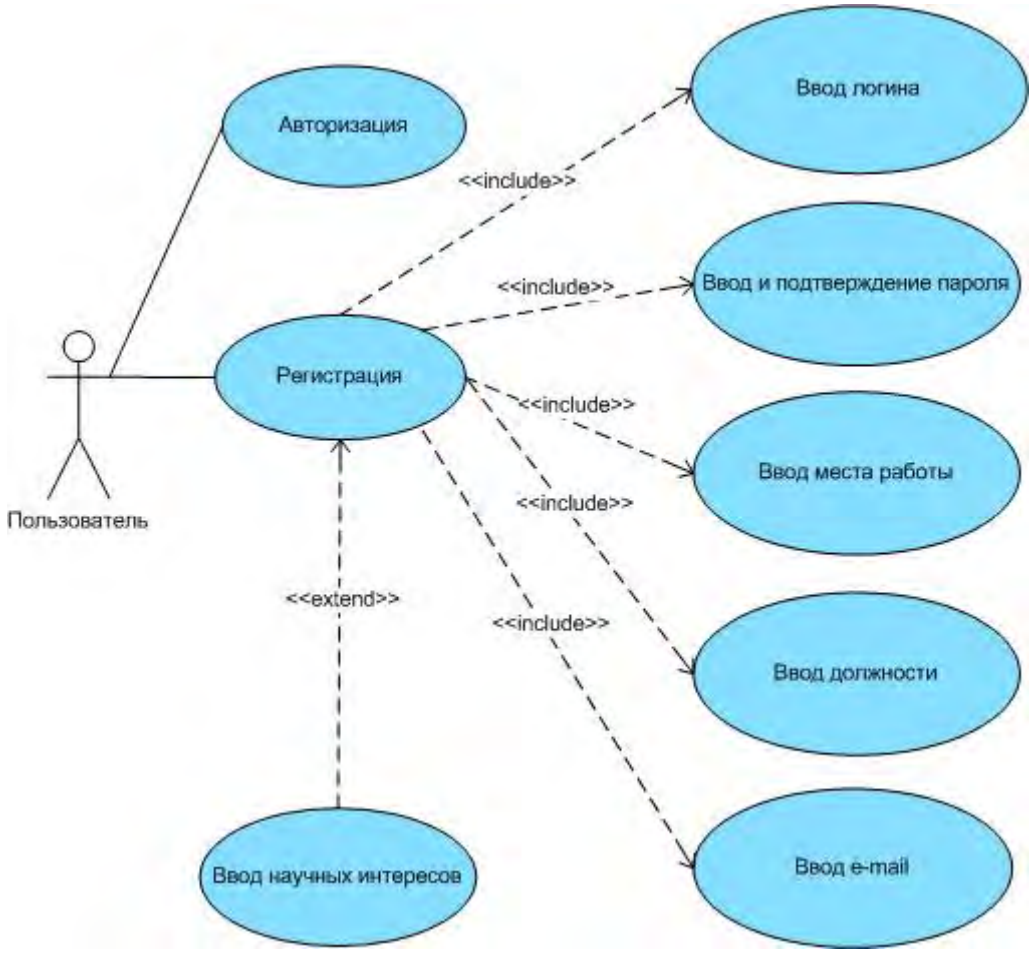


Рис. 1.4 – Схема взаимодействия пользователя с механизмом авторизации системы.

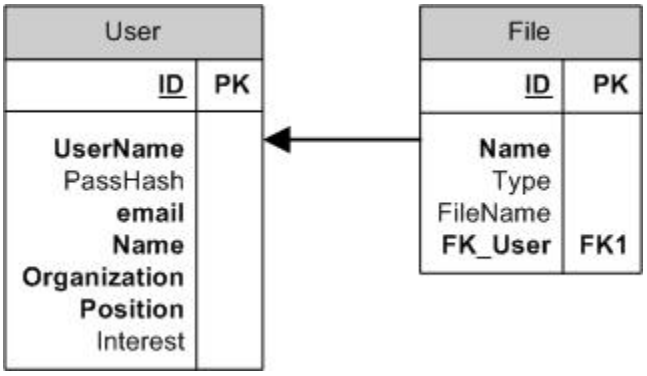


Рис. 1.5 – Схема хранения информации о файлах пользователя.

1.2. Взаимодействие пользователя с корпусом текстов.

Из системы «СтилеАнализатор-2» пользователь может осуществлять взаимодействие с корпусом текстов, хранящимся в БД. В соответствии с предоставленными ему правами пользователь может добавлять тексты в БД, объединять их в нужные ему для исследования группы (множества), редактировать их описание, разметку и т.д. (Рис. 2.1)

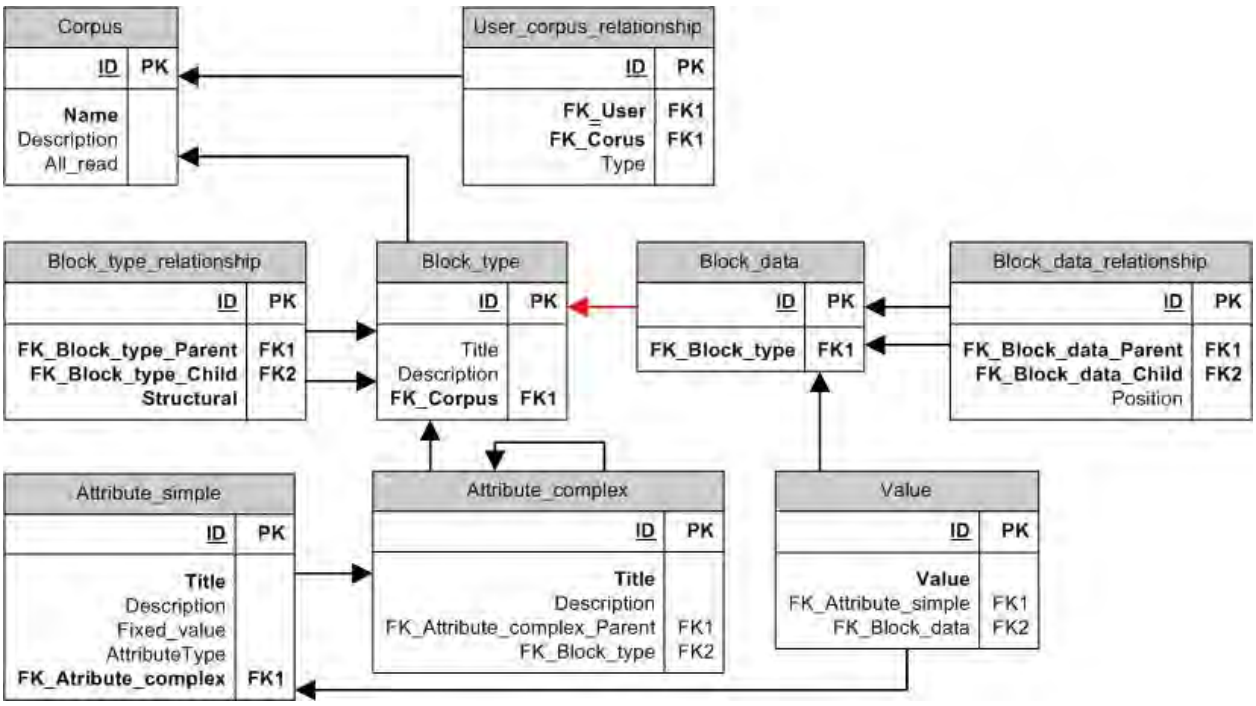


Рис. 2.1 – ER-диаграмма БД, отвечающая за хранение корпусов.

Все данные хранятся в одной сущности *«Values»*. Логическая организация этих данных осуществляется при помощи набора сущностей *«Atribute_Simple»*, *«Atribute_Complex»*, *«Block_type»* и таблиц организующих их связи. Сущность *«Block_Data»* организует связи между данными и логикой их хранения. За основу данной схемы были взяты принципы Универсальной модели данных.

В рамках системы «СтилеАнализатор 2.0 WEB» существует пять типов пользователей:

1. Владелец корпуса;
2. Администратор корпуса;
3. Модератор корпуса;
4. Участник корпуса;
5. Пользователь (сторонний);

Владелец корпуса – пользователь, создавший свой корпус в рамках системы (Рис 2.2).

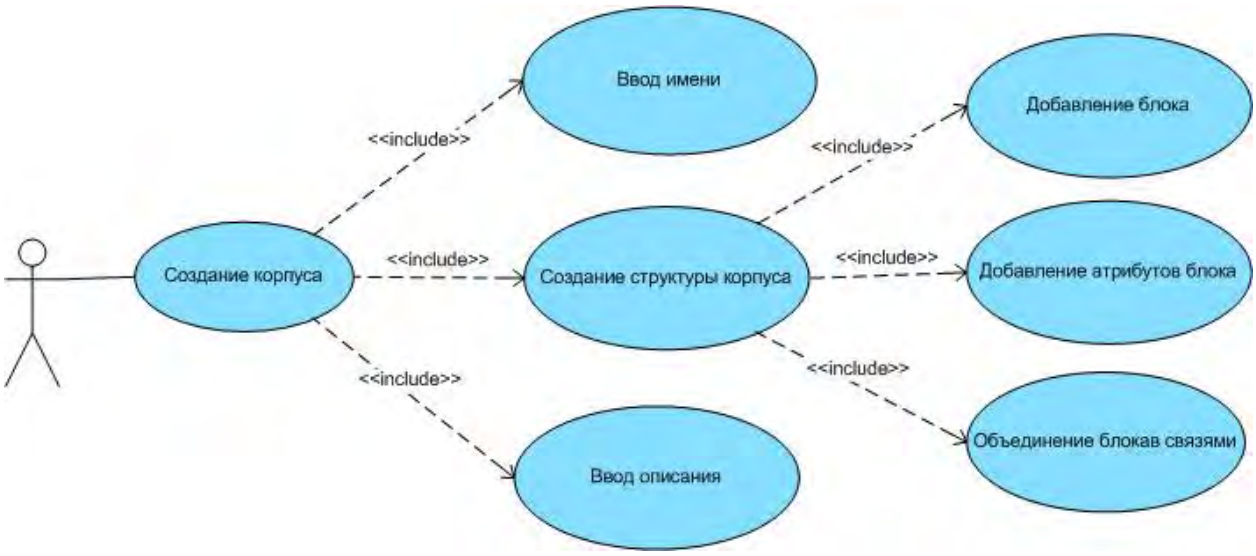


Рис. 2.2 – Схема взаимодействия пользователя с механизмом создания корпуса.

Структура корпуса в рамках системы «СтилеАнализатор 2.0 WEB» представляет собой набор блоков данных разных типов, объединенных иерархической системой связей. Каждый корпус должен состоять из двух основных типов блоков: единица корпуса (максимальная логическая единица корпуса, например, текст) и минимальная структурная единица корпуса (например, буква). Между данными блоками могут находиться промежуточные типы блоков, отвечающие за логическое представление текста. Таким образом, корпус есть набор блоков высшего уровня, каждый блок одного уровня состоит из блоков уровня ниже, спускаясь до низшего блока.

В «СтилеАнализаторе 2.0 WEB» существует базовый набор типов блоков, состоящий из: текста, предложения, слова, морфемы, буквы. Также пользователь может создавать свои типы блоков.

Данный механизм позволяет выделять в тексте композиционные элементы, такие как:

- заголовки;
- оглавления;
- предисловия;
- послесловия;
- диалоги;
- авторская речь;
- списки различных вставных объектов;
- библиографические списки литературы;
- таблицы;
- схемы;

– эпиграфы;
– цитаты из других текстов;
– редакционная информация о технических характеристиках изданного текста.

Блок «авторская речь», присутствует по умолчанию, и если не был введен создателем корпуса явно, то авторской речью считается все части текста, не относящиеся к другим блокам.

Каждый блок описывается набором атрибутов. Понятие атрибута разделено на «Комплексный атрибут» и «Простой атрибут». Комплексный атрибут является группой простых атрибутов, объединенных по какому либо признаку. Между комплексными атрибутами установлено отношение типа «Родитель-ребенок».

Связи между блоками делятся на два типа: структурные и не структурные. Структурная связь - эта связь, при которой все потомки представляют разбиение предка, например, объединение всех слов предложения восстанавливает исходное предложение.

По умолчанию, система содержит два структурных блока: «слово» и «текст», что позволяет на начальном этапе не формировать структуру корпуса, а заняться этим на готовом корпусе.

Для этого необходимо перейти на страницу «Мои корпуса» (Рис. 2.3).

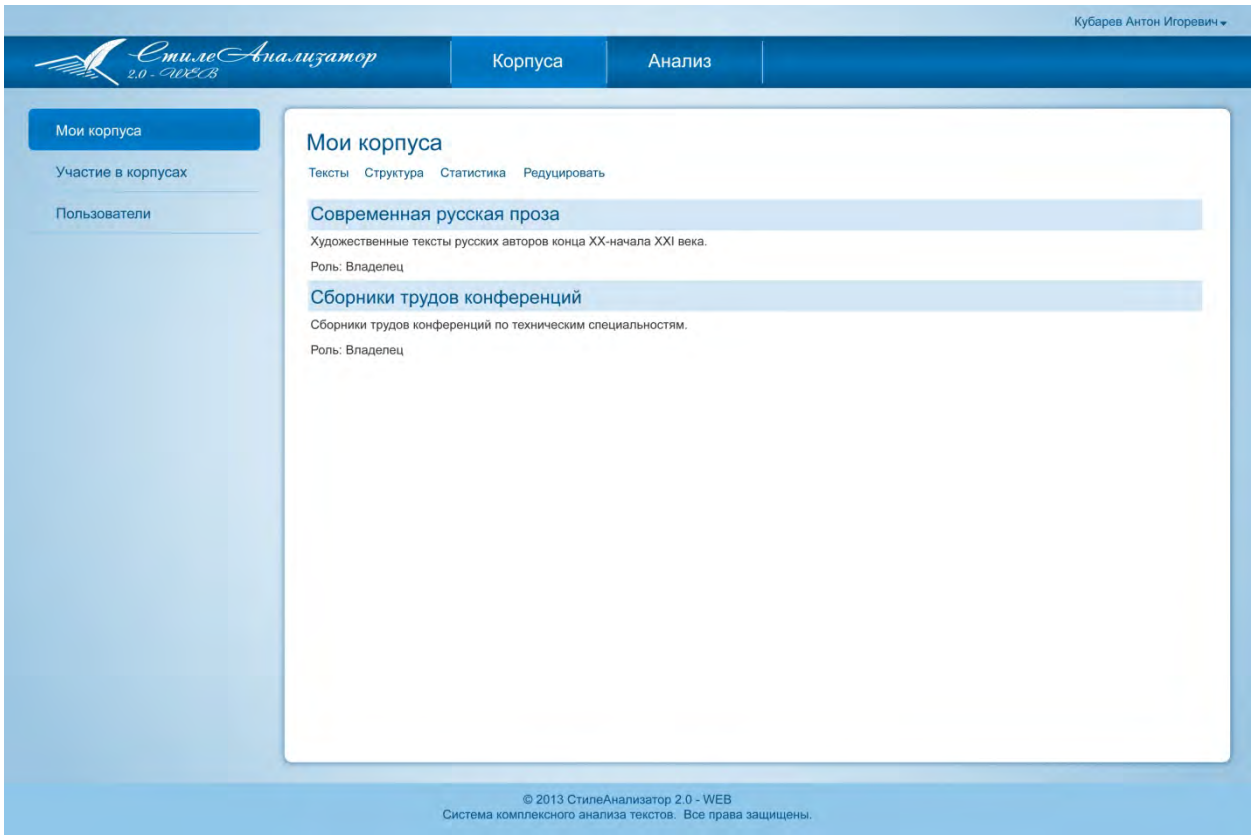


Рис 2.3 – Страница отображающая список корпусов пользователя

Выбрать из списка необходимый корпус и перейти на страницу «Структура». Интерфейс системы для удобства пользователей позволяет управлять структурой корпуса в графической нотации (Рис. 2.4).

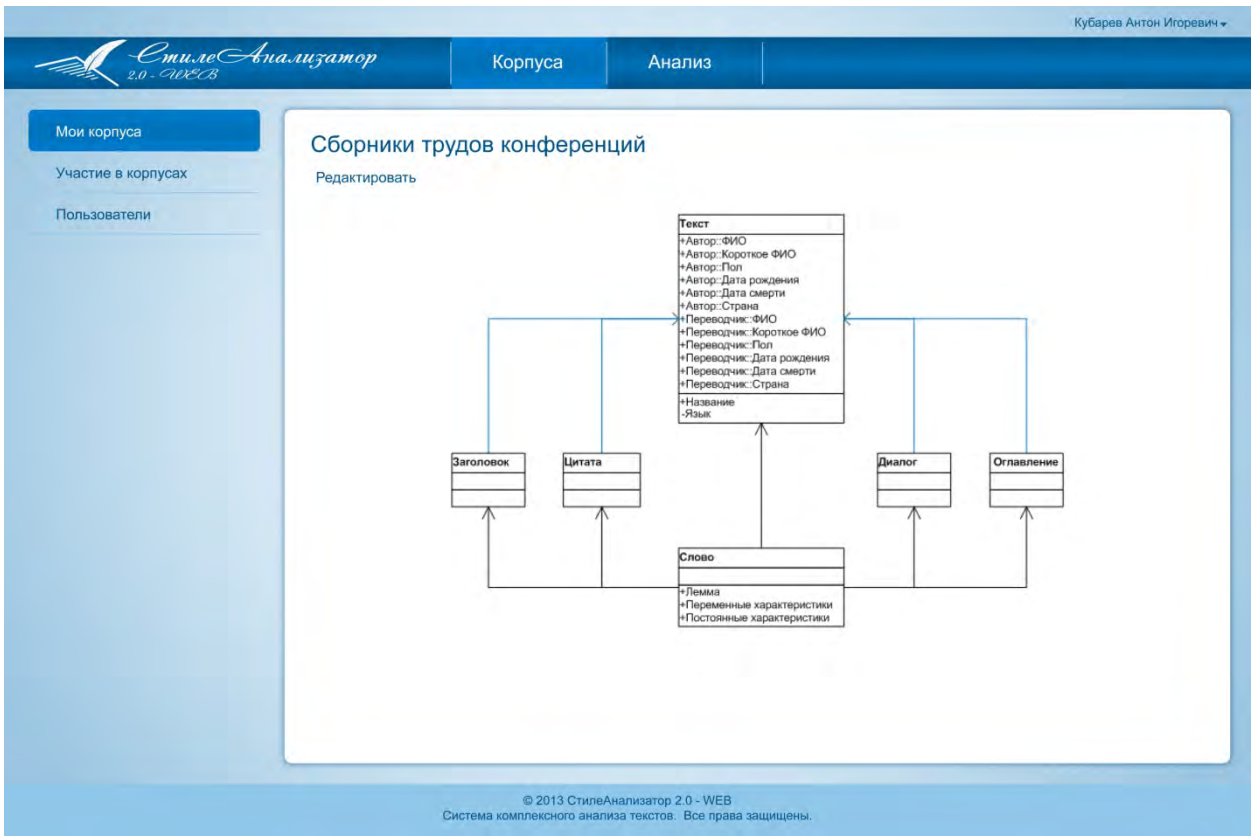


Рис.2.4 – Страница просмотра структуры корпуса.

В данной нотации типы блоков изображаются в виде прямоугольников с тремя разделами: название типа блока, комплексные атрибуты и простые атрибуты. Знаки плюс и минус возле атрибутов означают наличие возможности задавать пользователю произвольные значения атрибутов или выбирать из заранее заданного списка значений соответственно.

Структурные связи обозначаются сплошной стрелкой серого цвета, неструктурные - синего.

Описанный функционал представляет инструментарий для формирования корпусов текстов с произвольной паспортизацией и структурой.

Для выделения блоков того или иного типа пользователю необходимо перейти на страницу «Мои корпуса», выделить интересующий его корпус и перейти на страницу «Тексты», где выбрать пункт меню «Добавить» или «Редактировать» (предварительно выбрав текст для редактирования из таблицы). В открывшейся странице «Редактор текста» перейти на страницу «Редактор блоков текста» (Рис 2.5).

Данный редактор позволяет отображать блоки текста следующим образом. Пользователь из выпадающего списка «Тип блока» выбирает один из типов. После чего ниже появляется таблица, в которой отображаются параметры начала и конца блока. Выделив строку таблицы кликом, получим выделение блока в тексте. Пользователь, изменив выделение, может сохранить новые параметры начала и конца блока в тексте, воспользовавшись соответствующей ссылкой. Или удалить блок, отметив его галочкой и воспользовавшись ссылкой «Удалить». Также существует возможность автоматической разметки блоков, например, например, блоков, содержащих диалоги в художественных текстах («Выделить автоматически» + «Диалог»). Тип блоков будет изменен на диалог, в таблицу блоков будут добавлены все диалоги, найденные системой в тексте.

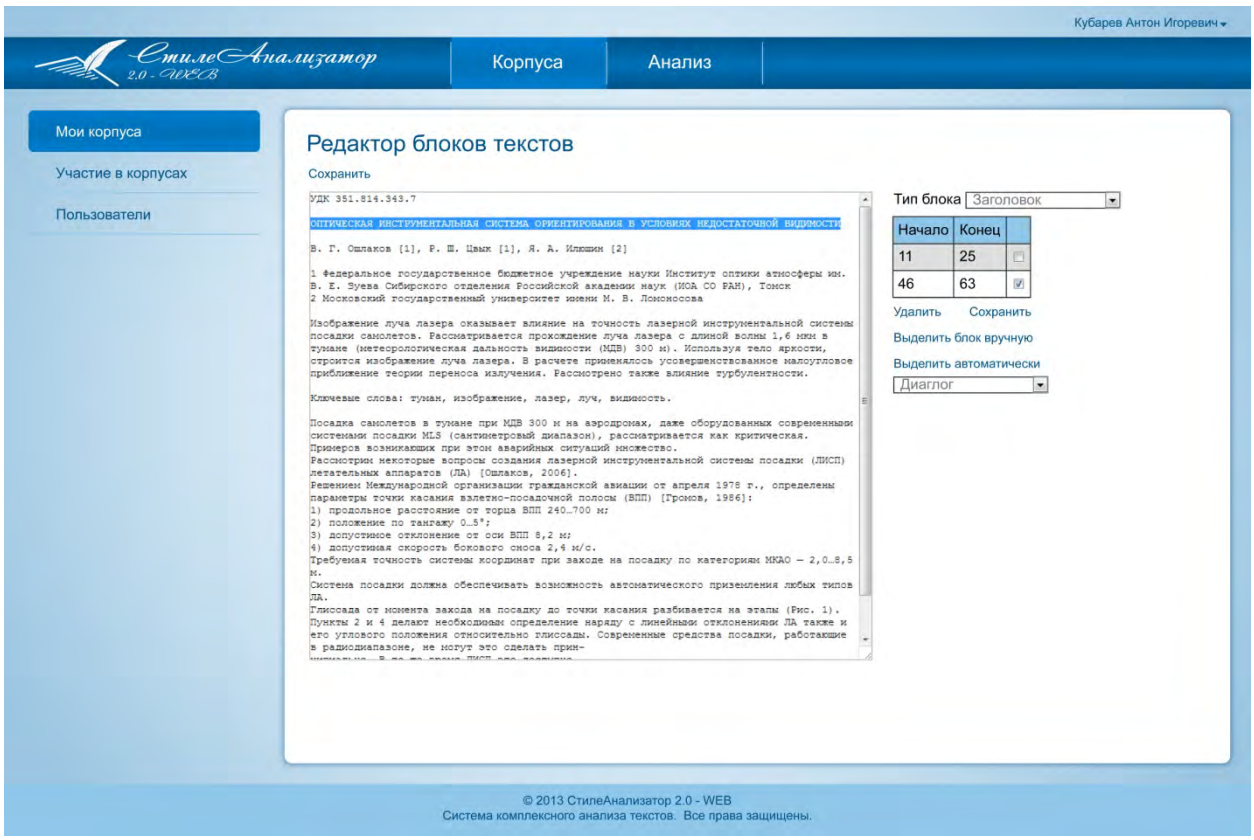


Рис.2.5 – Страница редактора блоков текста.

Также владелец может просмотреть статистику по корпусу, которая отображает общий объем слов в корпусе, общий объем различных слов, типы используемых блоков при разметке и их «объем участия» в корпусе (Рис 2.6).

Для создания редуцированных корпусов владельцу корпуса необходимо перейти на страницу «Редуцировать», которая позволяет выбрать необходимые блоки корпуса и сформировать из них новый корпус (или изменить текущий). Для создания нового корпуса пользователю необходимо заполнить поле «Имя корпуса». Если поле оставить пустым, система выдаст предупреждение о том, что данные корпуса будут изменены. В случае положительного ответа пользователя все не выбранные блоки будут удалены из корпуса безвозвратно, в противном случае пользователь вернется на предыдущую страницу для ввода информации о новом корпусе. (Рис 2.7)

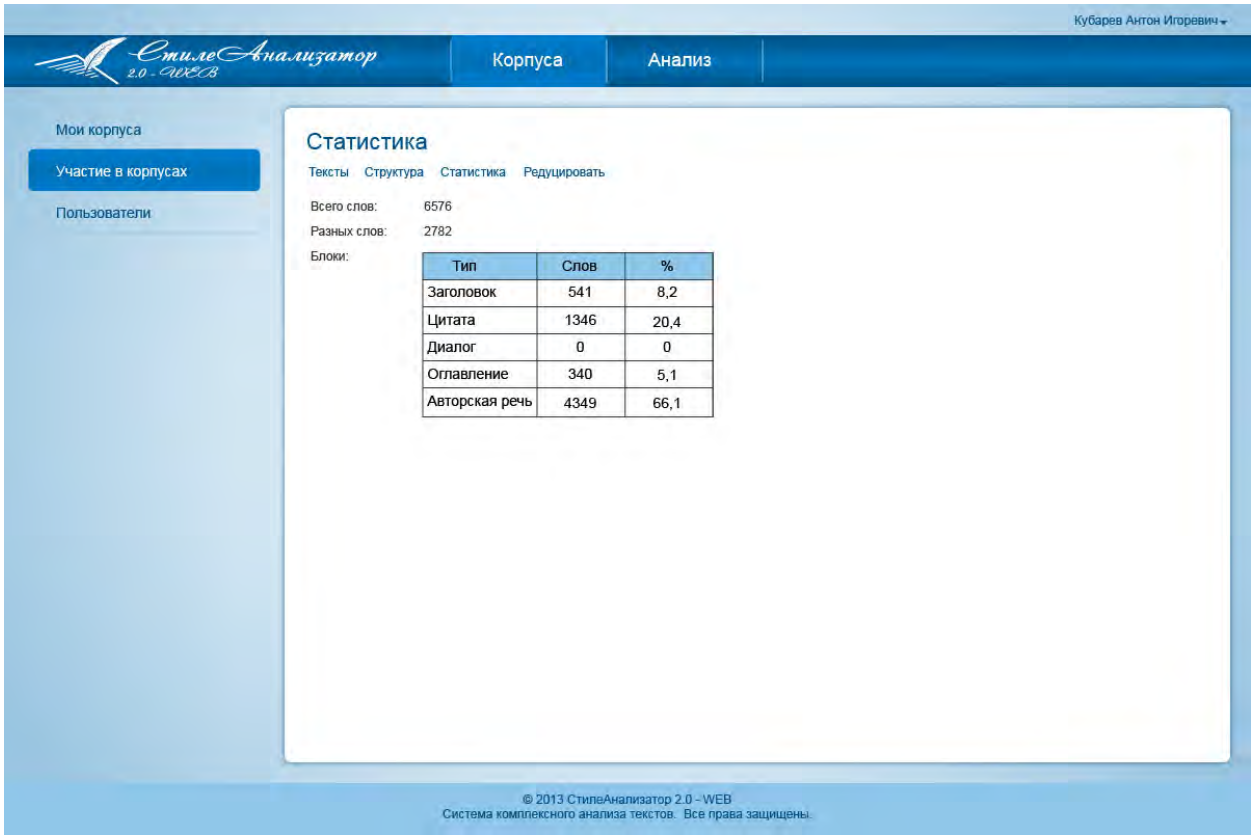


Рис 2.6 – Страница просмотра статистики корпуса

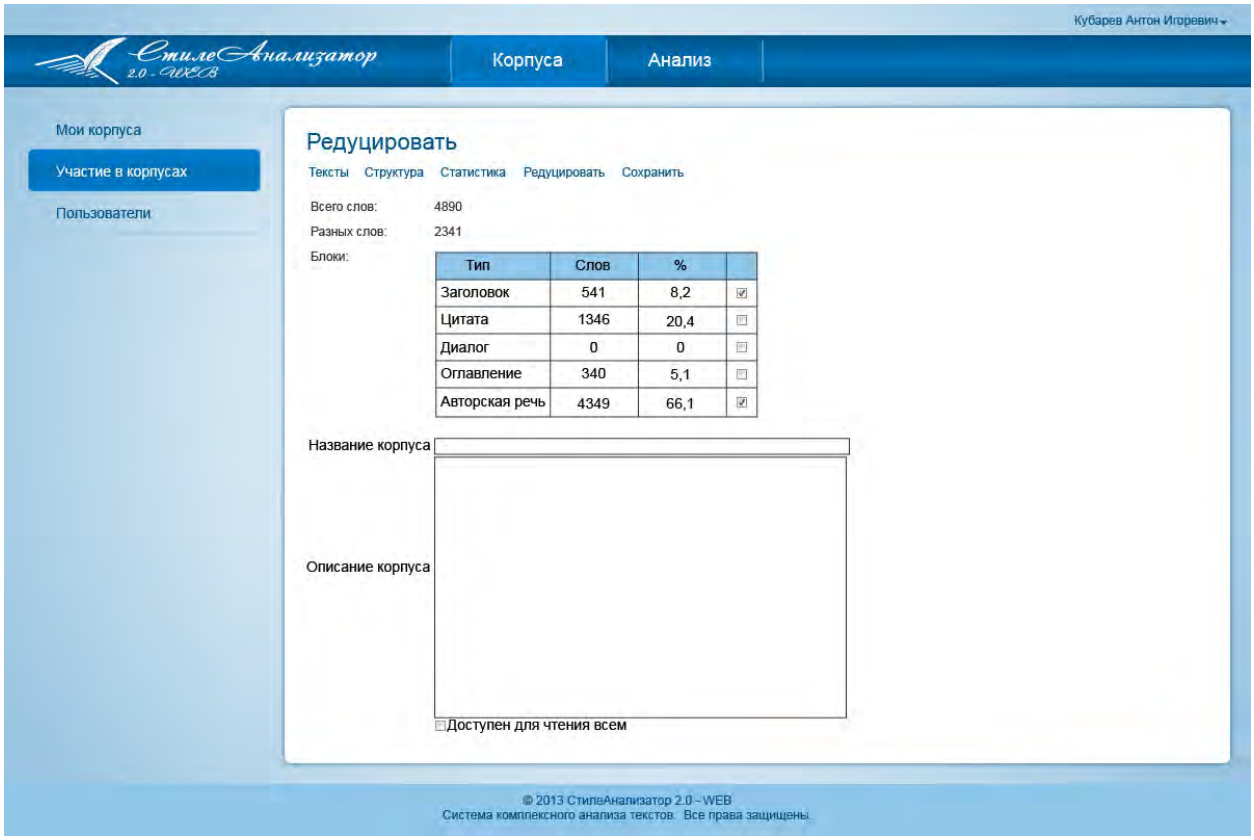


Рис 2.7 - Страница формирования редуцированного корпуса

Администратор/модератор/участник корпуса – пользователи, приглашенные владельцем для совместной работы над корпусом в рамках системы. Приглашение отправляется владельцем корпуса, также все пользователи системы могут обмениваться текстовыми сообщениями (Рис. 2.8, 2.9).

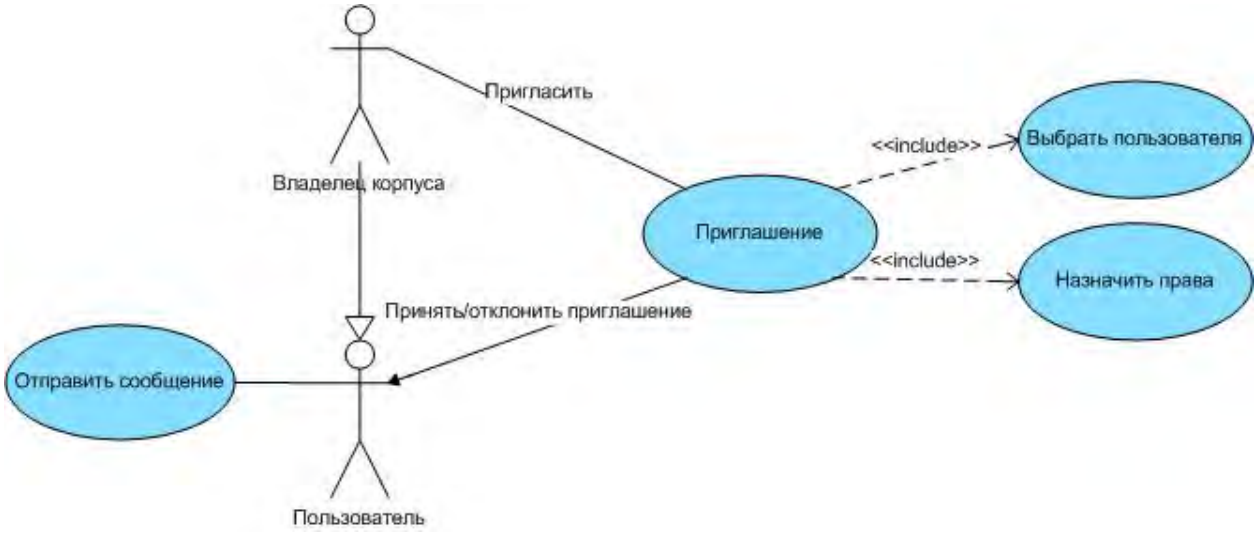


Рис. 2.8 – Схема коммуникации пользователей в системе.

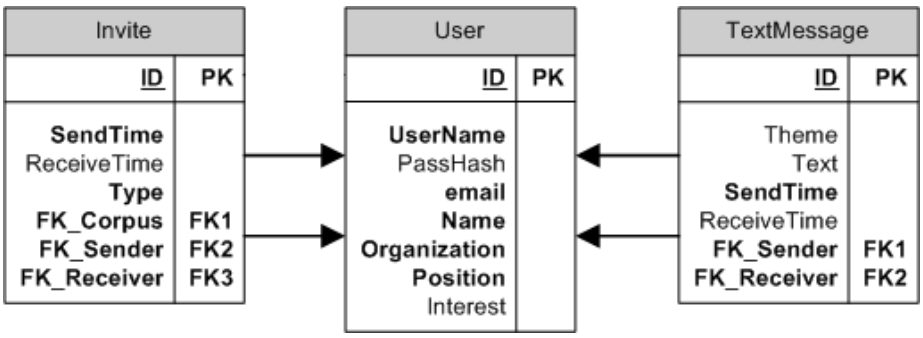


Рис. 2.9 – ER-диаграмма БД, отвечающая за коммуникацию пользователей в системе

Пользователь (сторонний) – все остальные пользователи системы, которым владелец корпуса может дать права на чтение данных и их анализ. (Рис. 2.10).

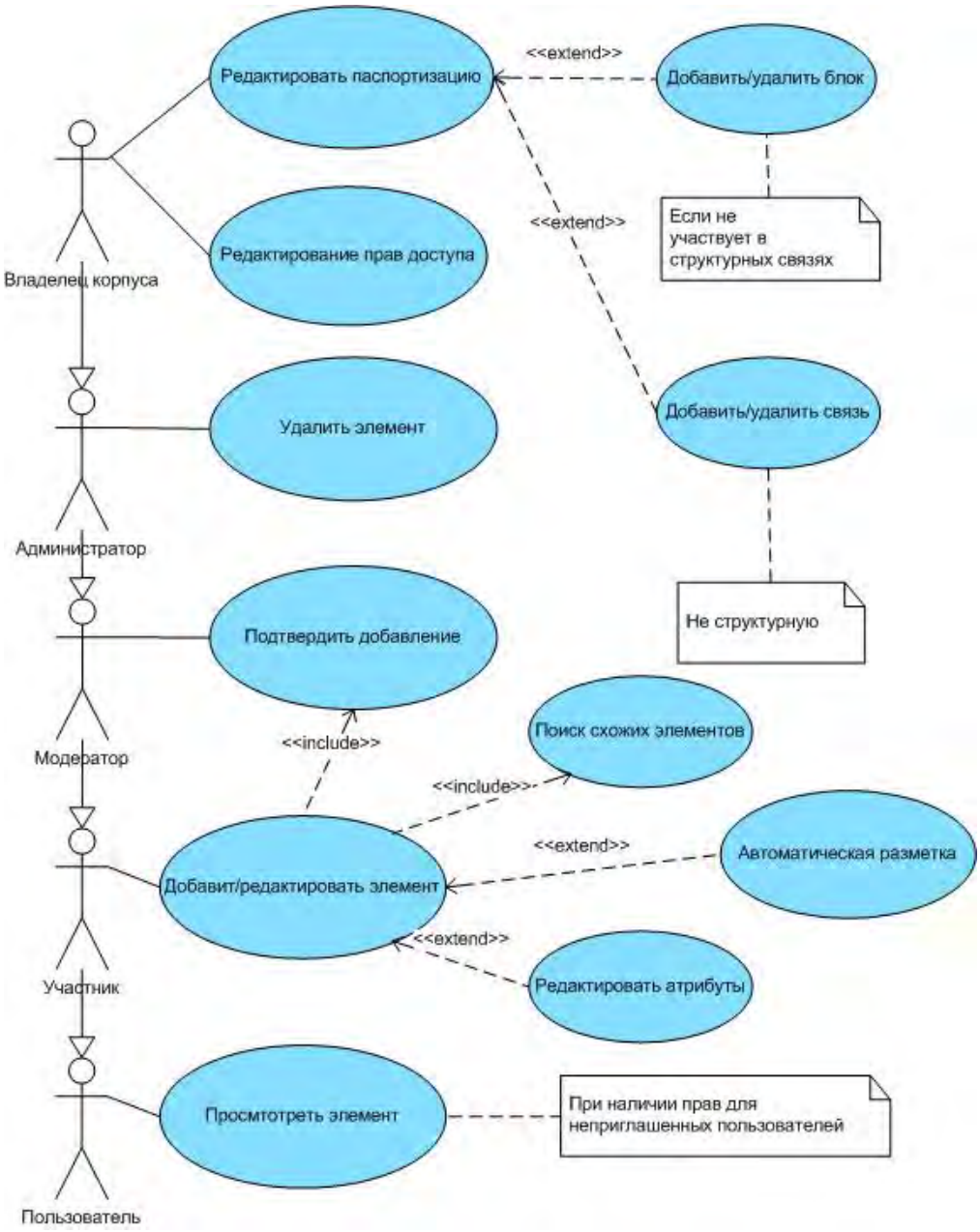


Рис. 2.10 – Схема функций взаимодействия пользователей с корпусом текстов.

Таким образом, участник корпуса может добавлять данные в корпус и изменять их атрибуты, при условии подтверждения добавления модератором. При выполнении данного функционала пользователь может использовать методы автоматической разметки (на основе поиска подобных элементов в корпусе, добавленных ранее), а также пользоваться разметкой на основе конкорданса, получаемого по всему корпусу или по каким либо типам блоков. Также существует возможность выделения неразмеченных блоков, т.е. блоков с пустыми значениями атрибутов. Правами на удаления данных из корпуса обладают только администратор и владелец корпуса.

Все типы пользователей наследуются по принципу «старший от младшего», т.е. модератор корпуса обладает правами участника корпуса, а владелец корпуса обладает всеми правами.

При необходимости в структуру корпуса можно включить типы блоков, отвечающих за композиционные элементы текстов: заголовки, оглавления, предисловия, послесловия,

диалоги, авторская речь и т.д., что на последующем этапе анализа позволит более гибко подготавливать текстовый материал. Размечаться такие блоки могут также в полуавтоматическом режиме.

Данная архитектура позволяет организовать распределенную работу коллектива с развитым диалогом пользователей и центральным сервером, хранящим все данные.

1.3. ХРАНЕНИЕ КОРПУСОВ ТЕКСТОВ ПРОИЗВОЛЬНОЙ СРУКТУРЫ И ПАСПОРТИЗАЦИИ В РЕЛЯЦИОННОЙ БАЗЕ ДАННЫХ

1.3.1. Введение и постановка задачи

Концепция системы «СтилеАнализатор 2.0 WEB» предполагает хранение анализируемых корпусов текстов на сервере системы в реляционной базе данных.

Изначальная схема базы данных позволяла хранить тексты, обладающие однотипной паспортизацией и включающие в себя одну структурную единицу – слово. Знаки препинания, числа и прочие символы также считались словами (Рис. 3.1).

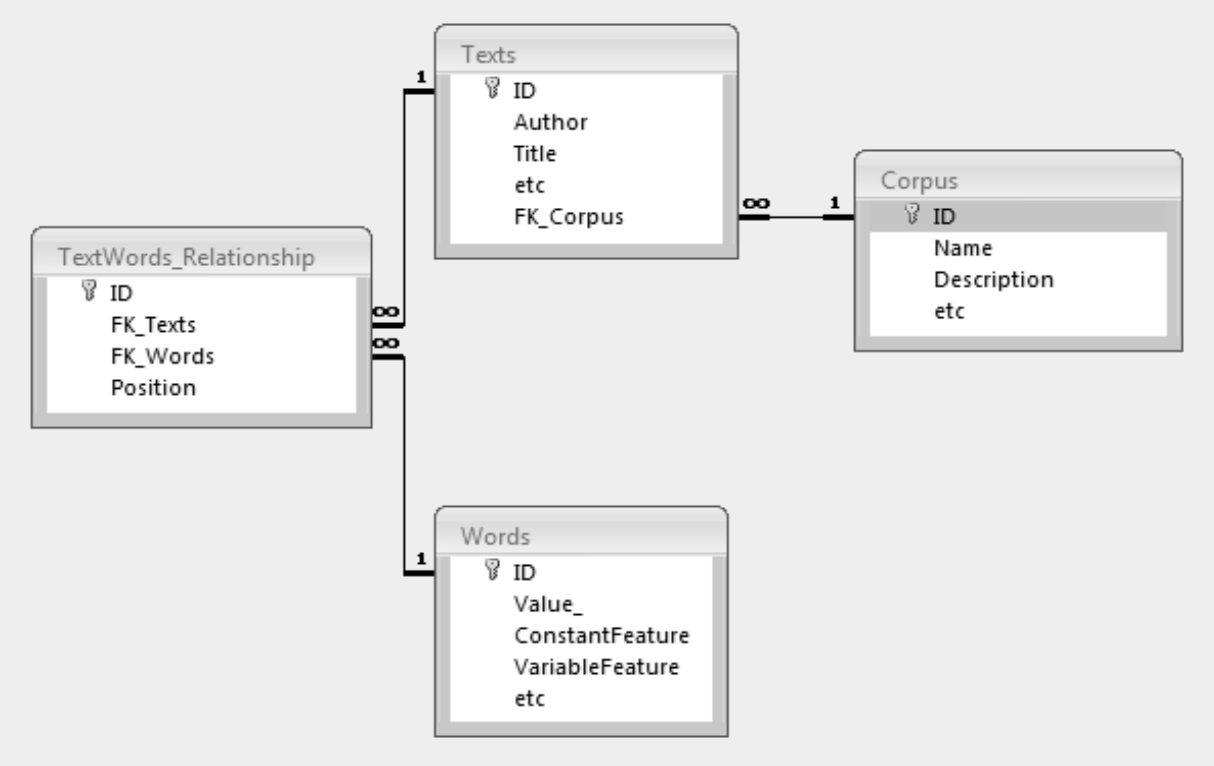


Рис. 3.1. Реляционная диаграмма изначальной схемы данных.

Схема отражает основные принципы хранения текстового материала, заложенные в систему «СтилеАнализатор» [1] – прототип системы «СтилеАнализатор 2.0 WEB».

При анализе требований потенциальных пользователей к системе в данной схеме были выявлены следующие недостатки:

- Отсутствие возможности хранения корпусов с различной паспортизацией текстов. Например, корпуса художественной литературы XIX в. и корпуса публицистики XXI в.
- Отсутствие гибкости структуры текстов, не допускающей хранения дополнительных единиц: предложений, словосочетаний, морфем и т.д.

- Единство принципов хранения, вне зависимости от различных требований пользователей.

Поиск решения возникших проблем привел к схеме данных, позволяющей хранить корпуса с различной паспортизацией и структурой текстов в рамках единой реляционной базы данных.

1.3.2. История изменений схемы данных

Изменения в схеме данных производились поэтапно, по мере поэтапного появления потенциальных пользователей и их требований к хранению корпусов текстов.

Первым этапом изменений был отказ от жестко заданной паспортизации текстов (Рис. 3.2).

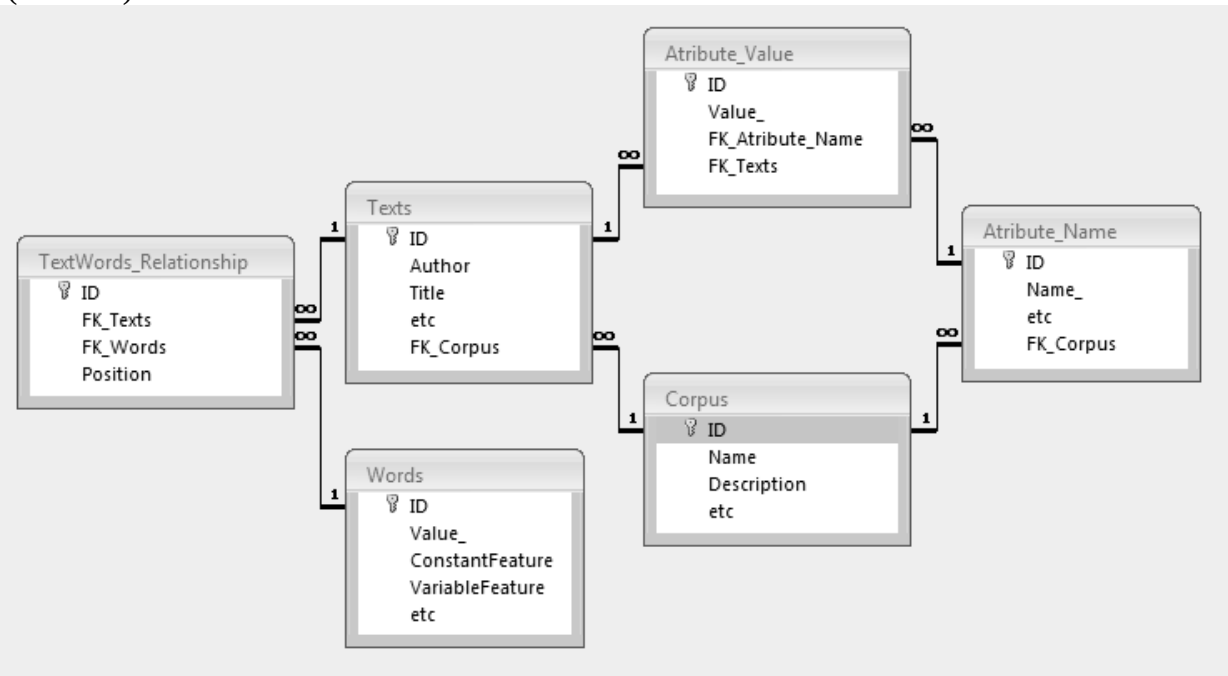


Рис. 3.2. Реляционная диаграмма схемы данных с добавлением атрибутов текстов.

В результате каждому корпусу в соответствие был поставлен набор атрибутов, задающих паспортизацию корпуса, а каждому тексту – набор значений данных атрибутов, составляющих паспорт текста.

Следующим этапом изменений стало введение дополнительных структурных единиц (Рис. 3.3). Первоначально на этом этапе были добавлены предложения и морфемы. Данная схема отвечала минимальному набору требований, хоть и не представляла в достаточно полном виде структурного состава любого текста.

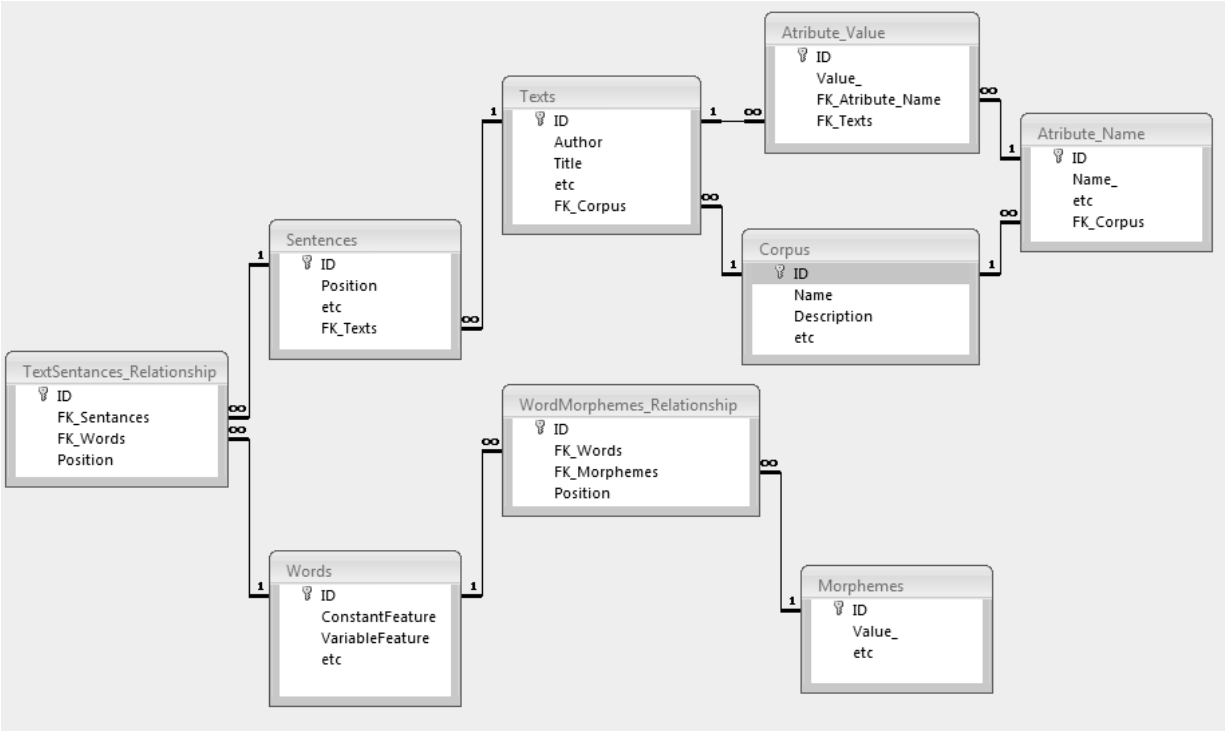


Рис. 3.3. Реляционная диаграмма схемы данных с добавлением предложений и морфем.

Для решения проблемы гибкости структурного состава текста, было введено понятие «блок». Блоком является любая структурная единица корпуса. В том числе и текст, таким образом, корпус есть набор блоков высшего уровня, каждый блок одного уровня состоит из блоков уровня ниже, спускаясь до низшего блока, представляющего символы.

1.3.3. Результирующая схема данных

После серии модификаций была получена следующая схема данных (Рис 3.4.).

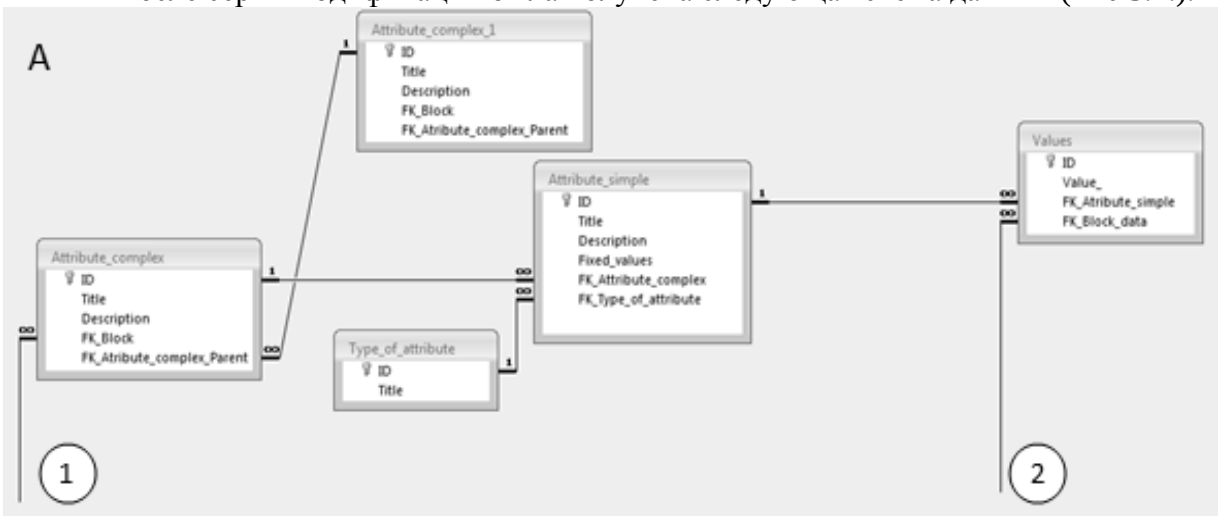


Рис. 3.4а. Реляционная диаграмма конечной схемы данных. А – Атрибуты и данные.

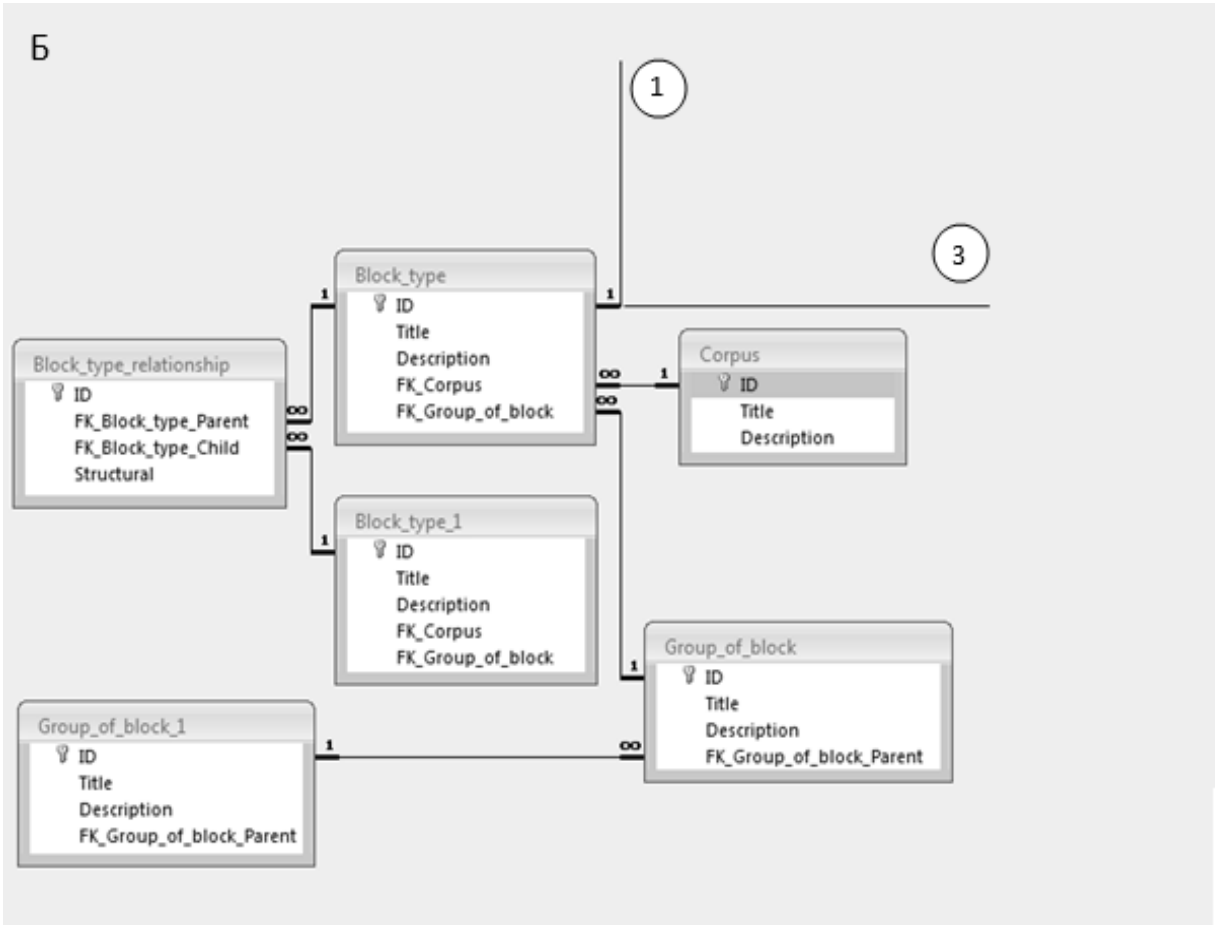


Рис. 3.4б. Реляционная диаграмма конечной схемы данных. Б – Типы блоков данных.

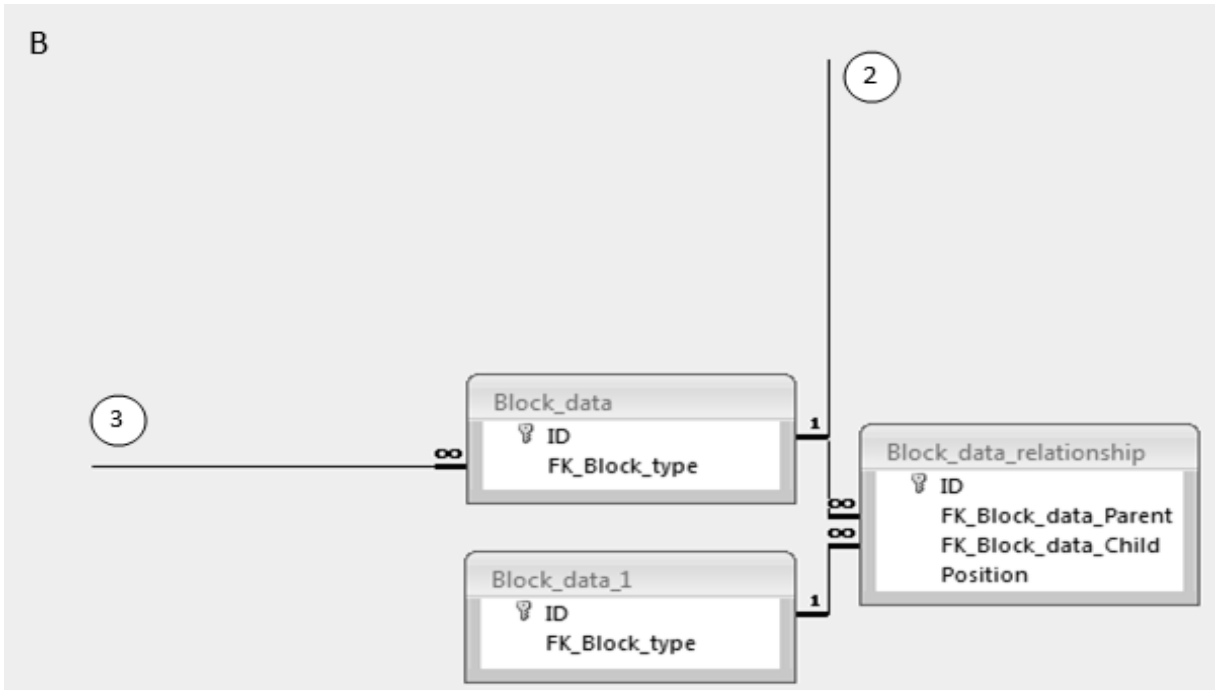


Рис. 3.4в. Реляционная диаграмма конечной схемы данных. В – Организация связей данных.

Основные отличия данной схемы состоят в следующем:

- Введено понятие «Группа блоков», объединяющее блоки в группы со схожим смыслом.
- Понятие атрибута разделено на «Комплексный атрибут» и «Простой атрибут». Комплексный атрибут является группой простых атрибутов, объединённых по какому либо признаку. Между комплексными атрибутами установлено отношение типа «Родитель-ребенок».

Таким образом, все данные хранятся в одной сущности *«Values»*. Логическая организация этих данных осуществляется при помощи набора сущностей *«Atribute_Simple»*, *«Atribute_Complex»*, *«Block_type»* и их связей. Сущность *«Block_Data»* организует связи между данными и логикой их хранения. За основу данной схемы были взяты принципы Универсальной модели данных [2, 3].

Полученная схема позволяет в рамках единой реляционной базы данных хранить корпуса с различной паспортизаций и различной структурой текстов.

Материал раздела опубликован в работе [4].

1.4. Построение DFT-таблиц

Система «СтилеАнализатор 2.0 WEB» в блоке статистического анализа использует частотные таблицы. Каждый пользователь системы может строить свои таблицы и открывать доступ для их использования всем участникам корпуса (Рис. 4.1).

В качестве объектов (строк) в DFT-таблицах выступают блоки данных корпуса, выбранные пользователем, например, блоки данных типа «Текст» с атрибутом «Автор::Фамилия», равным «Пушкин». Признаками выступают блоки данных определенного типа, связанные с блоками данных объектов таблицы (например, слова текста), которым принадлежит атрибут с определенным значением, например, блоки данных типа «Слово» с атрибутом «Часть речи», равным «Существительное».

Данные признаки могут объединяться в более сложные путем объединения в логическое выражение. Признаки, объединенные в одно выражение, должны затрагивать один тип блоков данных. Например, блоки данных типа «Слово» с атрибутом «Часть речи», равным «Существительное», и атрибутом «Число», не равным «Множественное».

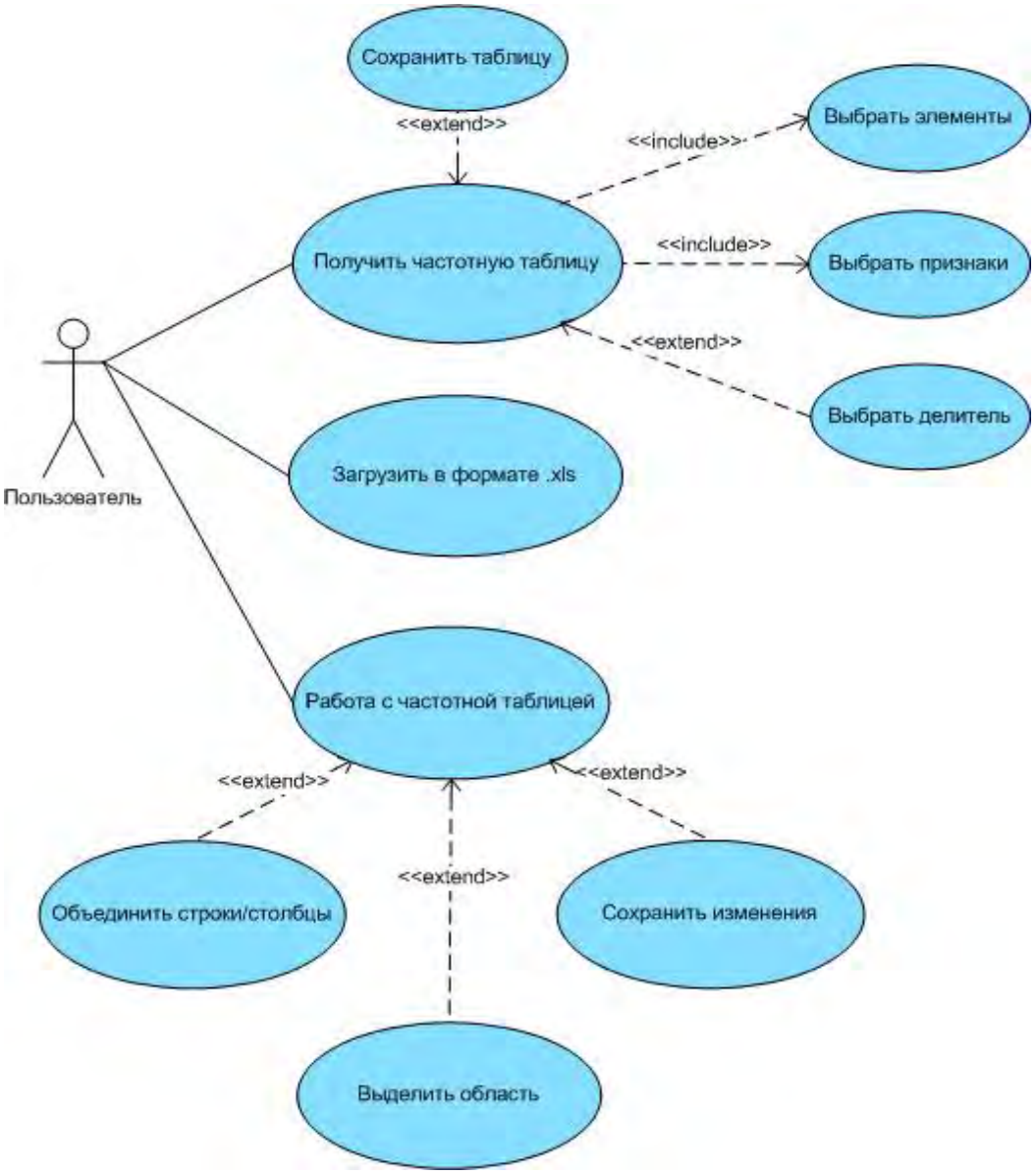


Рис. 4.1 – Схема функций взаимодействия пользователей
с блоком построения DFT-таблиц.

Результирующие DFT-таблицы хранятся в БД системы в виде набора таблиц (Рис. 4.2), дублирование связей между которыми введено с целью оптимизации скорости выполнения запросов на загрузку таблицы из БД при выполнении статистических методов анализа.

Описанный подход, позволяет производить оперативное переформирование корпусов текстов, предоставляя пользователю инструментарий для выбора анализируемых блоков по различным признакам. Например, при соответствующей структуре корпуса можно использовать в анализе только введения или не использовать диалоги, тем самым проводя анализ над «редуцированными текстами»

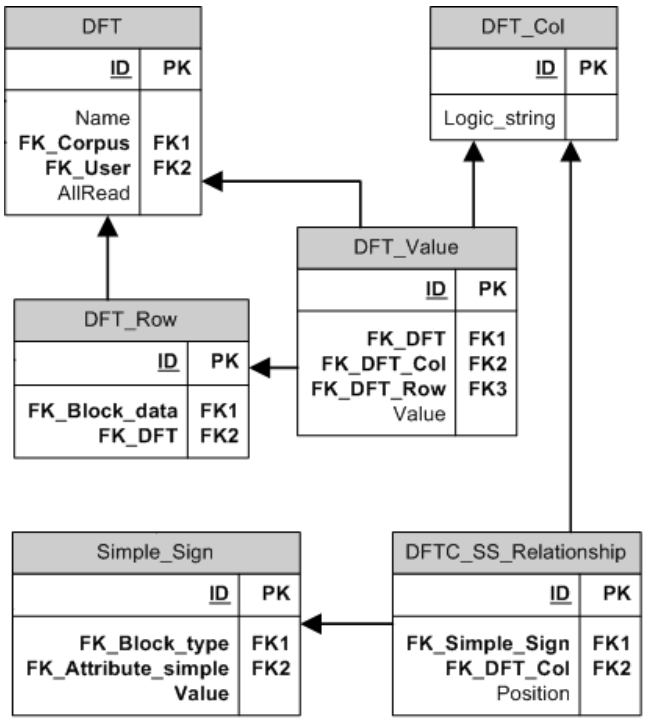


Рис. 4.2 – ER-диаграмма БД отвечающая за хранение DFT-таблиц.

Над полученной таблицей можно проводить операции типа суммирования строк и столбцов, выделение областей. Построение DFT-таблиц системой осуществляется следующим образом. Последовательно перебираются признаки, для каждого признака перебираются все объекты, формируется анализируемый материал (наборы блоков данных), вычисляется количество блоков, удовлетворяющих признаку (Рис 4.3).

Пользователь выбирает пункт меню «Подсчет значений признаков», вызывается метод Show(), выводятся параметры подсчета. Когда пользователь ввел необходимые данные, вызывается метод Show() класса CalcFeatures. Сразу же вызывается метод подсчета значений Calc(). Внутри него вызывается метод CalcNextFragment(). Результат сохраняется в БД, после чего формируется страница с представлением результата, которая отображается пользователю.

..

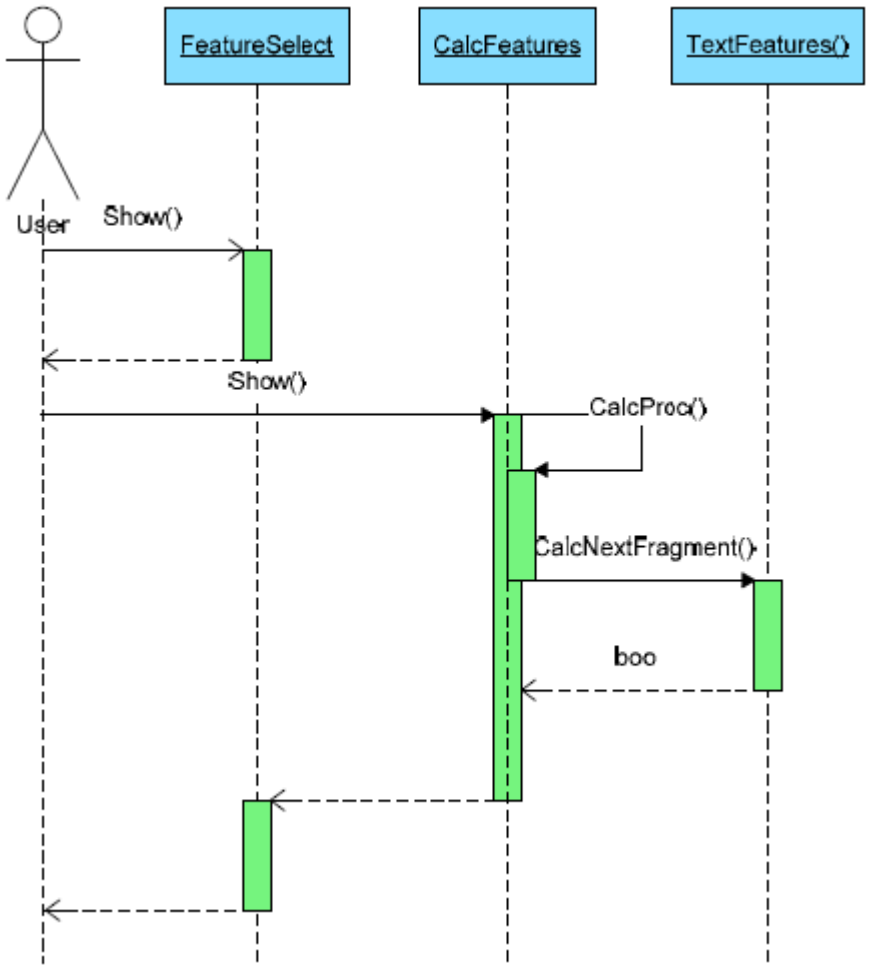


Рис. 4.3 – Диаграмма последовательности подсчета значений признаков.

1.5. Блок анализа

В «СтилеАнализаторе 2.0 WEB» был использован блок анализа системы «СтилеАнализатор», дополненный новыми методами анализа. Для проведения анализа пользователю необходимо выбрать DFT-таблицу, метод анализа и параметры, при необходимости их отличия от параметров по умолчанию (Рис. 5.1).

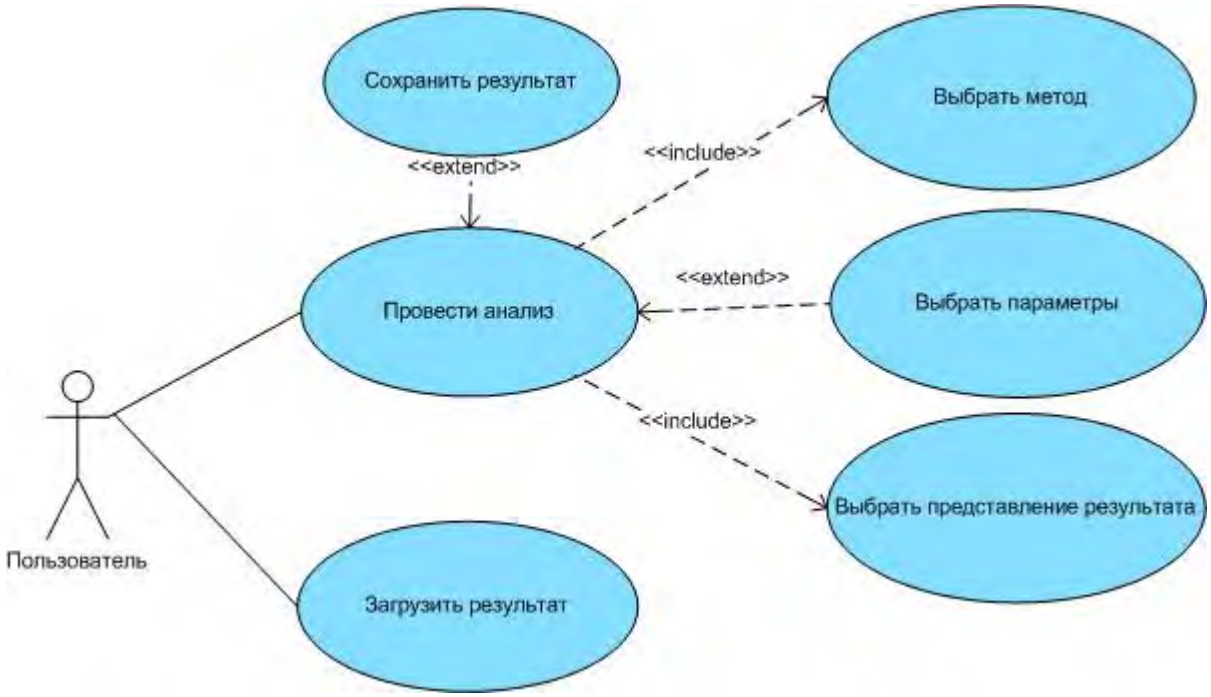


Рис. 5.1 – Схема функций взаимодействия пользователей
с блоком анализа.

Результат проведенного анализа пользователь может загрузить к себе на компьютер в виде электронной таблицы (*.xls) или диаграммы (*.png).

Блок анализа данных системы «СтилеАнализатор 2.0 WEB» включает в себя:
− статистический анализ текстов (подсчёт значений признаков, факторный анализ, в том числе метод главных компонент, дискриминантный анализ, кластерный анализ, методы байесовской классификации),
− информационный анализ и классификацию (деревья решений),
− логический анализ и тестовое распознавание,
− нейронные сети прямого распространения,
− самоорганизующиеся карты Кохонена (Self-Organizing Maps − SOM-сети),
− классификацию на основе суффиксных деревьев и др.
– построение гистограмм распределений признаков с наложением кривой гауссова распределения для визуального анализа.

Кластеризация

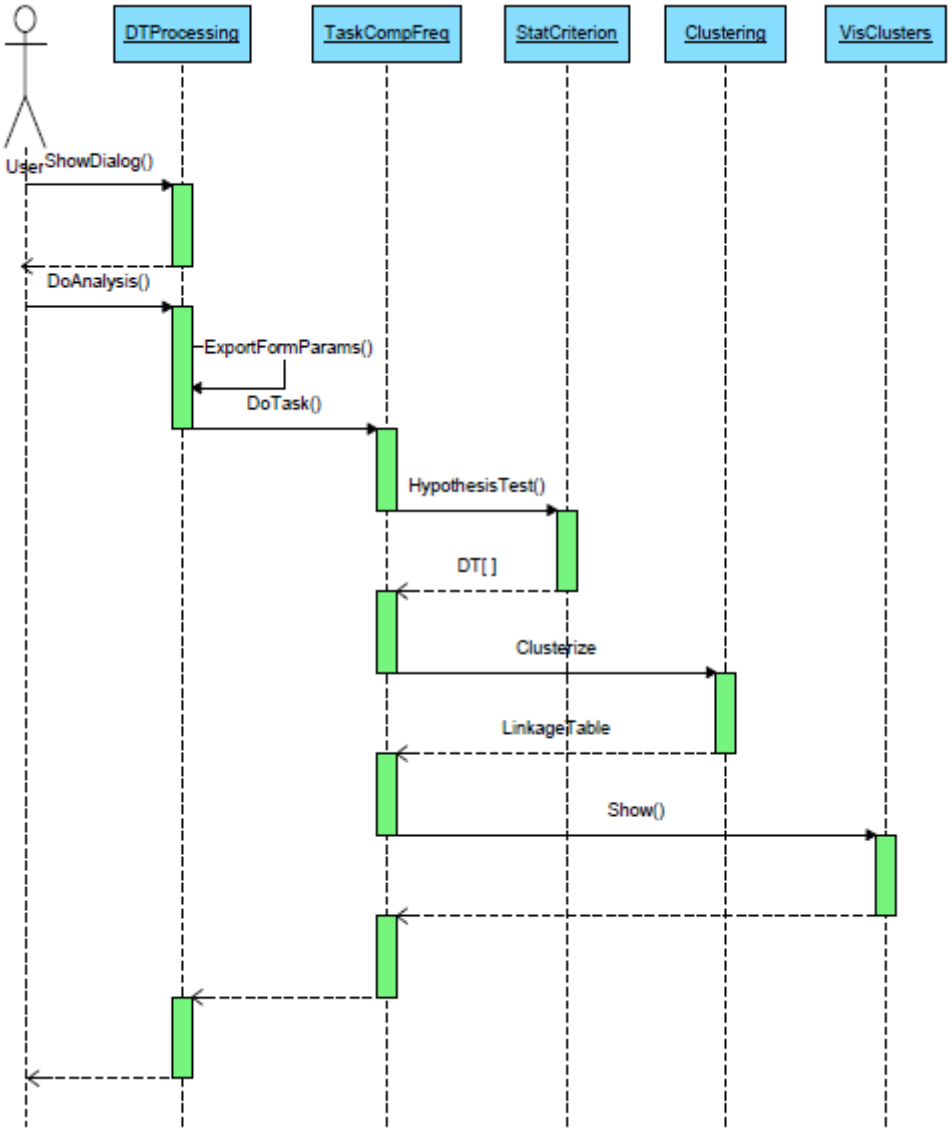


Рис. 5.2 - Диаграмма кластеризации.

Кластеризации текстов проводится на основе сравнения отдельных частот и на основе сравнения распределений. Метод ExportFromParams() выполняется для любого из вариантов анализа, далее начинаются различия.

Кластеризация с помощью SOM-сетей Кохонена

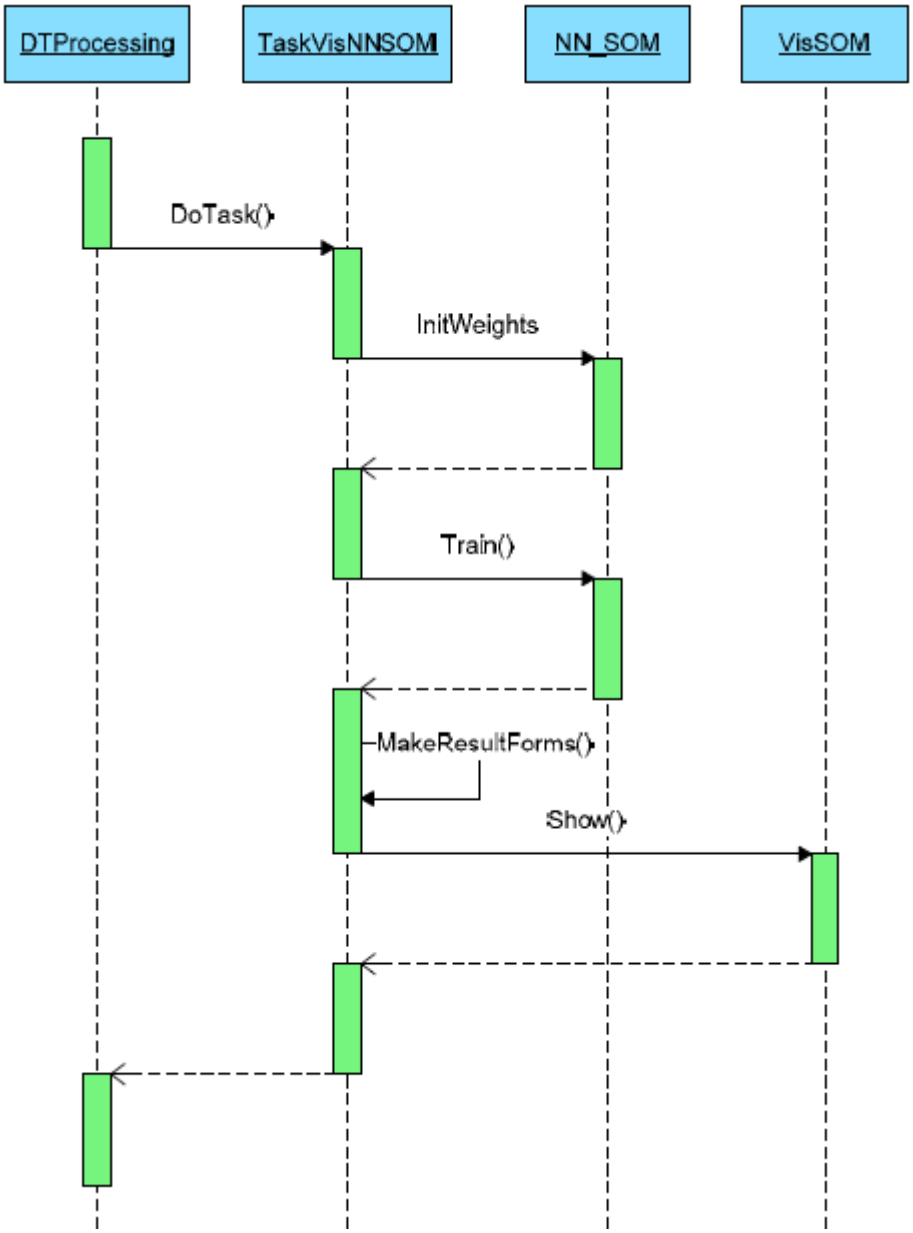


Рис. 5.3 - Диаграмма кластеризации с помощью SOM-сети.

Метод InitWeights() определяет веса. Метод Train() обучает сеть. Метод MakeResultForm() отображает результаты.

Классификация с помощью деревьев решений

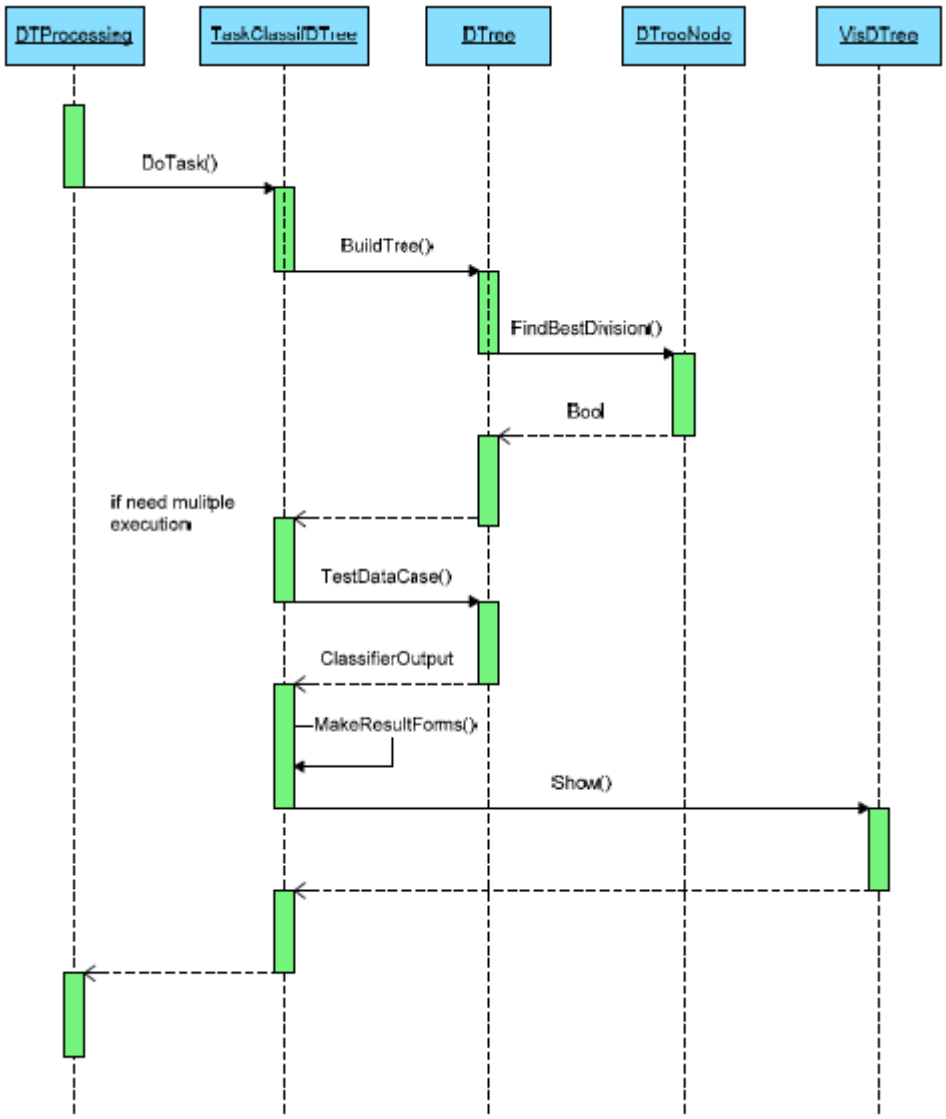


Рис. 5.4 - Диаграмма классификации с помощью деревьев решений.

Чтобы построить дерево решений, вызывается метод BuildTree(), внутри которого вызывается метод FindBestDivision(), который вычисляет наилучшее по информационному критерию разбиение множества признаков. После того, как дерево решений построено, метод MakeResultForms() отображает результаты.

Классификация на основе статистических и информационных мер

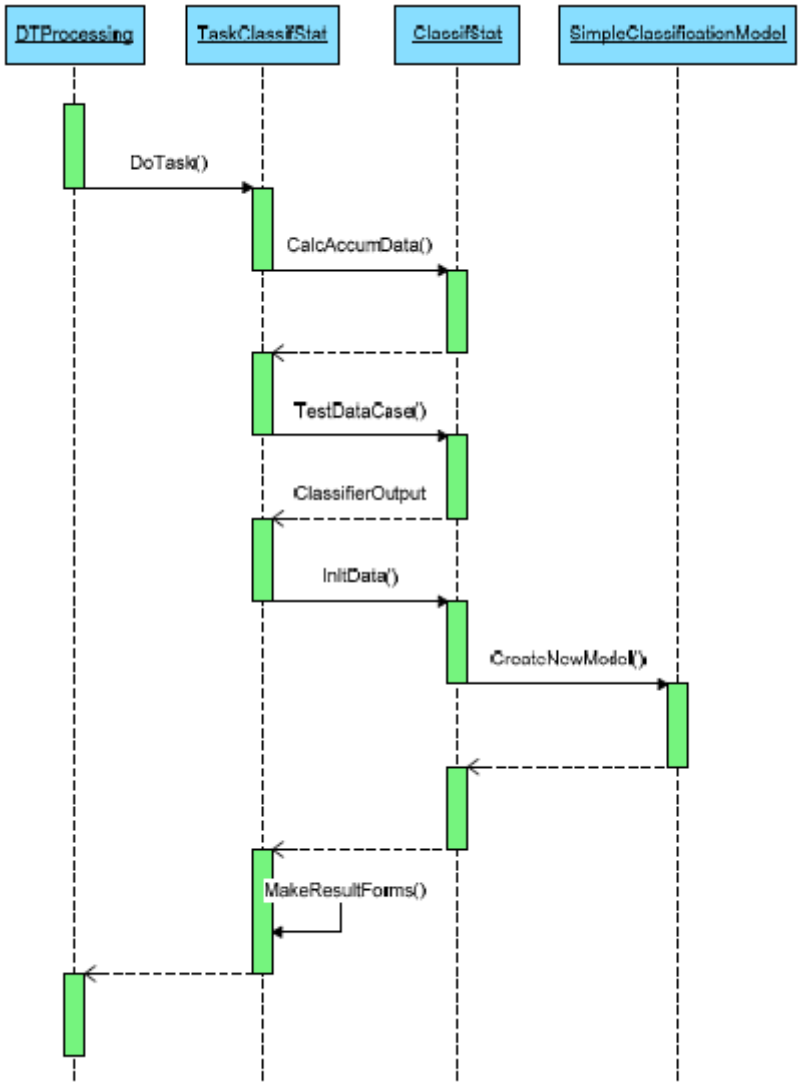


Рис. 5.5 - Классификация на основе статистических и информационных мер.

Метод CalcAcumData() обучает модель с помощью обучающей выборки (набора текстов, принадлежность которых к изучаемым классам известна). Метод TestDataCase() проверяет все образцы данных и составляет таблицу классификации, возвращая ее в качестве результата. В итоге работы метод MakeResultForms() отображает результаты.

Классификация с помощью нейронной сети прямого распространения

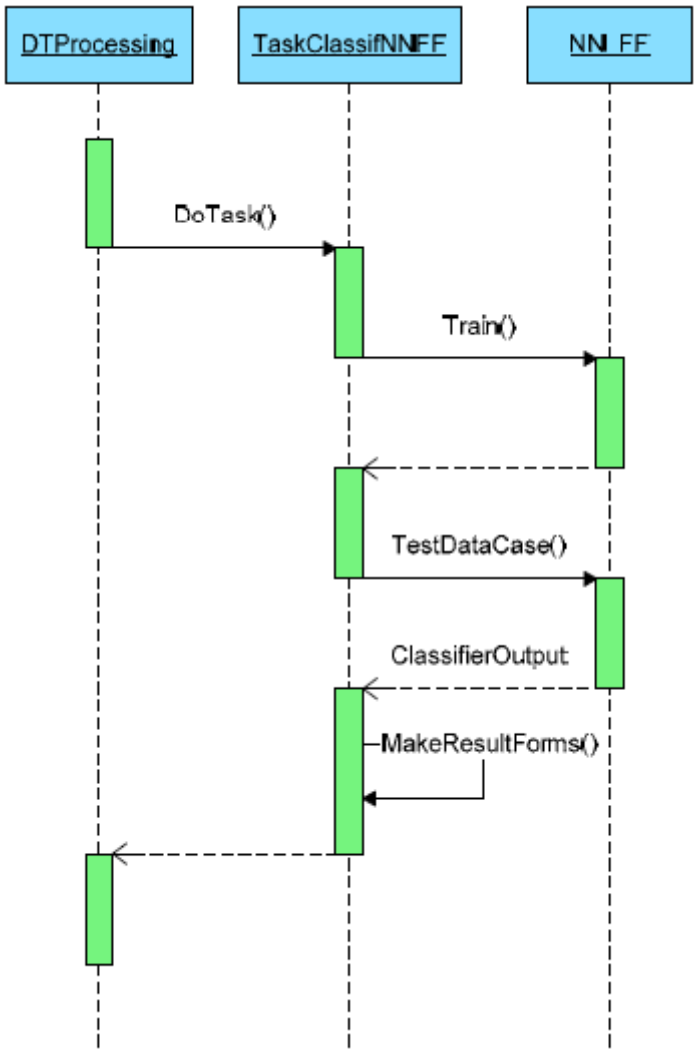


Рис. 5.6 - Диаграмма классификации с помощью нейронной сети прямого распространения.

Метод Train() обучает нейронную сеть. Метод TestDataCase() проверяет все образцы данных и составляет таблицу классификации, возвращая ее в качестве результата. Метод MakeResultForm() отображает результаты работы.

Классификация на основе суффиксных деревьев (потоковые методы)

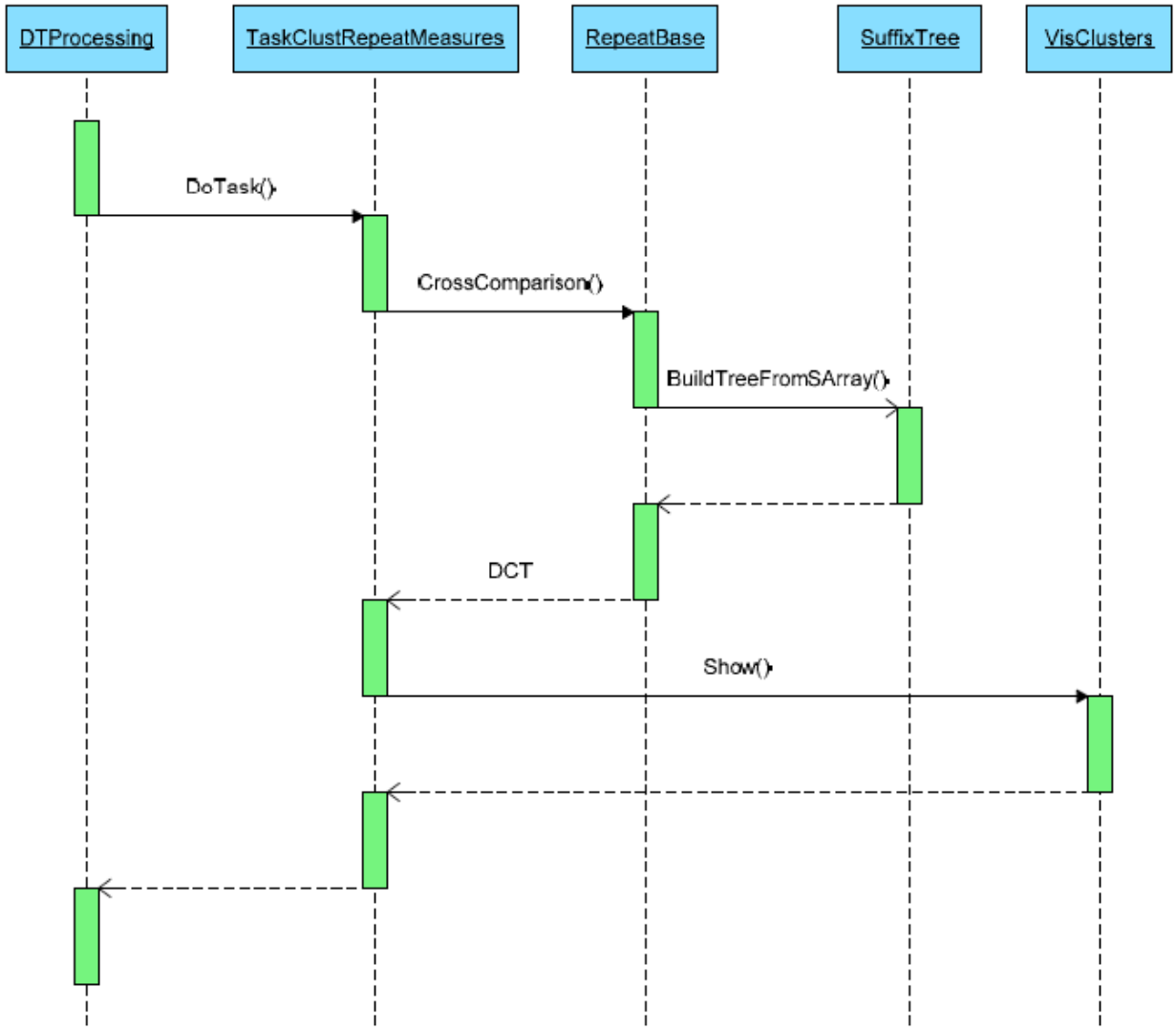


Рис. 5.7 - Диаграмма потокового метода классификации на основе суффиксных деревьев.

Метод CrossComparison() выполняет перекрестное сравнение суффиксных деревьев, внутри этого методы вызывается метод BuildTreeFromSArray(), который строит суффиксное дерево.

Построение гистограмм распределений

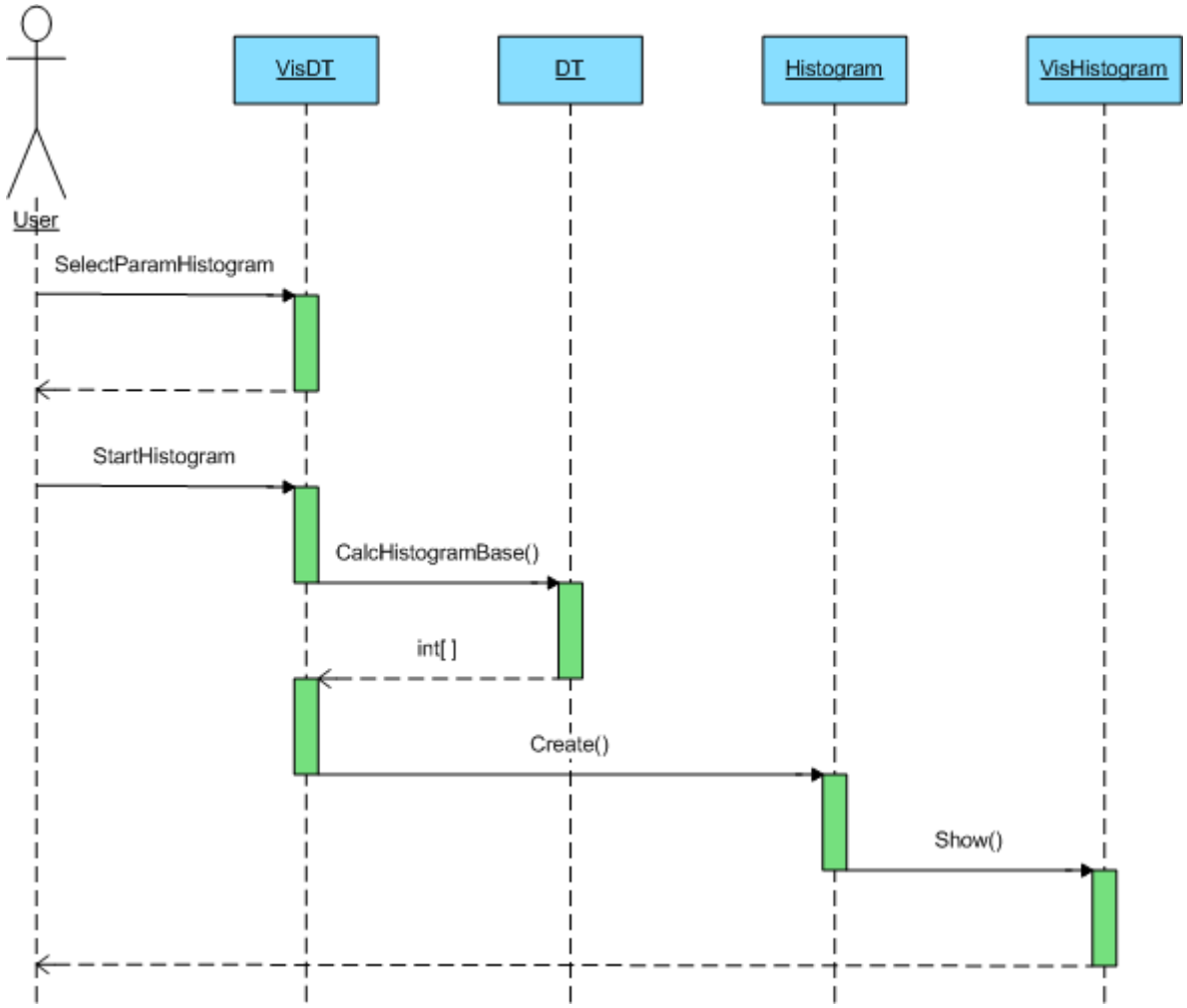


Рис. 5.8 - Диаграмма построения гистограммы

Литература к разделу

1. Шевелев, О.Г. Разработка и исследование алгоритмов сравнения стилей текстовых произведений: Автореф. дис. … канд. техн. наук / Том. гос. ун-т. – Томск, 2006. – 19 с.

2. Есин, В.И., Пергаменцев, Ю.А. Технология проектирования модели предприятия на основе универсальной модели данных. – [Электронный ресурс]. URL: http://citforum.ru/database/articles/udm/ (дата обращения 25.09.2012)

3. Муса-Оглы, Е., Бессарабов, Н. Универсальная модель данных в Oracle. – [Электронный ресурс]. URL: http://www.interface.ru/home.asp?artId=24052 (дата обращения 05.10.2012)

4. Кубарев А.И., Поддубный В.В. Хранение корпусов текстов произвольной структуры и паспортизации в реляционной базе данных // Информационные технологии и математическое моделирование (ИТММ-2013): Материалы XII Всероссийской научно-практической конференции с международным участием им. А. Ф. Терпугова (29-30 ноября 2013 г.). – Томск: Изд-во Том. ун-та, 2013. – Ч.1. – С.139–143.

2. РАЗВИТИЕ ИССЛЕДОВАТЕЛЬСКИХ ФУНКЦИЙ СИСТЕМЫ

2.1. БАЙЕСОВСКАЯ КЛАССИФИКАЦИЯ С ОБУЧЕНИЕМ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ КОПУЛА-ФУНКЦИЙ

2.1.1. Введение и постановка задачи

Рассмотрим множество n объектов, каждый из которых принадлежит одному из K классов и характеризуется набором m числовых признаков $a_1, \ldots, a_m$. Пусть имеется n_k объектов k-го класса, так что $n = \sum_{k=1}^{K} n_k$. Значение j-го признака i-го объекта из k-го класса обозначим x_{ijk}. Тогда этот объект можно охарактеризовать вектором-строкой $x_{ik} = \left(x_{i1k},...,x_{ijk},...,x_{imk}\right)$. Эту строку будем рассматривать как i-ю реализацию векторной случайной величины ξ_k, подчиняющейся распределению вероятностей с плотностью $p\left(x_1,...,x_m \mid k\right)$, своей для каждого класса k.

Пусть теперь наблюдается объект, принадлежность которого к какому-либо классу неизвестна. Возникает проблема определения его «классовой принадлежности». Для решения этой задачи используют байесовский классификатор [1], вычисляющий апостериорную вероятность класса наблюдаемого объекта по формуле Байеса и относящего объект к апостериорно наиболее вероятному классу $\hat{k}$:

$$\hat{k} = \arg\max_k P\left(k \mid x_1,...,x_m\right) = \arg\max_k p\left(x_1,...,x_m \mid k\right)P\left(k\right), \qquad (1)$$

где $P(k)$ – априорная вероятность того, что объект относится к k-му классу.

Широко распространённый и справедливо критикуемый «наивный» байесовский классификатор [2] исходит из предположения о независимости значений признаков объекта, представляя совместную плотность распределения $p\left(x_1,...,x_m \mid k\right)$ вектора признаков произведением $\prod_{j=1}^{m} p\left(x_j \mid k\right)$. Другая сторона «наивности» такого классификатора состоит в том, что плотность распределения любого j-го признака $p\left(x_j \mid k\right)$ предполагается известной. Эти два предположения резко сужают область применимости «наивного» байесовского классификатора.

Оба предположения «наивного» байесовского классификатора можно обойти, используя обучающую выборку наблюдений объектов, принадлежность которых к своим классам известна. Обучающая выборка в принципе даёт возможность получить статистическую оценку совместной плотности распределения признаков для каждого класса и построить байесовский классификатор уже без «наивных» предположений. Однако, для адекватной оценки совместной плотности распределения признаков может потребоваться слишком большая обучающая выборка, которой может не быть в природе. Поэтому в данной работе предлагается подход, основанный на гауссовых копула-функциях. Они позволяют учесть не всю статистическую зависимость признаков, а только её корреляционную составляющую, и привести в общем случае негауссовы маргинальные распределения наблюдений признаков к их совместно-гауссовой форме, обеспечивающей возможность построения байесовского классификатора.

2.1.2. Эмпирическая функция распределения обучающей выборки и её кусочно-линейная аппроксимация

Набор n_k выборок $(x_{i1k},...,x_{imk})$, $i=\overline{1,n_k}$, из k-го класса, используемых в качестве обучающих, позволяет построить маргинальную эмпирическую функцию распределения $\hat{F}_j(x\,|\,k)$ каждого j-го признака объекта k-го класса, являющуюся состоятельной статистической оценкой неизвестной маргинальной функции распределения $F_j(x\,|\,k)$ j-ой компоненты вектора признаков объекта k-го класса:

$$\hat{F}_j(x\,|\,k)=\begin{cases}0, & x\le x_{(1)},\\ \nu/n_k, & x_{(\nu)}<x\le x_{(\nu+1)},\quad \nu=\overline{1,n_k-1},\\ 1, & x>x_{(n_k)},\end{cases} \tag{2}$$

где $x_{(\nu)}$, $\nu=\overline{1,n_k}$, – ν-ая порядковая статистика вариационного ряда наблюдений j-го признака объектов k-го класса, ν – ранг соответствующего наблюдения.

Мажорируем сверху и снизу эту ступенчатую непрерывную слева функцию непрерывными кусочно-линейными функциями (ломаными линиями) с точками излома в угловых точках графика функции (2). Примем среднее арифметическое этих ломаных за сглаженную непрерывную эмпирическую оценку $\tilde{F}_j(x\,|\,k)$ неизвестной непрерывной маргинальной функции распределения $F_j(x\,|\,k)$. Плотность $\tilde{p}_j(x\,|\,k)$ сглаженного эмпирического распределения – кусочно-постоянная функция, значения которой на каждом интервале постоянства (между соседними значениями вариационного ряда) характеризуют скорости изменения сглаженной функции распределения $\tilde{F}_j(x\,|\,k)$ на этих интервалах.

Аналогично можно было бы построить и многомерную эмпирическую (в том числе сглаженную) функцию распределения $\hat{F}_{1,...,m}(x_1,...,x_m\,|\,k)$. Однако при $m>>1$ это практически невозможно, так как требует слишком большого объёма обучающей выборки. Поэтому мы ограничимся построением маргинальных эмпирических распределений, а для учёта статистической зависимости между признаками воспользуемся методом копула-функций [3].

2.1.3. Преобразование наблюдений при фиксированной гипотезе о классе к многомерному нормальному распределению методом копула-функций

Построив по обучающей выборке, соответствующей классу k, сглаженные эмпирические функции распределения $\tilde{F}_j(x\,|\,k)$ для всех признаков, $j=\overline{1,m}$, найдём эмпирические оценки математических ожиданий $\hat{\mu}_{jk}=1/n_k\sum_{i=1}^{n_k}x_{ijk}$ и дисперсий $\hat{\sigma}^2_{jk}=1/(n_k-1)\sum_{i=1}^{n_k}\left(x_{ijk}-\hat{\mu}_{jk}\right)^2$ признаков. Маргинальную нормальную функцию распределения с этими параметрами обозначим $\Phi_j(z\,|\,k)$. Потребуем, чтобы наблюдаемое значение x_j j-го признака испытуемого объекта и соответствующее ему значение z_j нормально распределённой случайной величины ζ имели одинаковые вероятности:

$$\widetilde{F}_j\left(x_j \mid k\right) = \Phi_j\left(z_j \mid k\right),\ \ j = \overline{1,m}\,. \tag{3}$$

Это требование (правило (3)) приведёт к однозначному нелинейному преобразованию наблюдений *x*, в общем случае негауссовых, в эквивалентные в вероятностном смысле гауссовы наблюдения *z*.

Потребуем дополнительно, чтобы эти нормально распределённые эквивалентные наблюдения разных признаков были совместно-гауссовыми. Поскольку нормальная функция распределения – монотонно возрастающая, ранги наблюдений при преобразовании (3) не изменяются, так что ранговые корреляции компонент вектора (x_1, …, x_m) и соответствующего ему вектора (z_1, …, z_m) будут одинаковыми. Это позволяет найти матрицу коэффициентов ранговых корреляций обучающей выборки и приписать её многомерному нормальному распределению, соответствующему преобразованию (3). Учитывая, что корреляционная матрица Пирсона *R* нормального распределения однозначно связана с ранговой корреляционной матрицей Спирмена R_S известным соотношением $R = 2\sin\left(\pi R_S / 6\right)$ [4], можно получить корреляционную матрицу нормального распределения, соответствующего преобразованию (3), из преобразованной по этому правилу обучающей выборки:

$$\left(\hat{\rho}_k\right)_{jl} = \frac{\left(\hat{R}_k\right)_{jl}}{\sqrt{\hat{\sigma}^2_{jk} \cdot \hat{\sigma}^2_{jl}}},\ \ \hat{R}_k = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}\left(z_{ik} - \hat{\mu}_k\right)^T\left(z_{ik} - \hat{\mu}_k\right), \tag{4}$$

где $z_{ik} = (z_{i1k}, \ldots, z_{imk})$ – вектор-строка преобразования (3) вектор-строки $x_{ik} = (x_{i1k}, \ldots, x_{imk})$ набора признаков *i*-го объекта *k*-го класса обучающей выборки, $\hat{\mu}_k = \left(\hat{\mu}_{1k}, \ldots, \hat{\mu}_{mk}\right)$ – вектор-строка эмпирических оценок математических ожиданий значений признаков, $\hat{\sigma}^2_{jk}$ – эмпирическая оценка дисперсии *j*-го признака, *T* – знак транспонирования.

За многомерное распределение обучающей выборки *k*-го класса примем гауссову копула-функцию [1, 7]:

$$C\left(x_1, \ldots x_m \mid k\right) = \Phi\left(\Phi_1^{-1}\left(\widetilde{F}_1\left(x_1 \mid k\right)\right), \ldots, \Phi_m^{-1}\left(\widetilde{F}_m\left(x_m \mid k\right)\right) \middle| \hat{\mu}_k, \hat{R}_k\right),\ \ k = \overline{1,K}\,. \tag{5}$$

Обозначив $z_k = \left(z_{1k}, \ldots, z_{mk}\right)$, где $z_{jk} = \Phi_j^{-1}\left(\widetilde{F}_j\left(x_j \mid k\right)\right),\ \ j = \overline{1,m}$, запишем (с точностью до несущественного слагаемого) логарифм плотности многомерного распределения величин $z_{1k}, \ldots, z_{mk}$, эквивалентных испытуемой выборке $x_1, \ldots, x_m$, в предположении её принадлежности *k*-му классу:

$$\ln p\left(z_k \mid k\right) = -\frac{1}{2}\ln\det\hat{R}_k - \frac{1}{2}\left(z_k - \hat{\mu}_k\right)\hat{R}_k^{-1}\left(z_k - \hat{\mu}_k\right)^T, \tag{6}$$

где z_k и $\hat{\mu}_k$ – вектор-строки длины *m*, $\hat{R}_k$ – $m \times m$-матрица.

2.1.4. Алгоритм байесовской классификации с обучением

Алгоритм байесовской классификации с обучением состоит из следующих этапов.

Этап обучения. По обучающей выборке $(x_{i1k},...,x_{imk})$, $i=\overline{1,n_k}$, $k=\overline{1,K}$, для каждого класса k и каждого признака a_j находится маргинальная эмпирическая функция распределения (2) и соответствующая ей сглаженная маргинальная эмпирическая функция распределения $\tilde{F}_j(x\,|\,k)$, а также выборочные эмпирические оценки $\hat{\mu}_{jk}$ и $\hat{\sigma}^2_{jk}$ математического ожидания и дисперсии j-го признака. Затем обучающая выборка преобразуется по правилу (3) в нормально распределённую выборку (z_{i1k}, …, z_{imk}) с $\hat{\mu}_{jk}$ и $\hat{\sigma}^2_{jk}$, по которой находится эмпирическая оценка корреляционной матрицы $\hat{R}_k$ для каждого класса k.

Обратим внимание, что объём обучающей выборки для каждого класса должен быть больше числа признаков. При $n_k \le m$ матрица $\hat{R}_k$ становится вырожденной, и классификатор на основе соотношения (6) не может работать. В этом случае следует от исходной системы признаков a_1, …, a_m перейти к системе меньшей размерности (не большей, чем минимальное n_k) с помощью некоторого линейного преобразования (например, используя метод главных компонент [5]). При использовании метода главных компонент обучающие выборки всех классов объединяются вместе. Если их общий объём n будет больше числа m исходных признаков, то по объединённой выборке можно построить эмпирическую оценку корреляционной матрицы исходного набора признаков. Собственные векторы этой матрицы образуют новый набор признаков – некоррелированные главные компоненты, а собственные числа будут их дисперсиями. Однако эти новые признаки внутри каждого класса будут в общем случае коррелированными. Выбрав некоторое небольшое число первых главных компонент (это число не должно превышать минимального объёма обучающей выборки класса), будем рассматривать их как новые исходные признаки, то есть будем работать с ними, как с исходными. Их корреляционные матрицы для всех классов уже не будут вырожденными.

Этап классификации. Наблюдаемый вектор-строка признаков (x_1, …, x_m) испытуемого объекта по правилу (3) преобразуется в нормально распределённые векторы $z_k=(z_{1k},...,z_{mk})$, $k=\overline{1,K}$ (для каждой гипотезы k о классе), по которым вычисляется логарифм функции правдоподобия каждого класса по формуле (6). В качестве класса, к которому принадлежит испытуемый объект, принимается максимально правдоподобный (или апостериорно максимально вероятный) класс $\hat{k}$ в соответствии с решающим правилом (1) для эквивалентных признаков z. Качество классификации может быть оценено F-мерой Ван Ризбергена [6] на основе имитационного моделирования объектов по алгоритму, рассмотренному в работе [7].

2.1.5. Численный пример

В качестве примера рассмотрена задача байесовской классификации по авторскому стилю прозаических текстов русских классиков 19 века. В качестве обучающей выборки использовалось 155 текстов художественных произведений 17 авторов. Набором признаков стиля служили частоты употребления 54 служебных слов русского языка. В табл.1 представлен фрагмент таблицы абсолютных частот появления служебных слов в русской художественной прозе 19 века. На рис.1 представлена гистограмма распределения относительных частот появления служебных слов в русской художественной прозе 19 века. На

рис.2 показана «каменистая осыпь» дисперсий главных компонент набора признаков – относительных частот появления служебных слов в русской художественной прозе 19 века. На рис.3 представлены эмпирические функции распределения двенадцати главных компонент (синие ступенчатые линии), их мажоранты (красные ломаные линии) и сглаженные эмпирические функции распределения (чёрные ломаные линии) этих двенадцати главных компонент. На рис.4 представлены эмпирические (синие ступенчатые линии), сглаженные эмпирические (чёрные ломаные линии) и соответствующие им нормальные (красные линии) функции распределения первых двух главных компонент признаков. На горизонтальной оси рис.4 чёрными точками помечены выборочные значения выборочных значений главных компонент, имеющих в общем случае ненормальное распределение, а красными точками помечены соответствующие им (имеющие такие же значения интегральной функции распределения) уже нормально распределённые величины для этих главных компонент. Построив по ним эмпирическую оценку корреляционной матрицы уже нормальных главных компонент, можно проводить байесовскую классификацию текстов по правилу (1).

Таблица 1

Фрагмент таблицы абсолютных частот появления служебных слов в русской художественной прозе 19 века

Блоки	в	на	с	за	к	по	из	у	от	для	во	без
ГогольНВ - Вечера на хуторе близ Диканьки. Ч1	611	502	272	165	130	147	109	111	91	23	26	22
ГогольНВ - Вечера на хуторе близ Диканьки. Ч2	817	675	437	189	183	177	158	156	101	39	36	49
ГогольНВ - Вий	292	212	135	56	61	78	65	44	22	12	25	10
ГогольНВ - Записки сумасшедшего	138	107	48	34	27	18	27	33	8	8	7	6
ГогольНВ - Повесть о том, как поссорился Иван Иванович с Иваном Никифоровичем	303	207	186	79	84	58	47	74	39	18	12	13
ГогольНВ - Мертвые души. Т1	1945	1189	833	325	351	363	293	360	158	140	143	93
ГогольНВ - Мертвые души. Т2	1868	992	841	317	423	350	257	436	242	152	119	125
ГогольНВ - Невский проспект	269	203	144	41	64	43	34	25	30	15	19	20
ГогольНВ - Нос	207	115	89	35	38	41	22	37	11	14	14	14

ГогольНВ - Портрет	499	267	200	82	84	65	67	82	51	34	31	26
Гог льНВ - Старосветские помещики	158	122	83	39	29	35	20	21	15	14	8	8
ГогольНВ - Тарас Бульба	859	737	386	202	158	204	190	151	113	56	41	39
ГогольНВ - Шинель	261	174	115	47	63	55	32	37	19	25	12	17
ГончаровИА - Обломов	3524	2401	2038	706	812	594	398	879	59	193	105	197
ГончаровИА - Обрыв	5088	3580	2991	1188	1470	821	696	1554	1100	284	204	281
ГончаровИА - Обыкновенная история	1794	1208	1174	422	489	321	236	534	385	180	73	143
ДостоевскийФМ - Бедные люди	761	481	333	158	197	152	111	315	169	73	38	48
ДостоевскийФМ - Белые ночи	332	202	130	78	64	32	31	69	56	26	21	22
ДостоевскийФМ - Бесы	5001	2468	2525	914	1151	840	717	944	655	393	246	303
ДостоевскийФМ - Братья Карамазовы	6833	3628	3062	1238	1736	942	905	1467	1045	563	419	302
ДостоевскийФМ - Вечный муж	1032	79	540	175	265	143	133	171	139	91	64	50
ДостоевскийФМ - Господин Прохарчин	201	126	117	47	53	38	35	43	36	20	14	16
ДостоевскийФМ - Двойник	1018	695	593	210	226	201	115	152	159	57	68	50
ДостоевскийФМ - Дядюшкин сон	862	471	434	240	211	105	149	168	147	86	3	39

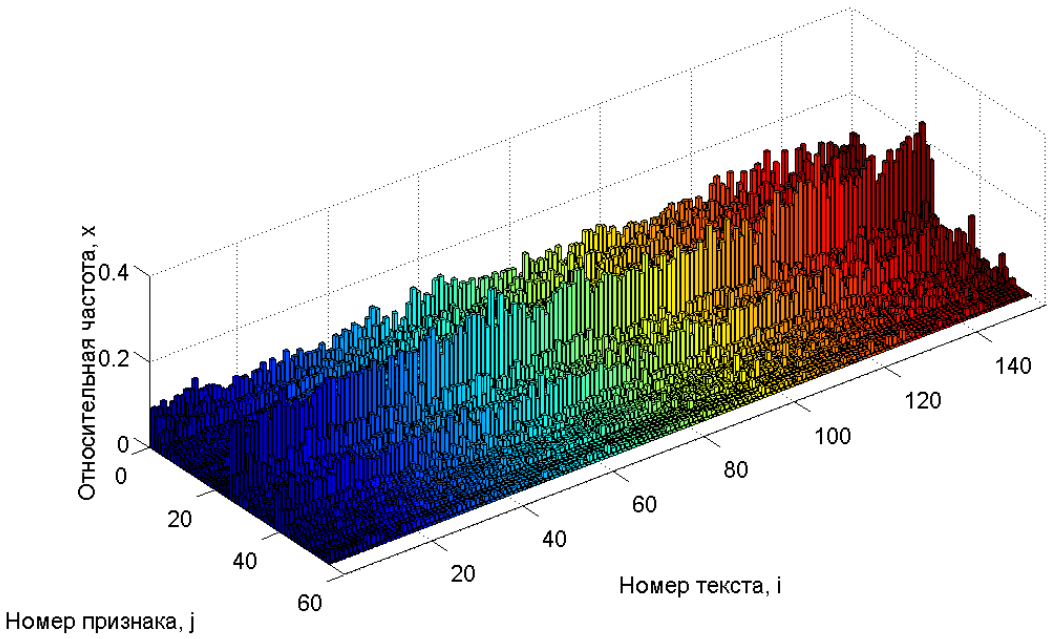


Рис.1. Гистограмма распределения относительных частот появления служебных слов в русской художественной прозе 19 века

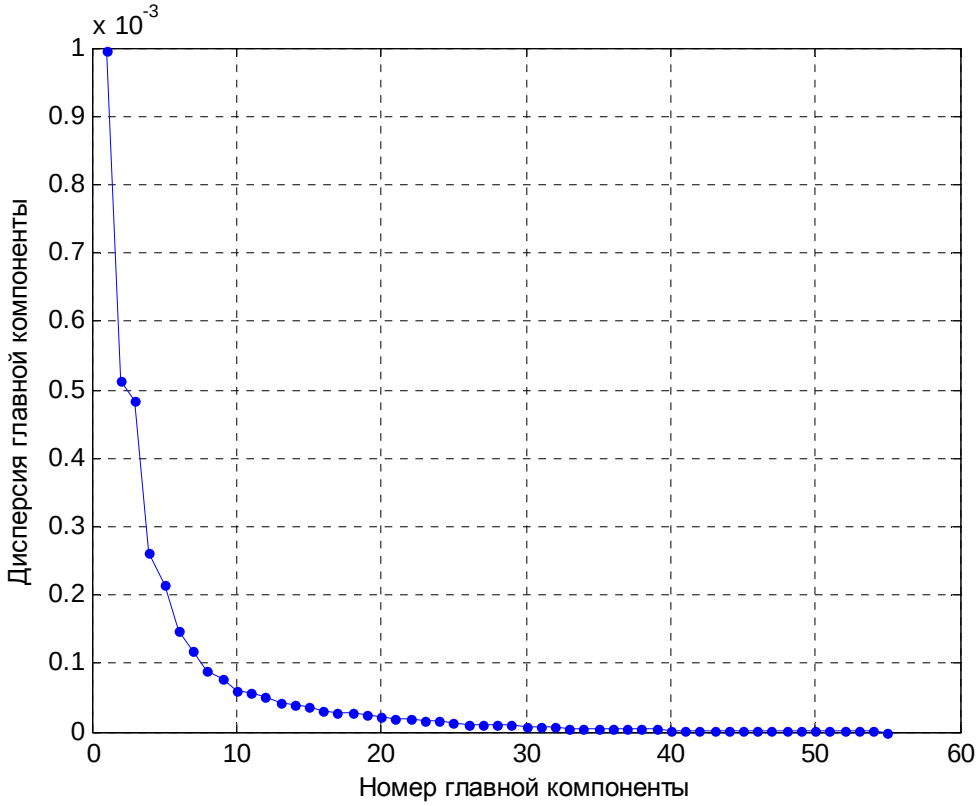


Рис.2. «Каменистая осыпь» дисперсий главных компонент

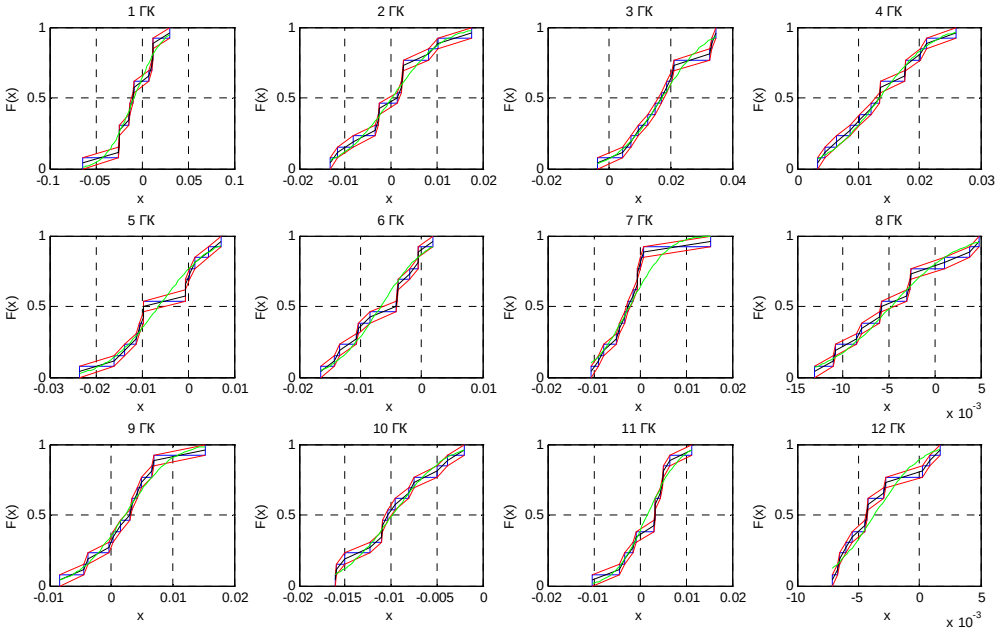


Рис.3. Эмпирические функции распределения (синие ступенчатые), их мажоранты (красные ломаные) и сглаженные эмпирические функции распределения (чёрные ломаные) двенадцати главных компонент

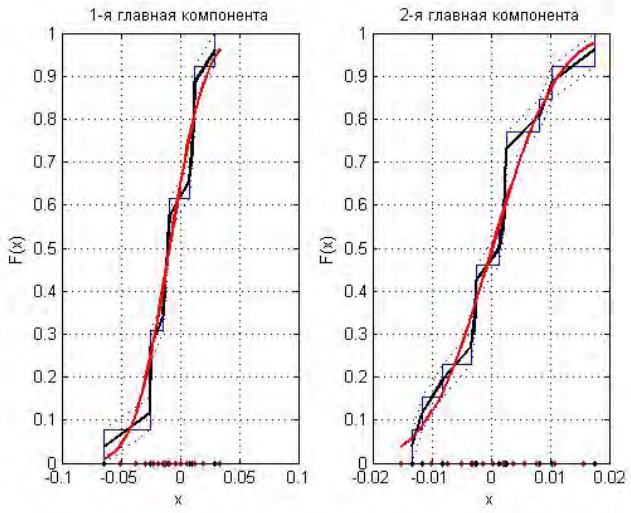


Рис.4. Эмпирические (синие), сглаженные эмпирические (чёрные) и соответствующие нормальные (красные) функции распределения первых двух главных компонент

Материал этого раздела опубликован в работе [8].

Литература к разделу

1. Прикладная статистика: Классификация и снижение размерности: Справочное издание / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин / Под ред. С.А. Айвазяна. – М.: Финансы и статистика, 1989. – 608 с.

2. McCallum, A., Nigam, K. A Comparison of Event Models for Naïve Bayes Text Classification. In AAAI/ICML-98 Workshop of Learning for Text Categorization, pp. 41–48. Technical Report WS-98-05. AAAI Press, 1998.

3. Nelsen, R.B. An Introduction to Copulas (Second Edition). – Berlin: Springer, 2006.

4. Ван дер Варден, Б.Л. Математическая статистика. – М.: Иностранная Литература, 1960. – 636 с.

5. Афифи, А., Эйзен, С. Статистический анализ: Подход с использованием ЭВМ. / Пер. с англ. – М.: Мир, 1982. – 488 с.

6. Van Rijsbergen, C.J. Information Retrieval. – London: Butterworths, 1979.

7. Поддубный, В.В., Пехтерев, А.С. Копулы сглаженных эмпирических распределений при наличии связок (совпадений) и их применение в имитационном моделировании // Труды XII Международной ФАМЭБ'2013 конференции. / Под ред. Олега Воробьёва. – Красноярск: НИИППБ, СФУ, 2013. – С. 312–321.

8. Кубарев А.И., Поддубный В.В. Байесовская классификация с обучением на основе использования копула-функций // Информационные технологии и математическое моделирование (ИТММ-2013): Материалы XII Всероссийской научно-практической конференции с международным участием им. А. Ф. Терпугова (29-30 ноября 2013 г.). – Томск: Изд-во Том. ун-та, 2013. – Ч.2. – С.126–130.

2.2. ПОТОКОВАЯ КЛАССИФИКАЦИЯ ТЕКСТОВ НА ОСНОВЕ С-МЕРЫ

2.2.1. Введение

Задачи автоматической классификации текстов возникают как при создании и совершенствовании поисковых систем, так и в научных исследованиях, связанных с выявлением стилевых особенностей художественных или публицистических произведений, определением авторства текстов и т.п.

Формальное определение задачи классификации текстов приведено, например, в [1]. Для решения данной задачи уже разработано множество методов, однако многие авторы (например, [3, 6, 7]) разделяют все эти методы на признаковые и потоковые. Признаковые методы (feature-based approaches) работают с текстами не напрямую, а лишь с их численными представлениями – векторами значений признаков. Для того, чтобы применить любой признаковый метод классификации для обработки текстов, необходимо, прежде всего, определить достаточный набор признаков, по которому будет проходить классификация. В случае неудачного выбора такого набора результат классификации может оказаться крайне неудовлетворительным.

В отличие от признаковых методов, потоковые методы (stream-based approaches) не требуют задания каких-либо признаков для классификации текстов. Они непосредственно используют элементы текста. Элементы текста обычно представлены в виде некоторой структуры для облегчения доступа к этим элементам и ускорения их обработки, так как текст X рассматривается как последовательность (поток) из n элементов $x_1, x_2, ..., x_n$ некоторого алфавита Q, при этом длина текстовой строки $n = |X|$. В качестве элемента текста x_n может быть выбран одиночный текстовый символ, слово, грамматический класс, любая группировка символов текста. В случае какой-либо группировки символов перед обработкой текста проводится его преобразование в новый вид – последовательность грамматических кодов.

Целью данного направления работы является исследование одного из методов потоковой классификации на примере текстов художественной литературы.

2.2.2. Потоковые методы классификации на основе C- и R-мер

В потоковых методах О.Г. Шевелев [1] выделяет два основных направления:

- подсчет повторений строк (R-, C- и другие меры);
- сжатие информации (off-the-shelf, PPM).

Первое направление видится как наиболее перспективное, так как позволяет наглядно анализировать характеристики исследуемого текста в момент классификации. Потоковые методы классификации используют вычисления меры близости между двумя текстами. Первоначально для использования этих методов производится конкатенация всех текстов класса, участвующих в обучении. Так формируются супертексты классов. Сам же процесс классификации происходит путем вычислением мер близости между исследуемым текстом и супертекстами каждого класса. Класс, имеющий наибольшее значение меры, и будет искомым классом.

Любая мера близости текста и супертекста является результатом подсчета определенных подстрок исследуемого текста, которые есть в супертексте.

Первоначально (по-видимому, В.Дж. Тианом и Д. Томасом в их неопубликованной работе, как указано в [1]) была предложена C-мера, которая подсчитывает только подстроки определенной длины:

$$c_k(X \mid S) = \sum_{i=k}^{n} c(x_{i-k+1} \ldots x_n \mid S),$$

где X – исследуемый текст, S – супертекст класса, k – некоторая длина подстроки поиска, $n = |X|$,

$$c(x_{i-k+1} \ldots x_n \mid S) = \begin{cases} 1, & x_{i-k+1} \ldots x_n \subset S \\ 0, & x_{i-k+1} \ldots x_n \not\subset S \end{cases}.$$

Чтобы убрать зависимость меры от длины текста, вводится нормализация для C-меры. Нормализованная C-мера представляется следующим образом:

$$c_k^{norm} = \frac{c_k(X \mid S)}{n-k}.$$

Д.В. Хмелёв [2, 4] предложил в качестве развития C-меры R-меру. В свою очередь она учитывает все возможные повторения всех подстрок одного текста в другом:

$$r(X \mid S) = \sum_{j=1}^{n} c_j(X \mid S).$$

Вычисления R- и C-мер сводятся к поиску подстрок текста X в супертексте S. Для C-меры и небольших строк такой поиск можно осуществлять прямым просмотром, во всех остальных случаях прямой просмотр будет слишком трудоемкой процедурой. Для ускорения подсчета можно использовать суффиксные массивы или суффиксные деревья. Построение как суффиксных деревьев [6], так и суффиксных массивов [7] происходит за линейное время. Преимущество деревьев в том, что поиск по ним будет в целом работать быстрее. Преимущество массивов состоит в том, что для их получения нужно меньше времени и памяти, их можно подготовить заранее и хранить на диске [7].

2.2.3. Исследование потокового классификатора, основанного на C-мере

Для первоначального анализа потоковых методов нами была выбрана классификация на основе С-меры, а для работы со строками – суффиксные массивы. В качестве классов для классификации выступили авторы русской прозы 19 в., исследуемые в работах [8, 9]. Всего в нашей работе использовалось 12 авторов, 156 произведений которых были выбраны в качестве «обучающей» выборки для построения соответствующих супертекстов, а несколько произведений тех же авторов использовались в качестве тестовой выборки.

Для проведения тестирования метода был спроектирован программный модуль, позволяющий проводить классификацию текстов на основе С-меры. Открытым оставался вопрос выбора длины подстроки поиска. Сами авторы С-меры отмечали [2], что данная мера неплохо работает лишь на подстроках длиной от 5 до 13 символов. Стоит отметить тот факт, что эти авторы проводили исследование метода, используя газетные статьи. Мы выбрали для анализа работы метода тексты художественных произведений и подстроки длины от 1 символа до 25 символов. Результаты классификации для текста, представленного в обучающей выборке, показаны на диаграмме рис.1 (для наглядности на диаграмме проиллюстрированы лишь пять классов):

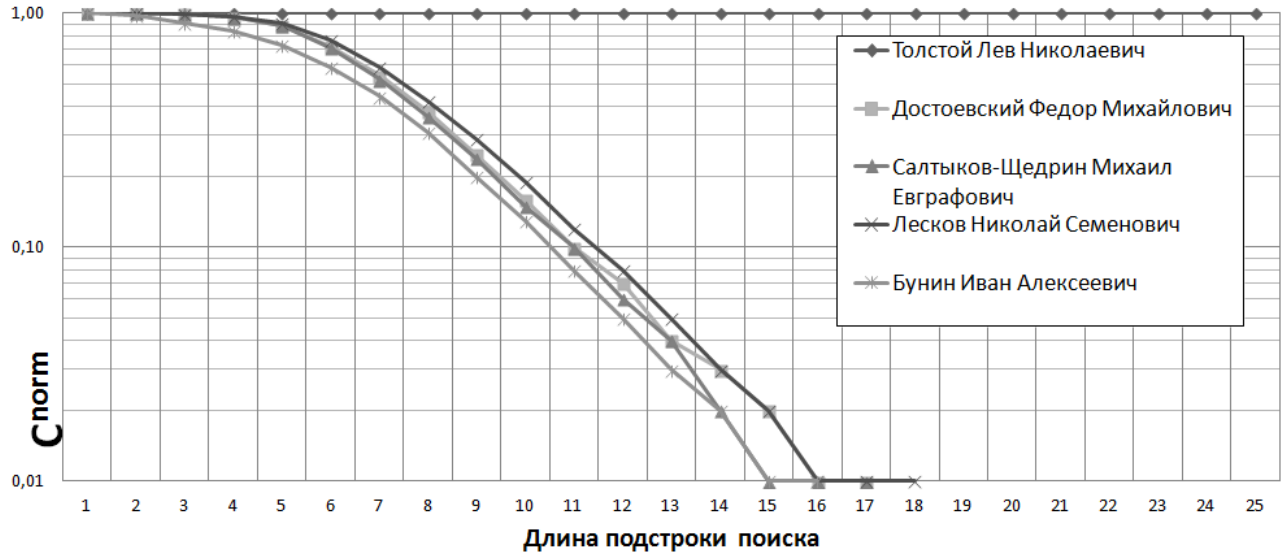


Рис. 1. Диаграмма значений С-меры для текста Л.Н. Толстого «Детство»

На диаграмме отлично видно, что какая бы длина подстроки ни была выбрана, С-мера дает полное совпадение с супертекстом своего класса, а, следовательно, идеальную принадлежность к своему классу.

Для текста же из тестовой выборки результаты представлены на диаграмме рис.2.

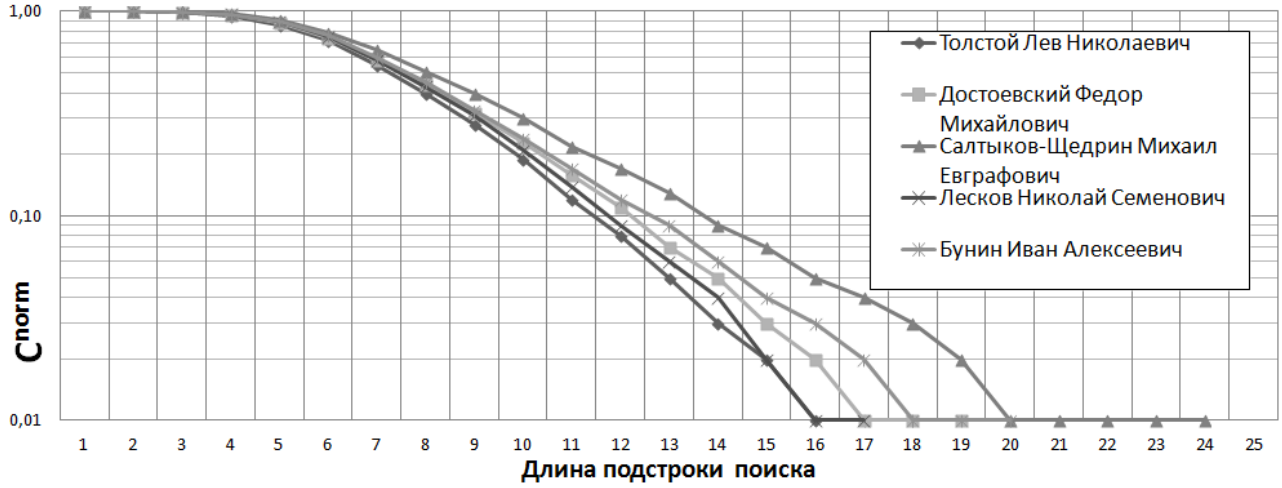


Рис. 2. Диаграмма значений С-меры для текста М.Е. Салтыкова-Щедрина «Пошехонская старина»

На диаграмме видно, что существенное различие в значениях С-меры наблюдается уже в подстроках длиной в 3 символа, а при достижении подстроки поиска в 24 символа различие значений становится крайне малым. При длине подстроки в 20 символов С-мера дает ошибку в определении класса, но по другим значимым параметрам дает правильный результат. Таким образом, С-мера при определенных параметрах длины строки может давать ошибки, что также отмечается в работе [2]. Для того, чтобы избежать ошибки такого типа, Д.В. Хмелёв и предложил использование R-меры.

В целом на художественных произведениях метод С-меры показал себя неплохо. Рабочим диапазоном длин для классификации методом С-меры стали длины подстроки от 3 до 20 символов.

Следующим этапом видится использование метода классификации на основе R-меры, являющегося развитием метода на основе С-меры. Однако использование данного метода сопряжено с определенной технической трудностью. Время, затрачиваемое для подсчета С-меры текста длиной порядка 500 тыс. символов (текст среднего объема художественного произведения), составляет порядка 2–3 сек для одного класса. Так как R-мера является суммой С-мер всех длин подстрок исследуемого текста, то, перемножая время подсчета одной С-меры на длину текста, получим общее время выполнения в наихудшем случае порядка 10 дней, что, естественно, неприемлемо для классификации в реальном времени.

2.2.4. Арбитражный метод классификации на основе С-меры

В качестве развития метода С-меры нами предлагается метод классификации, который использует С-меру на одном из его этапов. На первом этапе для каждого класса строится вектор значений С-мер длин подстрок от 3 до 25 символов: $c_{i,3}^{norm}$, $c_{i,4}^{norm}$, …, $c_{i,25}^{norm}$. После чего для каждого k выбираем класс-победитель, у которого значение С-меры максимально:

$$Vic_k = \max c_{i,k}^{norm}.$$

На заключительном этапе выбирается финальный победитель – класс, к которому и относится исследуемый текст – методом большинства по всем *k*.

Данный метод улучшает качество классификации текстов, так как чаще всего С-мера определяет истинный класс верно. Но порядка 20% тестовых текстов классифицируется ошибочно. Это связано с тем, что нередко значения С-меры сразу у нескольких классов весьма близки на длинах подстроки от 3 до 17 символов, и только при увеличении длины подстроки с 17 символов начинает все больше доминировать значение С-меры для истинного класса.

Материал этого раздела опубликован в работе [10].

Литература к разделу

1. Шевелев О.Г. Методы автоматической классификации текстов на естественном языке: Учебное пособие. Томск: ТМЛ-Пресс, 2007. - 144с.

2. Khmelev D.V., Teahan W.J. Verification of text collections for text categorization and natural language processing // Technical Report AIIA 03.1. School of Informatics, University of Wales. Bangor, 2003.

3. Humnisett D. and Teahan W.J. Context-based methods for text categorization // Proceedings of the 27th Annual International ACM SIGIR Conference (SIGIR), The University of Sheffield, UK, 2004.

4. Хмелев Д.В. Классификация и разметка текстов с использованием методов сжатия данных. Краткое введение. 2003. - [Электронный ресурс] // URL: http://compression.graphicon.ru/download/articles/classif/intro.html.

5. Shevelyov O.G., Poddubnyj V.V. Complex investigation of texts with the system "StyleAnalyzer" // Text and Lanquage / Ed. by P. Grzyber, E. Kelih, J. Macutek. - Wien: Praesens Verlag, 2010. - P. 207 - 212. - ISBN 978-3-7069-0625-8

6. Ukkonen E. Constructing Suffix-trees On-Line in Linear Time // Algorithms, Software, Architecture: Information Processing. - 1992. - № 1(92). - P. 484-92.

7. Kärkkäinen J. and Sanders P. Simple linear work suffix array construction // J.C.M. Baeten et al. (Eds.): ICALP 2003, LNCS 2719, pp. 943–955, 2003.

8. Поддубный В.В. Шевелев О.Г. Кравцова А.С. Фатыхов А.А. Словарно-аналитический блок системы "Стилеанализатор" // Научное творчество молодежи : Материалы XIV Всероссийской научн. - практ. конф. (15-16 апреля 2010 г.). - Томск : Изд-во Том. ун-та, 2010. - Ч. 1. - С. 138 – 140

9. Шевелев О.Г. Разработка и исследование алгоритмов сравнения стилей текстовых произведений: Автореф. дис. … канд. техн. наук / Том. гос. ун-т. – Томск, 2006. – 19 с.

10. Ашуров М.Ф., Поддубный В.В. Потоковый метод классификации текстов художественной литературы на основе С-меры // Информационные технологии и математическое моделирование (ИТММ-2013): Материалы XII Всероссийской научно-практической конференции с международным участием им. А. Ф. Терпугова (29-30 ноября 2013 г.). – Томск: Изд-во Том. ун-та, 2013. – Ч.2. – С.85–89.

3. НОВЫЕ ИССЛЕДОВАТЕЛЬСКИЕ РЕЗУЛЬТАТЫ, ПОЛУЧЕННЫЕ С ПОМОЩЬЮ ТЕКСТО-АНАЛИТИЧЕСКОЙ СИСТЕМЫ «СТИЛЕАНАЛИЗАТОР 2.0. WEB» И НА ОСНОВЕ ПРИМЕНЕНИЯ НЕИСПОЛЬЗОВАВШИХСЯ РАНЕЕ ПРИЗНАКОВ В 2013 ГОДУ

3.1. Анализ сочетания алгоритмов статистического и логического анализа при классификации текстов.

Была проведена серия экспериментов по проверке разработанного метода классификации текстов на основе частотных порогов, определяемых с помощью дерева решений. Применялась методика «просеивания» выборки на себя саму. Результаты «просеивания» сопоставлялись с другими методами - кластеризацией (на основе сравнения отдельных частот, метод дальнего соседа) и классификацией на основе статистических и информационных мер (Хи-квадрат, обучение с исключением одного).

Ниже представлены сопоставительные результаты определения авторства текста на основе этих трех методов. Материал - выборка 13 классиков, набор признаков - служебные слова (76 единиц).

Кластеризация	F-мера
Общая	55,15
БунинИА	52,94
ГогольНВ	40
ГончаровИА	72,73
ДостоевскийФМ	60,47
КупринАИ	66,67
ЛермонтовМЮ	50
ЛесковНС	54,55
ПушкинАС	75
СалтЩедрМЕ	75
ТолстойЛН	80
ТургеневИС	50,91
ЧернышНГ	100
ЧеховАП	42,11

Классификация

Блоки	Полнота	Точность	F-мера
БунинИА	0,8000	0,8889	0,8421
ГогольНВ	1,0000	0,8667	0,9286
ГончаровИА	1,0000	1,0000	1,0000
ДостоевскийФМ	0,8077	0,7778	0,7925
КупринАИ	1,0000	0,2222	0,3636
ЛермонтовМЮ	1,0000	0,8333	0,9091
ЛесковНС	0,7143	0,5556	0,6250
ПушкинАС	1,0000	0,7143	0,8333
СалтЩедрМЕ	1,0000	0,5556	0,7143
ТолстойЛН	0,8750	1,0000	0,9333
ТургеневИС	0,8500	0,8947	0,8718
ЧернышНГ	1,0000	0,4286	0,6000
ЧеховАП	0,3714	0,8667	0,5200
Итого	0,8783	0,7388	0,7641

Всего 45 текстов из 180 распознано неверно.
«Просеивание» через пороги частот, определенные с помощью деревьев решений.

Автор	*Свой автор*	*Два автора*	*Три автора*
Бунин		2	18
Гоголь	5	7	1
ГончаровИА	2	3	1
ДостоевскийФМ	18	8	
КупринАИ		2	
ЛермонтовМЮ	3	2	
ЛесковНС	7		
ПушкинАС		3	7
СалтЩедрМЕ	5		
ТолстойЛН	5	3	
ТургеневИС	18	22	

ЧернышНГ	1	1	1
ЧеховАП	20	12	3
Всего текстов	**84**	**65**	**31**

При «просеивании» выборки саму на себя исключено непопадание автора в число кандидатов. Поэтому критерием оценки здесь является к-во однозначно распознанных текстов, отсутствие текстов, которым приписано более трех кандидатов и пр.

На первый взгляд, классификация с исключением одного дает более точные результаты: однозначно правильно распознаны 135 текстов (см. выше). Однако у этого метода есть ряд существенных недостатков по сравнению с «просеиванием» текста через частотные пороги: он обязательно приписывает тексту автора; настоящий автор может быть определен неверно; неизвестен вклад каждого конкретного признака в принятие решения; авторский стиль не получает качественной и количественной характеристики, т.е. остается не описан. При просеивании выборки саму на себя эти недостатки устраняются, и появляется возможность получить описание стиля автора. Ср. ниже описание устойчивых признаков (восклиц. знаки помечают уникальные признаки):

Автор:БунинИА
все ниже :бы/<0,00280142817907168>=
все ниже :б&бы/<0,00280373831775701>=
все ниже :даже/<0,00195848021934978>=
все ниже :иль/<8,20442136267234E-06>=
все ниже :коли/<3,33533453405377E-05>=
все ниже :коль/<2,80051641522697E-06>=
все ниже :коли&коль/<3,33533453405377E-05>=
!все ниже :о&об&обо/<0,00277824408259191>=
все ниже :сколь/<2,80051641522697E-06>=
!все ниже :также/<5,41711809317443E-05>=
все ниже :у/<0,00445831475702185>=
!все ниже :хоть&хотя/<0,00046946834767527>=
все ниже :чтоб/<0,00028631964725 4195>=
Автор:ГогольНВ
все выше :бы/<0,00280142817907168>=
все выше :б&бы/<0,00280373831775701>=
все выше :ж/<0,00021039343572 4805>=
все выше :если/<0,00099928312297 6995>=
все выше :как/<0,00572501173158142>=
все ниже :коль/<2,80051641522697E-06>=
все выше :если&коли&коль/<0,00112839410949708>=
все выше :не/<0,0113156885308784>=
!все ниже :от&ото/<0,00268326714607706>=
все выше :перед&передо/<0,00057483063026 0727>=
все ниже :под&подо/<0,00149250374812 59>=
все выше :при/<0,00044888340253 6191>=
все ниже :столь/<9,61107195489204E-05>=
все выше :также/<5,41711809317443E-05>=
все выше :только/<0,00246519721577726>=
все ниже :у/<0,00445831475702185>=
!все выше :уже/<0,00221729490022173>=
все выше :хотя/<8,65975810409029E-05>=
все выше :хоть/<6,51706384550214E-05>=
все выше :хоть&хотя/<0,00046694683476 7527>=
все ниже :через/<0,00100315276583548>=
все выше :чтобы/<0,00099695626232034>=
Автор:ГончаровИА
все выше :будто/<0,00037258199321 2966>=
все ниже :в&во/<0,0204980249692228>=
все выше :вдруг/<0,00063371356147 0215>=
все ниже :ведь/<0,00104661460429917>=
все выше :вон/<0,00016287676562 9383>=
все выше :вот/<0,00147928994082 84>=
все ниже :даже/<0,00195848021934978>=
все выше :до/<0,00131790226315 63>=
все ниже :же/<0,00371068425017573>=
все выше :ж/<0,00021039343572 4805>=
все ниже :за/<0,00405530941832214>=
все выше :зато/<1,32787135582305E-05>=
все выше :если/<0,00099928312297 6995>=
все ниже :и/<0,03456012595945 68>=
все ниже :иль/<8,20442136267234E-06>=
все ниже :к&ко/<0,00490825923832 04>=
все выше :как/<0,00572501173158142>=
все ниже :коль/<2,80051641522697E-06>=
все выше :зато/<1,32787135582305E-05>=
все выше :кроме/<1,57025312480372E-05>=
все выше :ли/<0,00137709433096167>=
все ниже :ль/<2,89809470973511E-05>=
все выше :ли&ль/<0,00139529891596007>=
все выше :лишь/<1,60521373420871E-05>=
все ниже :на/<0,01330376940133 04>=
все выше :над/<9,79240109674892E-05>=
все выше :не/<0,01131568853087 84>=
!все ниже :но/<0,00428756431346 47>=
все ниже :под&подо/<0,00149250374812 59>=
все ниже :сколь/<2,80051641522697E-06>=
все ниже :столь/<9,61107195489204E-05>=
все выше :также/<5,41711809317443E-05>=
все выше :то_есть/<6,03536725209729E-05>=
все выше :как_бы/<6,40738130326136E-05>=
все выше :тоже/<0,00016498927569708>=
все ниже :уже/<0,00221729490022173>=
все выше :хотя/<8,65975810409029E-05>=
все выше :хоть/<6,51706384550214E-05>=
все выше :хоть&хотя/<0,00046946834767527>=
все выше :хоть_бы/<5,53412618914536E-06>=
все ниже :через/<0,00100315276583548>=
все ниже :чтобы/<0,00099695626232034>=
все выше :чтоб/<0,00028631964725 4195>=
Автор:ДостоевскийФМ
все выше :будто/<0,00037258199321 2966>=
все ниже :вон/<0,00016287676562 9383>=
все выше :ж/<0,00021039343572 4805>=
все выше :коли&коль/<3,33533453405377E-05>=
все ниже :на/<0,01330376940133 04>=
все выше :над/<9,79240109674892E-05>=
все ниже :под&подо/<0,00149250374812 59>=
все выше :про/<0,00015573898146 7061>=
все выше :то_есть/<6,03536725209729E-05>=
все выше :тоже/<0,00016498927569708>=
все выше :уж/<0,00084530853761623>=
все выше :хотя/<8,65975810409029E-05>=
все выше :хоть/<6,51706384550214E-05>=
все выше :хоть&хотя/<0,00046946834767527>=
все ниже :через/<0,00100315276583548>=
все выше :чтоб/<0,00028631964725 4195>=

Автор:КупринАИ
все ниже :без&безо/<0,000850665645867892>=
все ниже :бы/<0,00280142817907168>=
все ниже :б/<0,000214732780895119>=
все ниже :б&бы/<0,00280373831775701>=
все ниже :в&во/<0,0204980249692228>=
все выше :вдруг/<0,000633713561470215>=
все ниже :вон/<0,000162876765629383>=
все выше :вот/<0,0014792899408284>=
все выше :вроде/<1,16188743634517E-05>=
все ниже :да/<0,00268273790406378>=
все ниже :даже/<0,00195848021934978>=
все выше :для/<0,00117907674717736>=
все выше :еще/<0,00212754916714234>=
все ниже :же/<0,00371068425017573>=
!все ниже :ж/<0,000210393435724805>=
все ниже :за/<0,00405530941832214>=
все выше :зато/<1,32787135582305E-05>=
все выше :если/<0,000999283122976995>=
все выше :и/<0,0345601259594568>=
все выше :из&изо/<0,00236173981499705>=
все выше :из-за/<3,20780137293899E-05>=
все выше :или/<0,00165612767238783>=
все ниже :иль/<8,20442136267234E-06>=
все выше :или&иль/<0,00165612767238783>=
все ниже :к&ко/<0,00490825923832 04>=
!все ниже :как/<0,00572501173158142>=
все ниже :коль/<2,80051641522697E-06>=
все выше :если&коли&коль/<0,00112839410949708>=
все выше :зато/<1,32787135582305E-05>=
все выше :кроме/<1,57025312480372E-05>=
все ниже :ль/<2,89809470973511E-05>=
все выше :либо/<5,53412618914536E-06>=
все выше :лишь/<1,60521373420871E-05>=
все ниже :на/<0,0133037694013304>=
все выше :над/<9,79240109674892E-05>=
все выше :не/<0,0113156885308784>=
все выше :но/<0,0042875643134647>=
все выше :перед&передо/<0,00057483063026 0727>=
все выше :по/<0,00318593792914035>=
все ниже :под&подо/<0,00149925037481259>=
все выше :при/<0,000448883402536191>=
все ниже :столь/<9,61107195489204E-05>=
все выше :также/<5,41711809317443E-05>=
все выше :как_бы/<6,40738130326136E-05>=
все выше :тоже/<0,000164989275697 08>=
все ниже :уж/<0,00084530853761623>=
все ниже :уже/<0,00221729490022173>=
все выше :хотя/<8,65975810409029E-05>=
все выше :хоть/<6,51706384550214E-05>=
все выше :хоть&хотя/<0,000466946834767527>=
все выше :хотя_бы/<1,65109137139649E-05>=
все выше :хоть_бы/<5,53412618914536E-06>=
все ниже :через/<0,00100315276583548>=
!все ниже :что/<0,0114020148370522>=
все выше :чтобы/<0,00099695626232034>=
все ниже :чтоб/<0,000286319647254195>=
все ниже :чтобы&чтоб/<0,00113442994895065>=
Автор:ЛермонтовМЮ
все выше :б/<0,000214732780895119>=
все ниже :ведь/<0,00104661460429917>=
все ниже :вон/<0,000162876765629383>=
все ниже :да/<0,00268273790406378>=
все ниже :даже/<0,00195848021934978>=
все ниже :еще/<0,00212754916714234>=
все выше :если/<0,000999283122976995>=
все выше :из&изо/<0,00236173981499705>=
все ниже :к&ко/<0,00490825923832 04>=
все ниже :коль/<2,80051641522697E-06>=
все выше :если&коли&коль/<0,00112839410949708>=
все ниже :с&со/<0,0101210663246303>=
все ниже :сколь/<2,80051641522697E-06>=
все ниже :столь/<9,61107195489204E-05>=
все ниже :тоже/<0,00016498927569708>=
все ниже :у/<0,00445831475702185>=
все выше :хотя/<8,65975810409029E-05>=
все выше :хоть&хотя/<0,000466946834767527>=
все ниже :через/<0,00100315276583548>=
все выше :чтоб/<0,000286319647254195>=

Автор:ЛесковНС
!все выше :а/<0,00564295485636115>=
все ниже :бы/<0,00280142817907168>=
все ниже :б/<0,000214732780895119>=
все ниже :б&бы/<0,00280373831775701>=
все выше :вон/<0,000162876765629383>=
все выше :вроде/<1,16188743634517E-05>=
все ниже :даже/<0,00195848021934978>=
!все выше :за/<0,00405530941832214>=
все выше :зато/<1,32787135582305E-05>=
все выше :из-за/<3,20780137293899E-05>=
все ниже :коли/<3,33533453405377E-05>=
все ниже :коли&коль/<3,33533453405377E-05>=
все выше :зато/<1,32787135582305E-05>=
все выше :кроме/<1,57025312480372E-05>=
все ниже :ли/<0,00137709433096167>=
все ниже :ль/<2,89809470973511E-05>=
все ниже :ли&ль/<0,00139529891596007>=
все выше :лишь/<1,60521373420871E-05>=
все ниже :на/<0,0133037694013304>=
все выше :над/<9,79240109674892E-05>=
все выше :не/<0,0113156885308784>=
все ниже :под&подо/<0,00149925037481259>=
все выше :при/<0,000448883402536191>=
все выше :про/<0,000155738981467061>=
все выше :как_бы/<6,40738130326136E-05>=
все выше :тоже/<0,00016498927569708>=
все ниже :уже/<0,00221729490022173>=
все выше :хотя/<8,65975810409029E-05>=
все выше :хоть/<6,51706384550214E-05>=
все выше :хоть&хотя/<0,000466946834767527>=
все ниже :через/<0,00100315276583548>=
все выше :чтобы&чтоб/<0,00113442994895065>=
Автор:ПушкинАС
все ниже :бы/<0,00280142817907168>=
все ниже :б&бы/<0,00280373831775701>=
все ниже :ведь/<0,00104661460429917>=
все ниже :вот/<0,0014792899408284>=
!все ниже :вроде/<1,16188743634517E-05>=
все ниже :даже/<0,00195848021934978>=
все выше :ж/<0,000210393435724805>=
все ниже :либо/<5,53412618914536E-06>=
!все ниже :лишь/<1,60521373420871E-05>=
все выше :над/<9,79240109674892E-05>=
все выше :не/<0,0113156885308784>=
все ниже :под&подо/<0,00149925037481259>=
все ниже :сколь/<2,80051641522697E-06>=
!все ниже :то/<0,00257953568357696>=
все ниже :тоже/<0,00016498927569708>=
!все ниже :только/<0,00246519721577726>=
все ниже :чтобы/<0,00099695626232034>=
Автор:СалтЩедрМЕ
все выше :без&безо/<0,000850665645867892>=
все выше :б/<0,000214732780895119>=
!все выше :в&во/<0,0204980249692228>=
все выше :вроде/<1,16188743634517E-05>=
!все выше :даже/<0,00195848021934978>=
все выше :для/<0,00117907674717736>=
все выше :до/<0,00131790226315 63>=
все ниже :еще/<0,00212754916714234>=
все ниже :же/<0,00371068425017573>=
все выше :ж/<0,000210393435724805>=
все ниже :за/<0,00405530941832214>=
все выше :зато/<1,32787135582305E-05>=
все ниже :и/<0,0345601259594568>=
все выше :из-под&из-подо/<1,65605411984864E-05>=
все ниже :к&ко/<0,00490825923832 04>=
все выше :как/<0,00572501173158142>=
!все выше :коль/<2,80051641522697E-06>=
все выше :коли&коль/<3,33533453405377E-05>=
все выше :зато/<1,32787135582305E-05>=
все выше :кроме/<1,57025312480372E-05>=
все выше :ли/<0,00137709433096167>=
все ниже :ль/<2,89809470973511E-05>=
все выше :ли&ль/<0,00139529891596007>=
все выше :либо/<5,53412618914536E-06>=
все выше :лишь/<1,60521373420871E-05>=
все ниже :на/<0,0133037694013304>=
все выше :над/<9,79240109674892E-05>=

все выше :не/<0,0113156885308784>=
!все выше :ни/<0,00164502467537013>=
все выше :но/<0,0042875643134647>=
все выше :о&об&обо/<0,00277824408259191>=
все выше :перед&передо/<0,00057483063026072 7>=
все выше :по/<0,00318593792914035>=
все ниже :под&подо/<0,00149925037481259>=
все выше :при/<0,000448883402536191>=
все ниже :с&со/<0,0101210663246303>=
все выше :то/<0,00257953568357696>=
все выше :то_есть/<6,03536725209729E-05>=
все выше :как_бы/<6,40738130326136E-05>=
все выше :тоже/<0,00016498927569708>=
все выше :только/<0,00246519721577726>=
все ниже :уже/<0,00221729490022173>=
все выше :хотя/<8,65975810409029E-05>=
все выше :хоть/<6,51706384550214E-05>=
все выше :хоть&хотя/<0,000466946834767527>=
все выше :хоть_бы/<5,53412618914536E-06>=
все ниже :через/<0,00100315276583548>=
все выше :что/<0,0114020148370522>=
все выше :чтоб/<0,000286319647254195>=
все выше :чтобы&чтоб/<0,00113442994895065>=
Автор:ТолстойЛН
!все ниже :а/<0,00564295485636115>=
все ниже :без&безо/<0,000850665645867892>=
все ниже :бы/<0,00280142817907168>=
все ниже :б/<0,000214732780895119>=
все ниже :б&бы/<0,00280373831775701>=
все ниже :ведь/<0,00104661460429917>=
все ниже :вон/<0,000162876765629383>=
все ниже :вот/<0,0014792899408284>=
все ниже :да/<0,00268273790406378>=
все ниже :даже/<0,00195848021934978>=
все выше :для/<0,00117907674717736>=
все выше :ж/<0,000210393435724805>=
все ниже :за/<0,00405530941832214>=
все выше :и/<0,0345601259594568>=
все выше :из-за/<3,20780137293899E-05>=
все выше :из-под&из-подо/<1,65605411984864E-05>=
!все выше :к&ко/<0,0049082592383204>=
все выше :кроме/<1,57025312480372E-05>=
все ниже :ли/<0,00137709433096167>=
все ниже :ль/<2,89809470973511E-05>=
все ниже :ли&ль/<0,00139529891596007>=
все выше :над/<9,79240109674892E-05>=
все выше :не/<0,0113156885308784>=
все выше :но/<0,0042875643134647>=
все выше :о&об&обо/<0,00277824408259191>=
!все ниже :по/<0,00318593792914035>=
все ниже :под&подо/<0,00149925037481259>=
все выше :при/<0,000448883402536191>=
все выше :про/<0,000155738981467061>=
!все выше :с&со/<0,0101210663246303>=
все выше :то/<0,00257953568357696>=
все выше :то_есть/<6,03536725209729E-05>=
все выше :как_бы/<6,40738130326136E-05>=
все выше :тоже/<0,00016498927569708>=
все выше :только/<0,00246519721577726>=
все ниже :у/<0,00445831475702185>=
все ниже :уж/<0,00084530853761623>=
все выше :хотя/<8,65975810409029E-05>=
все выше :хоть/<6,51706384550214E-05>=
все ниже :через/<0,00100315276583548>=
все выше :что/<0,0114020148370522>=
все выше :чтобы/<0,00099695626232034>=
все выше :чтобы&чтоб/<0,00113442994895065>=
Автор:ТургеневИС
все ниже :чтобы&чтоб/<0,00113442994895065>=
Автор:ЧернышНГ
все выше :без&безо/<0,000850665645867892>=
все выше :будто/<0,000372581993212966>=
все выше :бы/<0,00280142817907168>=
все выше :б/<0,000214732780895119>=
все выше :б&бы/<0,00280373831775701>=
все ниже :в&во/<0,0204980249692228>=
!все ниже :вдруг/<0,000633713561470215>=
!все выше :ведь/<0,00104661460429917>=
все ниже :вон/<0,000162876765629383>=
все выше :вот/<0,0014792899408284>=
все выше :вроде/<1,16188743634517E-05>=
!все выше :да/<0,00268273790406378>=
все ниже :даже/<0,00195848021934978>=
все выше :для/<0,00117907674717736>=
все выше :до/<0,0013179022631563>=
все выше :еще/<0,00212754916714234>=
!все выше :же/<0,00371068425017573>=
все выше :ж/<0,000210393435724805>=
все ниже :за/<0,00405530941832214>=
все выше :зато/<1,32787135582305E-05>=
все выше :если/<0,000999283122976995>=
все ниже :и/<0,0345601259594568>=
все выше :из&изо/<0,00236173981499705>=
все выше :из-за/<3,20780137293899E-05>=
все выше :или/<0,00165612767238783>=
все ниже :иль/<8,20442136267234E-06>=
все выше :или&иль/<0,00165612767238783>=
все ниже :к&ко/<0,0049082592383204>=
все выше :как/<0,00572501173158142>=
все ниже :коль/<2,80051641522697E-06>=
все выше :если&коли&коль/<0,00112839410949708>=
все выше :зато/<1,32787135582305E-05>=
все выше :кроме/<1,57025312480372E-05>=
все выше :ли/<0,00137709433096167>=
все ниже :ль/<2,89809470973511E-05>=
все выше :ли&ль/<0,00139529891596007>=
все выше :лишь/<1,60521373420871E-05>=
все ниже :на/<0,0133037694013304>=
все выше :над/<9,79240109674892E-05>=
все выше :не/<0,0113156885308784>=
!все ниже :ни/<0,00164502467537013>=
все выше :но/<0,0042875643134647>=
все выше :о&об&обо/<0,00277824408259191>=
!все выше :от&ото/<0,00268326714607706>=
все ниже :под&подо/<0,00149925037481259>=
все выше :при/<0,000448883402536191>=
все выше :про/<0,000155738981467061>=
все ниже :с&со/<0,0101210663246303>=
все ниже :сколь/<2,80051641522697E-06>=
все ниже :столь/<9,61107195489204E-05>=
все выше :также/<5,41711809317443E-05>=
все выше :то/<0,00257953568357696>=
все выше :то_есть/<6,03536725209729E-05>=
все выше :как_бы/<6,40738130326136E-05>=
все выше :тоже/<0,00016498927569708>=
все выше :только/<0,00246519721577726>=
все выше :уж/<0,00084530853761623>=
все ниже :уже/<0,00221729490022173>=
все выше :хотя/<8,65975810409029E-05>=
все выше :хоть/<6,51706384550214E-05>=
все выше :хоть&хотя/<0,000466946834767527>=
все выше :хотя_бы/<1,65109137139649E-05>=
все выше :хоть_бы/<5,53412618914536E-06>=
все выше :что/<0,0114020148370522>=
все выше :чтобы/<0,00099695626232034>=
все выше :чтобы&чтоб/<0,00113442994895065>=
Автор:ЧеховАП
все ниже :коль/<2,80051641522697E-06>=
все ниже :либо/<5,53412618914536E-06>=
все ниже :сколь/<2,80051641522697E-06>=

Важное преимущество реализованного метода просеивания - возможность непопадания текста, не входящего в выборку, ни в какой класс. При просеивании текстов на пороги, полученные на иной выборке, им может не приписываться ни один автор. Это позволяет установить, что текст не соответствует ни одному из стилей. Ниже в Таблице 1 даны результаты анализа просеивания маленьких рассказов классиков по частотным

порогам, полученным для их романов и повестей. Он показывает, что имеется явная зависимость частот употребления служебных слов в текстах одного автора от жанра и объема текста (см. рассказы Бунина, которые, судя по частотам употребления служебных слов, явно отличаются по стилю от его повестей и романов).

Таблица 1.

Анализ просеивания маленьких рассказов классиков по частотным порогам, полученным для их романов и повестей.

Результат		а	без&безо	будто	бы
	БунинИА	0,00278067431352103	0,00139033715676051	0,00034758428919012 9	0,00034758428919012 9
	БунинИА	0,00249221183800623	0,00062305295950155 8	0	0,00062305295950155 8
	БунинИА	0,0050761421319797	0	0,00126903553299492	0,00084602368866328 3
	БунинИА	0,00352422907488987	0,00088105726872246 7	0	0,0014684287812041 1
	БунинИА	0,00370096225018505	0,00074019245003701	0,00074019245003701	0,00074019245003701
	БунинИА	0,00448807854137447	0,00028050490883590 5	0,00056100981767180 9	0
	БунинИА	0,00383141762452107	0,00051085568326947 6	0,00025542784163473 8	0,00051085568326947 6
	БунинИА	0,00665188470066519	0	0	0
	БунинИА	0,00518403317781234	0,00207361327112494	0	0,00207361327112494
ПушкинАС	БунинИА	0,00557103064066852	0	0	0,00278551532033426
	БунинИА	0,00113122171945701	0	0	0,00113122171945701
	БунинИА	0,00887409872434831	0,00221852468108708	0	0,00277315585135885
	БунинИА	0,00391389432485323	0,00065231572080887 1	0,00065231572080887 1	0
	БунинИА	0,00561797752808989	0,00084269662921348 3	0	0,00084269662921348 3
	БунинИА	0,0198412698412698	0	0	0
	БунинИА	0,00237529691211401	0,00158353127474268	0	0
	ЛермонтовМЮ	0,00398633257403189	0,00028473804100227 8	0	0,00142369020501139
	ЛермонтовМЮ	0,0065402223675605	0	0	0,0026160889470242
ПушкинАС	ПушкинАС	0,00451467268623025	0,00075244544770504 1	0	0
	ТургеневИС	0,00462504129501156	0,00066072018500165 2	0,00033036009250082 6	0,00165180046250413

3.2. Оптимизация признакового пространства текстов

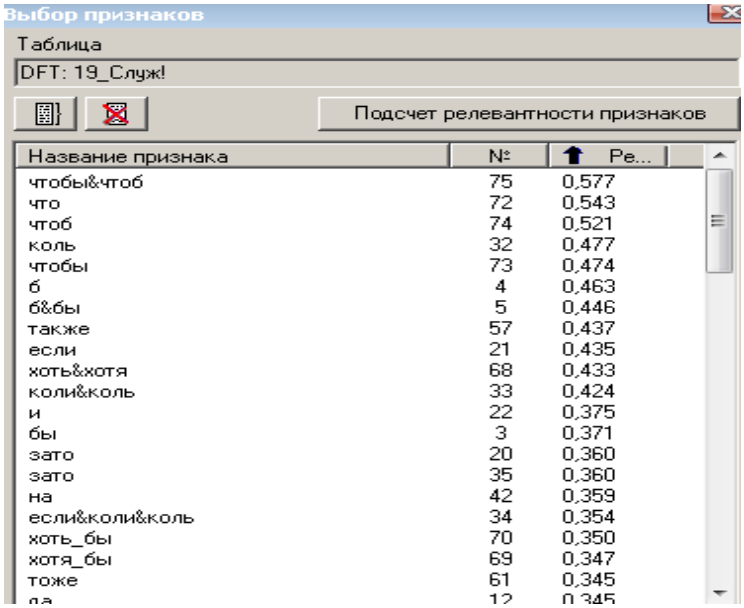
В настоящее время оптимизация может производиться тремя способами:

1) Через функцию «Подсчет релевантности признаков».

Здесь имеются два метода:

а) «Информативность признака в дереве решений» (выделение корней):

Этот метод позволяет отсортировать признаки по степени их релевантности в дереве решений. См. ниже: самые значимые служебные слова – *чтобы (*вместе с вариантом *чтоб), что, чтоб, коль, чтобы.*

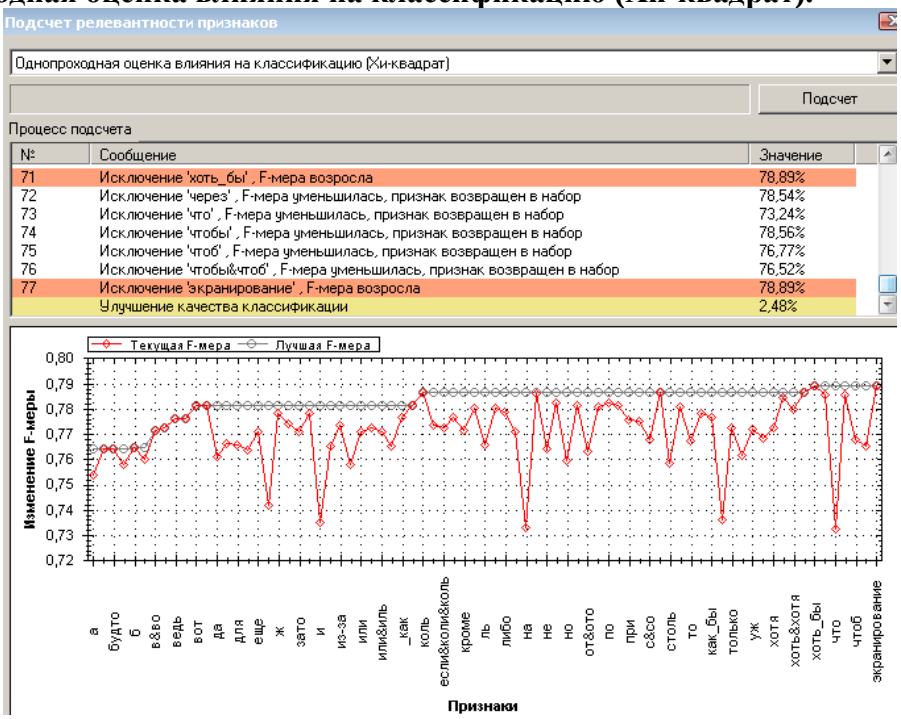


Как показал анализ, отсортированный таким образом и сокращенный список в некоторых случаях дает увеличение F-меры при кластеризации.

б) Однопроходная оценка влияния на классификацию (Хи-квадрат). Позволяет выкинуть из набора те признаки, которые ухудшают результаты классификации. **См. Рисунок № 3.1.**

Рисунок № 3.1.

Однопроходная оценка влияния на классификацию (Хи-квадрат).



С результатами кластеризации этот метод коррелирует плохо. Так, например, при исключении выделенных признаков F-мера понизилась с 55,15 до 52,67.

2) Третий способ оптимизации признакового пространства, разработанный нами в ходе реализации данного проекта в 2011-2013 гг., базируется на выделении устойчивых и уникальных признаков на основе отношения к порогу, установленному по дереву решений. См. Рисунок № 3.2.

Рисунок № 3.2.

Выделение устойчивых и уникальных признаков на основе отношения к порогу, установленному по дереву решений

	а	бы	б&бы	да
Порог	<0,00564295485636115>=	<0,00280142817907168>=	<0,00280373831775701>=	<0,00268273790406378>=
БунинИА		все ниже	все ниже	
ГогольНВ		все выше	все выше	
ГончаровИА				
ДостоевскийФМ				
КупринАИ		все ниже	все ниже	все ниже
ЛермонтовМЮ				все ниже
ЛесковНС	все выше	все ниже	все ниже	
ПушкинАС		все ниже	все ниже	
СалтЩедрМЕ				
ТолстойЛН	все ниже	все ниже	все ниже	все ниже
ТургеневИС				
ЧернышНГ		все выше	все выше	все выше
ЧеховАП				

На **рис. № 3.2.** красным выделен автор (Лесков), у которого союз «а» во всех текстах >= частотного порога; синим - автор (Толстой), у которого союз «а» во всех текстах ниже порога. Только у этих авторов признак ведет себя устойчиво. У остальных он колеблется (нет выделения). На этой основе можно составлять для каждого типа текстов (отличающихся автором, жанром, тематикой и пр.) наборы устойчивых признаков и наборы признаков, обладающих наибольшей различающей силой. См. например, Таблицу 2 с количественными данными по текстам того же корпуса 13 классиков 19 века.

Таблица 2.

Признак	**>=**	**<**	**Устойчив у N типов (здесь: авторов)**
а	1	1	2 из 13 типов
без&безо	2	2	4 из 13 типов
будто	3	0	3 из 13 типов
бы	2	5	7 из 13 типов
б	3	3	6 из 13 типов
б&бы	2	5	7 из 13 типов
в&во	1	3	4 из 13 типов
вдруг	2	1	3 из 13 типов

Очевидно, что наиболее значимы признаки, показывающие устойчивость у наибольшего числа авторов и при этом противопоставляющие их по принципу «выше/ниже». Ср., например, *бы*, *б*, *б&бы*.

3.3. Методика просеивания выборки на саму себя.

Для общей оценки эффективности набора признаков может быть использовано также методика просеивания выборки на саму себя и результаты, полученные при этом просеивании на наборе служебных слов (76 единиц).

Этот набор дает идеальное распознавание - 100% однозначно распознанных текстов) на текстах классиков достаточно большого объема. Такое распознавание было получено на выборке романов и повестей классиков (89) и на выборке их больших рассказов (больше 4500 слов, 39 текстов). Из маленьких текстов (52 текста) было однозначно распознано 42 (остальные получили двух авторов)

На полной выборке (180 текстов разных жанров и объемов) этот набор признаков дает 84 случая однозначного распознавания и ни одного набора авторов больше 3. Этот результат, с учетом большой степени устойчивости и уникальности служебных слов, можно считать эталонным для выборки из худож. текстов разного жанра, объема и авторов. На его основе может оцениваться распознающая сила остальных наборов признаков.

3.4. Исследование влияния диалогов на авторский стиль.

Был проведен сопоставительный анализ результатов распознавания авторов на текстах полных, этих же текстах с вырезанными диалогами и квази-текстах, состоящих из одних диалогов. Проверялись следующие гипотезы:

1) Диалоги представляют собой особый тип языка (разговорную речь), в котором авторские стили нейтрализуется;
2) Устранение диалогов из текстов должно существенно повысить их распознаваемость.

Были получены следующие результаты.

а) Кластеризация

Кластеризация	С выраз. диалогами.	Одни диалоги	Полные тексты
F-мера	50, 21	44,67	**55,15**
БунинИА	47,62	41, 38	52,94
ГогольНВ	43,75	42,42	40
ГончаровИА	53,33	42,86	72,73
ДостоевскийФМ	64,15	36,84	60,47
КупринАИ	66,67	80	66,67
ЛермонтовМЮ	57,14	80	50
ЛесковНС	44,44	57,14	54,55
ПушкинАС	40, 15	33,33	75
СалтЩедрМЕ	80	57,16	75
ТолстойЛН	93,33	54,55	80
ТургеневИС	40,5	49,02	50,91
ЧернышНГ	100	60	100
ЧеховАП	36,23	36,36	42,11

б) Классификация

Без диалогов

Блоки	Полнота	Точность	F-мера
БунинИА	0,9000	0,6429	0,7500
ГогольНВ	0,8462	0,5500	0,6667
ГончаровИА	0,8333	0,7143	0,7692
ДостоевскийФМ	0,7692	0,8333	0,8000
КупринАИ	1,0000	0,1818	0,3077
ЛермонтовМЮ	1,0000	0,7143	0,8333
ЛесковНС	0,4286	0,7500	0,5455
ПушкинАС	0,9000	0,8182	0,8571
СалтЩедрМЕ	1,0000	0,8333	0,9091
ТолстойЛН	0,8750	1,0000	0,9333
ТургеневИС	0,7000	0,9655	0,8116
ЧернышНГ	1,0000	0,3333	0,5000
ЧеховАП	0,4286	0,8824	0,5769
Итого	0,8216	0,7092	0,7123

Одни диалоги

Блоки	Полнота	Точность	F-мера
БунинИА	0,3889	0,2692	0,3182
ГогольНВ	0,3846	0,4545	0,4167
ГончаровИА	0,0000	0,0000	0,0000
ДостоевскийФМ	0,4000	0,5000	0,4444
КупринАИ	1,0000	0,2857	0,4444
ЛермонтовМЮ	0,6667	0,4000	0,5000
ЛесковНС	0,2857	0,5000	0,3636
ПушкинАС	0,8000	0,1250	0,2162
СалтЩедрМЕ	0,2000	0,0500	0,0800
ТолстойЛН	0,1250	0,1429	0,1333
ТургеневИС	0,3250	0,7647	0,4561
ЧернышНГ	1,0000	0,3750	0,5455
ЧеховАП	0,1515	0,6250	0,2439
Итого	0,4406	0,3455	0,3202

Полные тексты

Блоки	Полнота	Точность	F-мера
БунинИА	0,8000	0,8889	0,8421
ГогольНВ	1,0000	0,8667	0,9286
ГончаровИА	1,0000	1,0000	1,0000
ДостоевскийФМ	0,8077	0,7778	0,7925
КупринАИ	1,0000	0,2222	0,3636
ЛермонтовМЮ	1,0000	0,8333	0,9091
ЛесковНС	0,7143	0,5556	0,6250
ПушкинАС	1,0000	0,7143	0,8333
СалтЩедрМЕ	1,0000	0,5556	0,7143
ТолстойЛН	0,8750	1,0000	0,9333
ТургеневИС	0,8500	0,8947	0,8718
ЧернышНГ	1,0000	0,4286	0,6000
ЧеховАП	0,3714	0,8667	0,5200
Итого	0,8783	0,7388	0,7641

Текстов-диалогов меньше, так как не во всех произведениях есть диалоги.

в) «Просеивание» на полной выборке (180 текстов)

	Без Диалогов			*Одни диалоги*					*Полные тексты*		
	1 авт.	2 авт.	3 авт.	1 авт.	2 авт.	3 авт.	4 авт.	5 авт.	1 авт.	2 авт.	3 авт.
Бунин		8	12		5	9	2	2		2	18
Гоголь	4	6	3	1	2	7	3		5	7	1
ГончаровИА		6				1	2	3	2	3	1
ДостоевскийФМ	25	1		7	12	2	2	2	18	8	
КупринАИ		2		2						2	
ЛермонтовМЮ	3	2			2	1			3	2	
ЛесковНС	5	2		4	2		1		7		
ПушкинАС	7	3			2		3			3	7
СалтЩедрМЕ	3	2			4	1			5		
ТолстойЛН	6	1	1	3	2	1	1	1	5	3	
ТургеневИС	22	18		5	9	7	13	6	18	22	
ЧернышНГ	3			2	1				1	1	1
ЧеховАП	10	18	7	1	12	12	8		20	12	3
	88	**69**	**23**	**25**	**53**	**41**	**35**	**14**	**84**	**65**	**31**

Выводы по разделу 3.4.

1. Гипотеза 1 подтвердилась: на диалогах распознаваемость у всех трех методов резко падает. Ср. F-меру на классификации (0,32).
2. Гипотеза о существенном увеличении распознаваемости при устранении диалогов не подтвердилась. Результаты кластеризации и классификации упали по сравнению с полными текстами (кластеризация 50,21 и 55, 15; классификация 0,71 и 0,76). Только просеивание дает незначительно улучшение за счет повышения однозначного распознавания (88 и 84 текста).

Возможно, отсутствие улучшения результатов связано с тем, что при автоматическом вырезании диалогов в них попадает и авторская речь, находящаяся внутри прямой. Т.е. необходимо дальнейшее совершенствование алгоритма и новые эксперименты с более полным контролем производимых изменений в текстах.

3. 5. Развитие экспериментов при использовании системы «Стилеанализатора2-web» по проверке новых, экспериментально ещё не исследованных признаков

3.5.1. Использование в качестве признаков результатов морфемного членения и словообразовательного анализа.

Было проведено исследование некоторых морфемно-словообразовательных моделей, которые могут представлять наибольший интерес с точки зрения авторского стиля. Это R_чик, R_ци-я, R_ушк-а, R_ть-е, R_ств-о, R_стви-е, R_очк, R_ость, R_нь-е, R_ни-е, R_нк-а, R_к-а, R_ист, R_ирова-ть, R_ик, R_изм, R_еньк, R_ейш-ий, R_ость_ейший, R_ист-ик, R_ист_нь-е, R_ист_ость, R_ци_ость.

Этими моделями были размечены те лексемы, которые присутствуют в тексте. См. Рисунок № 3.3.

Рисунок № 3.3. Фрагмент текста, размеченного морфемно-словообразовательными моделями.

На основе данной разметки оценивались такие качества индивидуального стиля, как

а) использование заимствованных суффиксальных моделей;

б) использование модели превосходной степени;

в) использование ум.-ласк. моделей;

г) использование номинализаций.

Обнаружились следующие распознающие возможности комплекса моделей:

- кластеризация - 40,58;
- классификация (с искл. одного) - очень низкая - 0,17

Блоки	Полнота	Точность	F-мера
БунинИА	0,5000	0,2703	0,3509
ГогольНВ	0,1538	0,1818	0,1667
ГончаровИА	0,5000	0,2727	0,3529
ДостоевскийФМ	0,2308	0,4286	0,3000
КупринАИ	0,5000	0,0588	0,1053
ЛермонтовМЮ	0,2000	0,0476	0,0769
ЛесковНС	0,0000	0,0000	0,0000
ПушкинАС	0,1000	0,0833	0,0909
СалтЩедрМЕ	0,2000	0,1111	0,1429
ТолстойЛН	0,0000	0,0000	0,0000
ТургеневИС	0,1500	0,5000	0,2308
ЧернышНГ	0,3333	0,2000	0,2500
ЧеховАП	0,1143	0,2857	0,1633
Итого	0,2294	0,1877	0,1716

- просеивание - однозначно распознаны только 5 текстов (из 180).

При этом мы все равно получаем данные о специфике авторских стилей в отношении исследуемых признаков, так как выявляем устойчивость и уникальность каждого из них. См. Рисунок № 3.4.

Рисунок № 3.4. Фрагмент таблицы с данными об устойчивости и уникальности исследуемых словообразовательных признаков

	R_к-а	R_ик	R_еньк	R_ист-ик	R_нка
Порог	3,18777932916372E-06>=	<1,82165953183335E-05>=	<0,000564812199943519>=	<0,000185554576239737>=	<1,364647443333301E-05>=
БунинИА				все ниже	все ниже
ГогольНВ			все выше	все ниже	все ниже
ГончаровИА		все выше	все выше	все ниже	все ниже
ДостоевскийФМ			все выше	все ниже	все ниже
КупринАИ			все выше	все ниже	все ниже
ЛермонтовМЮ				все ниже	все ниже
ЛесковНС		все выше		все ниже	все ниже
ПушкинАС				все ниже	все ниже
СалтЩедрМЕ				все ниже	
ТолстойЛН		все выше	все выше		все ниже
ТургеневИС				все ниже	все ниже
ЧернышНГ	се ниже	все ниже	все выше	все ниже	все ниже
ЧеховАП					все ниже

Так, комплексные категории, собранные из нескольких моделей, ведут себя так (см. Рисунок № 3.5).

Рисунок № 3.5. **Комплексные словообразовательные категории (Ум.-ласк., Номинализация, ИноязычнМодели) как признаки уникальности и устойчивости стиля автора**

	Ум-ласкат	Номинализации	ИноязычнМодели
Порог	<0,0023665138293151 9>=	<0,0142739745605259>=	<5,49299642955232E-05>=
БунинИА			
ГогольНВ	все ниже		все выше
ГончаровИА	все ниже		все выше
ДостоевскийФМ			все выше
КупринАИ		все выше	все выше
ЛермонтовМЮ	все ниже		
ЛесковНС	все ниже		все выше
ПушкинАС	все ниже	все выше	
СалтЩедрМЕ		все выше	все выше
ТолстойЛН		все выше	все выше
ТургеневИС			
ЧернышНГ		все выше	все выше
ЧеховАП			

Из этих данных следует, что выделяется индивидуальный стиль со словообразовательным признаком устойчиво низкой «ум.-ласк.» (Гоголь, Гончаров, Лермонтов,Лесков, Пушкин) и с колеблющимся признаком (остальные авторы); стиль с устойчивой высокой степенью номинализаций, т.е. большей степенью сложности и абстрактности и т.п.

3.5.2. Использование наборов лексических единиц для психолингвистического анализа особенностей личности автора.

С этой целью были проведены эксперименты по определению эмотивно-смысловой доминанты текстов автора на основе классификации литературных текстов, предложенной В.П. Белянiным (см. «Основы психолингвистической диагностики. Модели мира в литературе»). Анализировались группы лексем, позволяющих, по мнению автора, выделять 5 типов текстов: ***светлые, темные, печальные, красивые, веселые***, соответствующие определенной акцентуации личностей (светлые - паранойальность, темные - эпилептоидность, печальные - депрессивность, красивые - истероидность; веселые - маниакальность). Они сводились в один признак. Ставилась задача определить, как полученные автоматически и по нашим методам результаты соотносятся с оценками е В.П.Белянина, сделанными в ходе наблюдений над сюжетом и особенностями текстов. Априорно было ясно, что автоматический анализ не может дать столь же качественных результатов, как анализ эксперта. Но было важно установить, может ли он что-либо дать на этом типе данных, без снятия многозначности и др. шума.

Были получены следующие результаты. См. Рисунок № 3.6.

Рисунок № 3.6. **Автоматическое отнесение текстов к категориям *светлые, темные, печальные, красивые, веселые* по лексическим данным**

ТипыТекстов По Белянину	темные_тексты	светлые_тексты	печальные_тексты	красивые_тексты	веселые_тексты
Порог	<0,00831946755407654>=	<0,0108699915668085>=	<0,00762388818297332>=	<0,00837893296853625>=	<0,0058870810071036>=
БунинИА	20,00%(4), 80,00%(16)	35,00%(7), 65,00%(13)	10,00%(2), 90,00%(18)	35,00%(7), 65,00%(13)	45,00%(9), 55,00%(11)
ГогольНВ	100,00%(13), 0,00%(0)	100,00%(13), 0,00%(0)	38,46%(5), 61,54%(8)	69,23%(9), 30,77%(4)	100,00%(13), 0,00%(0)
ГончаровИА	100,00%(6), 0,00%(0)	33,33%(2), 66,67%(4)	66,67%(4), 33,33%(2)	50,00%(3), 50,00%(3)	33,33%(2), 66,67%(4)
ДостоевскийФМ	100,00%(26), 0,00%(0)	92,31%(24), 7,69%(2)	61,54%(16), 38,46%(10)	92,31%(24), 7,69%(2)	80,77%(21), 19,23%(5)
КупринАИ	100,00%(2), 0,00%(0)	50,00%(1), 50,00%(1)	0,00%(0), 100,00%(2)	0,00%(0), 100,00%(2)	0,00%(0), 100,00%(2)
ЛермонтовМЮ	100,00%(5), 0,00%(0)	60,00%(3), 40,00%(2)	0,00%(0), 100,00%(5)	0,00%(0), 100,00%(5)	0,00%(0), 100,00%(5)
ЛесковНС	100,00%(7), 0,00%(0)	100,00%(7), 0,00%(0)	71,43%(5), 28,57%(2)	85,71%(6), 14,29%(1)	100,00%(7), 0,00%(0)
ПушкинАС	100,00%(10), 0,00%(0)	90,00%(9), 10,00%(1)	30,00%(3), 70,00%(7)	70,00%(7), 30,00%(3)	70,00%(7), 30,00%(3)
СалтЩедрМЕ	100,00%(5), 0,00%(0)	100,00%(5), 0,00%(0)	100,00%(5), 0,00%(0)	100,00%(5), 0,00%(0)	100,00%(5), 0,00%(0)
ТолстойЛН	100,00%(8), 0,00%(0)	100,00%(8), 0,00%(0)	62,50%(5), 37,50%(3)	50,00%(4), 50,00%(4)	12,50%(1), 87,50%(7)
ТургеневИС	92,50%(37), 7,50%(3)	97,50%(39), 2,50%(1)	35,00%(14), 65,00%(26)	55,00%(22), 45,00%(18)	80,00%(32), 20,00%(8)
ЧернышНГ	100,00%(3), 0,00%(0)	66,67%(2), 33,33%(1)	33,33%(1), 66,67%(2)	100,00%(3), 0,00%(0)	0,00%(0), 100,00%(3)
ЧеховАП	85,71%(30), 14,29%(5)	80,00%(28), 20,00%(7)	31,43%(11), 68,57%(24)	42,86%(15), 57,14%(20)	57,14%(20), 42,86%(15)

В скобках указано количество текстов, в которых группа слов, характеризующих тип текста, имеет частоту < или >= порога.

На основе этих данных нами производится оценка авторского стиля с точки зрения выделенных групп (учитывается только поведение в отношении «>=порога»)

Бунин	печальный (18 т.).; темный (16); светлый (13); красивый (13); веселый (11).
Гоголь	печальный тип (8 т.)
Гончаров	светлый (4) и веселый (4) тип
Дост	нет доминанты (во всех текстах частоты ниже порога)
куприн	печальный, красивый, веселый типы (2 т.)
Лермонтов	печальный, красивый, веселый типы (2 т.)
Лесков	нет доминанты (во всех текстах частоты ниже порога)
Пушкин	печальн тип (7 т.).
Толстой	веселый (7 т), красивый (4 т.)
Тургенев	печальный тип (26 т.)
Черн	веселый (3) и печальный (2) тип
Чехов	печальн (24 т.) и красивый (20 т.) тип

В.П. Белянин отмечал, что печальный тип текстов часто встречается у Гоголя, Бунина, Тургенева. Это совпадает с полученными результатами. Светлым, по его мнению, является Обломов (см. представленность типа «светлый» у Гончарова). Однако у Достоевского, при его явной эпилептоидности, признаки темных текстов нашим методом не выявляются.

Как видно, данный тип признаков нуждается в дальнейшем исследовании.

3.5.3. Среднее число слогов в слове и средняя длина предложений

Среднее число слогов в слове считается важным показателем легкости понимания стиля. Среднее число для каждого из авторов равно:

БунинИА	1,665
ГогольНВ	1,661
ГончаровИА	1,460
ДостоевскийФМ	1,553
КупринАИ	1,665
ЛермонтовМЮ	1,606
ЛесковНС	1,602
ПушкинАС	1,682
СалтЩедрМЕ	1,646

ТолстойЛН	1,617
ТургеневИС	1,563
ЧернышНГ	1,590
ЧеховАП	1,607

Этот параметр достаточно близок у всех классиков. Наименьшая длина слова у Гончарова и Достоевского. Наибольшая - у Бунина.

Средняя длина предложения для каждого автора составляет:

БунинИА	13,20304
ГогольНВ	12,87174
ГончаровИА	9,161571
ДостоевскийФМ	10,94739
КупринАИ	10,99297
ЛермонтовМЮ	10,24113
ЛесковНС	11,54955
ПушкинАС	11,11638
СалтЩедрМЕ	11,59827
ТолстойЛН	12,51189
ТургеневИС	9,905075
ЧернышНГ	13,67995
ЧеховАП	10,32521

Наименьшая средняя длина предложения у Гончарова и Тургенева; наибольшая у Бунина и Чернышевского.

	Макс. длина текста	Мин. длина текста
Бунин	23,3	6
Гоголь	23,8	9,7
Гончаров	9,5	7,8
Достоевский	17,4	5,7
Куприн	11,31	10,9
Лермонтов	13,8	8,9
Лесков	16,34	9,5
Пушкин	14,4	9,4

Салт-Щедрин	14,4	10,9
Толстой	15,3	11,7
Тургенев	15,7	7,3
Черныш	15,8	11,87
Чехов	18,9	5,22

Таким образом, **Гончаров и Бунин** составляют два полюса по длине единиц: у Гончарова самые короткие словоформы и предложения; у Бунина самые длинные. Разброс длин предложений в текстах у Бунина максимален, а у Гончарова минимален (вместе с Куприным).

Распознавание по этим параметрам неэффективно, поскольку дает очень плохое результаты на всех методах. Так, при выделении длин предложений с интервалом в 5 слов (от 1 до 100) кластеризация = 39,79; классификация 0,17; просеивание - ни одной однозначной атрибуции.

3.5.4. Использование тезаурусных групп для распознавания индивидуального стиля.

В связи с важностью и сложностью автоматизации прагматического анализа текста были проведены исследования такого признака, как тип эмоции. Для этого на основе «Словаря эмоций» Л.Бабенко были подготовлены анализаторы-запросы на 37 типов эмоций: Беспокойство, Вдохновение, Вера, Влечение, Высокомерие, Горе, Грусть, Доброта, Дружба, Жалость, Желание, Жестокость, Злость, Искренность, Лицемерие, Любовь, Любопытство, Наглость, Надежда, Неверие, Недовольство, Неприязнь, Обида, Одиночество, Одобрение, Протест, Равнодушие, Радость, Смелость, Смирение, Сомнение, Спокойствие, Страх, Стыд, Уважение, Удивление, Удовольствие.

Ниже представлен фрагмент «эмоциональной карты», полученной с помощью выявления устойчивых признаков на основе дерева решений (см. Рисунок 3.7).

Рисунок 3.7.

19_Эмоции	**Радость**	**Горе**	**Страх**	**Злость**	**Любовь**	**Вера**	**Неверие**	**Неприязнь**	**Одобрение**	**Протест**	**Обида**
Порог	<0,00571891641583 7>=	<0,00799128223755903>=	<0,00174186046 51163>=	<0,00245 0065827 91808>=	<0,00867340 723596053>=	<0,00062810 124992 1487>=	<2,12666411 466973E-05>=	<0,00300 1598092 22565>=	<0,00031 1477962 934122>=	<4,3045930 0073178E-05>=	<0,00065 4022236 75605>=
БунинИА											
ГогольНВ									все выше	все выше	
ГончаровИА							все выше	все ниже	все выше	все выше	
ДостоевскийФМ			все выше			все выше			все выше		
КупринАИ	все выше	все ниже	все выше	все выше	все выше	все выше	все выше	все выше	все выше	все выше	все выше
ЛермонтовМЮ	все выше					все выше	все выше		все выше		
ЛесковНС	все ниже			все ниже	все ниже	все выше	все выше	все ниже	все выше	все выше	все выше
ПушкинАС									все выше		
СалтЩедрМЕ	все ниже	все ниже		все ниже	все ниже	все выше	все выше		все выше	все выше	
ТолстойЛН				все ниже				все ниже	все выше	все выше	
ТургеневИС								все ниже			
ЧернышНГ	все ниже	все выше	все ниже	все ниже	все выше	все выше	все выше		все выше	все выше	все выше
ЧеховАП											

Методы и подходы, использованные в ходе выполнения данного проекта

Та комбинация этих методов, которая реализована в проекте, является оригинальной, нигде ранее не представленной. Использовались следующие методы:

— методы Web-технологий; - методы компьютерного моделирования;

— методы современной математической статистики, а также методы искусственного интеллекта;

— методы компьютерного анализа лингвистических данных с помощью лингвистических анализаторов, словарей и характеристических списков морфем, слов, словосочетаний различного рода;

— методы и теоретические модели современной системной лингвистики, позволяющие прогнозировать и оценивать получаемые экспериментальные данные по классификации текстов различных классов с помощью признаков различного рода.

Особенно надо выделить оригинальность той комбинации методов современной математической статистики и методов современной теории искусственного интеллекта, которая была разработана одним из участников нашего творческого коллектива профессором факультета информатики Томского государственного университета Василием Васильевичем Поддубным.

Методы современной системной лингвистики и компьютерного анализа лингвистических данных с помощью лингвистических анализаторов, словарей и характеристических списков n-грамм, морфем, слов, словосочетаний различного рода также характеризуются широтой и оригинальностью, что и позволяло для одних и тех же анализируемых текстов получать широкий спектр их отображений по разным характеристическим признакам разных единиц этих текстов.

Степень новизны полученных результатов

Все полученные результаты (как исходные характеристические данные, так и результаты их анализа), полученные в ходе создания и экспериментальной эксплуатации тексто-аналитической системы являются новыми.

Сопоставление полученных результатов с мировым уровнем:

В настоящем проекте ставилась задача создания такой относительно универсальной тексто-аналитической системы, которая бы базировалась на современных идеях создания мощных и гибких распределённых вычислительных комплексов, на использовании широкого арсенала статистических и логико-математических инструментов анализа признаков текстов, а также современного лингвистического обеспечения для создания анализируемой признаковой базы. В данном случае разрабатывалось лингвистическое обеспечение для анализа текстов русского языка. Однако многие функции существующей системы

«Web-СтилеАнализатор» уже и сейчас являются языко-независимыми, т.к. в этой системе возможна безболезненная замена лингвистических анализаторов одного языка на анализаторы любого другого, использующего в своей графематической системе фонолого-фонетический принцип.

На основе полученной исходной количественной информации в системе «Web-СтилеАнализатор» в ходе использования статистических методов и методов искусственного интеллекта могут производиться различного рода распознавательные и категоризационно-классификационные процедуры с текстами – определение основных тем текста, основных его модальностей, отношения автора к обозначаемым объектам, определение стиля текста (функционального стиля и жанра текста – монография, статья, репортаж, речь на митинге и т.п.), типа автора (пол, возраст, образование и т.п.), конкретного автора, близости стиля и содержания текстов различных авторов по различным критериям (при наличии текстового материала для сопоставления конкурирующих гипотез) и т.п. При этом может осуществляться оптимизация списков характеристических признаков (за счёт исключения малоинформативных и шумовых), а также оптимизация списков текстов при формировании эталонных выборок текстов (за счёт выявления и элиминации нетипичных текстов).

Какой-либо другой аналогичный проект ни в готовом виде, ни в стадии подготовки нам не известен.

За границей в чём-то аналогичным по масштабу, но сильно отличающимся по намеченным функциям (в сторону их относительной редукции) проектом занимаются в Тюбингенском университете. Их продукт называется «Dynamic Corpus Analyzer» (DCA) [http://www.thomas-zastrow.de/dca.php]. (См., например: Marie Hinrichs, Thomas Zastrow and Erhard Hinrichs. WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure, in: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valetta, Malta. Published by European Language Resources Association (ELRA). – 2010; Ulrich Heid, Fabienne Fritzinger, Erhard Hinrichs, Marie Hinrichs and Thomas Zastrow. Term and Collocation Extraction by Means of Complex Linguistic Web Services, in: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valetta, Malta. Published by European Language Resources Association (ELRA). – 2010). Этот проект также выполняется на Java в виде веб-приложения, однако на базе другой технологии - ZK Webframework (http://www.zkoss.org/).

Основные функциональные и структурные отличия «Dynamic Corpus Analyzer» от нашего проекта:

1. Ориентация их проекта на веб-сервисы. Многие функции их системы выполнены в виде отдельных и зачастую сторонних приложений, располагающихся на других серверах. С одной стороны, это обеспечивает блоковость и масштабируемость системы, но, с другой, существенно замедляет его работу за счет более медленных связей между блоками, а также необходимости в наличии общих универсальных форматов передачи данных.

2. Большой акцент их проекта на стандартизацию в ущерб оптимизации. Тексты и их грамматическая разметка представлены у них в XML-файлах, которые за счет избыточного формата занимают много памяти и существенно затрудняют аналитическую обработку.

3. Необходимо отметить, что при сильной проработке в их проекте корпусных функций наблюдается ограниченный набор аналитических инструментов. Например, включение методов классификации и кластеризации пока даже не планируется. В нашей системе «Web-СтилеАнализатор» это не просто планируется, но уже реализован и используется широкий спектр современных средств статистического анализа, в т.ч. различение и распознавание текстов по стилям, авторам и т.д. методами дискриминантного и факторного анализа.

Кроме того в «Web-СтилеАнализаторе» используются также логические и логико-вероятностные методы искусственного интеллекта и распознавания образов.

4. В системе «Dynamic Corpus Analyzer» представлен ограниченный набор средств анализа:
— простая статистика символов и слов,
— переходные вероятности на уровне основных грамматических классов слов (частей речи) и отображение распределений частей речи,
— поиск по выбранным словоупотреблениям с разными параметрами, включая нечеткий поиск, пар слов с заданным расстоянием, части речи словоупотребления, комбинаций частей речи,
— представление текстовых частотных законов (например, закона Ципфа),
— расчет связей объёма текста и соответствующего ему объёма словаря (тип-токен) и др.
В ближайшее время авторы «Dynamic Corpus Analyzer» не планируют делать упор на аналитические функции, в их планах - улучшение работы с пользователями, добавление новых форматов данных, более тонкая настройка существующей функциональности, добавление новой функциональности в общем и улучшение процедур экспорта.

В то же время, в нашей системе планируется использовать широкий спектр лингвистических анализаторов текста (часть из которых уже реализована и прошла проверку), что определяет более широкий спектр возможных приложений нашей системы, а также возможность получения относительно более интересных, более точных, более надёжных результатов, чем в системе «Dynamic Corpus Analyzer» - на основе взаимной перепроверки независимо получаемых классификационных результатов по разным методам, разным критериям, разным лингвистическим единицам и их признакам).

Список опубликованных работ
в течение выполнения проекта по теме проекта

1. Кукушкина О.В., Е.В. Суровцева Е.В., Л.В.Лапонина, Д.Ю.Рюдигер. Под общ. ред. Поликарпова А.А. (2012). Частотный грамматико-семантический словарь языка художественных произведений А.П. Чехова. – М: МаксПресс, 2012. - 572 с. ISBN 978-5-317-64249-3.
2. Ашуров М.Ф., Поддубный В.В. Потоковый метод классификации текстов художественной литературы на основе С-меры // Информационные технологии и математическое моделирование (ИТММ-2013): Материалы XII Всероссийской научно-практической конференции с международным участием им. А. Ф. Терпугова (29-30 ноября 2013 г.). – Томск: Изд-во Том. ун-та, 2013. – Ч.2. – С.85–89.
3. Кубарев А.И., Кукушкина О.В., Поддубный В.В., Шевелёв О.Г. Построение таблиц стилей текстовых произведений с использованием алгоритмов классификации на основе деревьев решений // Вестник Томского государственного университета. «Управление, вычислительная техника и информатика». 2012. № 4 (21). С. 79-88.
4. Кубарев А.И., Поддубный В.В. Хранение корпусов текстов произвольной структуры и паспортизации в реляционной базе данных // Информационные технологии и математическое моделирование (ИТММ-2013): Материалы XII Всероссийской научно-практической конференции с международным участием им. А. Ф. Терпугова (29-30 ноября 2013 г.). – Томск: Изд-во Том. ун-та, 2013. – Ч.1. – С.139–143.
5. Кубарев А.И., Поддубный В.В. Байесовская классификация с обучением на основе использования копула-функций // Информационные технологии и математическое моделирование (ИТММ-2013): Материалы XII Всероссийской научно-практической конференции с международным участием им. А. Ф. Терпугова (29-30 ноября 2013 г.). – Томск: Изд-во Том. ун-та, 2013. – Ч.2. – С.126–130.
6. Поддубный В. В., Кубарев А. И., Шевелёв О. Г., Кукушкина О. В. Построение таблиц стилей текстовых произведений с использованием алгоритмов классификации на основе деревьев решений // Международная конференция "Современные проблемы математики, информатики и биоинформатики", посвященная 100-летию со дня рождения члена-корреспондента АН СССР Алексея Андреевича Ляпунова (11 - 14 октября 2011 г., Академгородок , Новосибирск, Россия). № гос. регистрации – 0321103262, ISBN 978-5- 905569-03-6. [Электронный ресурс]. - URL: http://conf.nsc.ru/Lyap-00/ru/reportview/74806.
7. Поддубный, В.В., Пехтерев, А.С. Копулы сглаженных эмпирических распределений при наличии связок (совпадений) и их применение в имитационном моделировании // Труды XII Международной ФАМЭБ’2013 конференции. / Под ред. Олега Воробьёва. – Красноярск: НИИППБ, СФУ, 2013. – С. 312–321.
8. Поддубный В.В., Пехтерев А.С. Минимизация числа информативных признаков методом тестового распознавания в задаче классификации текстов по дереву решений // Труды XVI Международной конференции по эвентологической математике и смежным вопросам (ЭМ’2012, Красноярск, 7-8 декабря 2012 г.) / Под ред. О. Воробьёва. - Красноярск: СФУ, НИИППБ, ТЭИ, 2012. - С. 173-178.
9. Поддубный В.В., Пехтерев А.С. Анализ влияния ошибок на качество логического распознавания // Новые информационные технологии в исследовании сложных

структур: Материалы Девятой Российской конференции с международным участием. - Томск: Изд-во НТЛ, 2012. - С. 76.

10. Поддубный В.В., Пехтерев А.С. Статистико-эвентологический подход к анализу влияния одиночных ошибок описания объектов на качество тестового распознавания // Труды XV Международной ЭМ’2011 конференции / Под ред. О. Воробьёва. - Красноярск: СФУ, НИИППБ, КГТЭИ, 2011. - С. 157-162.
11. Поддубный В.В., Поликарпов А.А. Диссипативная стохастическая динамическая модель развития языковых знаков // Компьютерные исследования и моделирование. 2011. - Т. 3. № 2. - С. 103-124.
12. Поликарпов А.А., Поддубный В.В. Вывод закона синхронного полисемического распределения языковых знаков на основе диссипативной стохастической динамической модели эволюции знаковых ансамблей // Синхронное и диахронное в сравнительно-историческом языкознании. - М: Добросвет, Изд-во "КДУ", 2011. - С. 182-190.
13. Poddubny, V. V., Polikarpov, A. A. (2013). Stochastic Dynamic Model of Evolution of Language Sign Ensembles // Methods and Applications of Quantitative Linguistics - Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO), edited by Ivan Obradovic, Emmerich Kelih and Reinhard Kohler. – Belgrad: University of Belgrade and Academic Mind Publishers, 2013. - Pp. 69-83. ISBN 9768-86-7466-465-0.
14. Polikarpov, A. A., Poddubny, V. V. (2012). From Signs’ Life Cycle Regularities toMathematical Modelling of Language Evolution: Explaining the Mechanism for the Formation of Words’ Synchronous Polysemy and Frequency of Use Distributions // Five Approaches to Language Evolution Proceedings of the Workshops of the 9th International Conference on the Evolution of Language. – Kyoto, 2012. - Pp. 114-123.