

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «НАЦИОНАЛЬНЫЙ
ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «МЭИ»

На правах рукописи

Козлов Павел Юрьевич



**НЕЙРО-НЕЧЕТКИЕ МЕТОДЫ И АЛГОРИТМЫ АНАЛИЗА
ЭЛЕКТРОННЫХ НЕСТРУКТУРИРОВАННЫХ
ТЕКСТОВЫХ ДОКУМЕНТОВ**

Специальность 05.13.17 – Теоретические основы информатики

Диссертация
на соискание ученой степени
кандидата технических наук

Научный руководитель:
доктор технических наук, профессор
М. И. Дли

Москва – 2017

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....	5
1 АНАЛИЗ СОВРЕМЕННЫХ ПОДХОДОВ К	
АВТОМАТИЗИРОВАННОМУ АНАЛИЗУ ТЕКСТОВЫХ ДОКУМЕНТОВ	13
1.1 Общие процедуры и основные задачи автоматизированного анализа	
текстовых документов	13
1.2 Анализ современных методов автоматизированного рубрицирования	
текстовых документов	20
1.3 Перспективы использования методов автоматизированного анализа	
текстов для рубрицирования электронных неструктурированных текстовых	
документов.....	27
1.4 Выводы по главе	38
2 РАЗРАБОТКА МЕТОДОВ И МОДЕЛЕЙ АНАЛИЗА ЭЛЕКТРОННЫХ	
НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ И	
МОНИТОРИНГА РУБРИК.....	40
2.1 Мультимодельный метод анализа и рубрицирования электронных	
неструктурированных текстовых документов	40
2.2 Каскадная нейро-нечеткая модель анализа коротких электронных	
неструктурированных текстовых документов с использованием экспертной	
информации	50
2.2.1 Структура каскадной нейро-нечеткой модели для рубрицирования	
коротких ЭНТД	50
2.2.2 Модель рубрицирования ЭНТД с использованием весовых	
коэффициентов.....	52
2.2.3 Модель формализации ЭНТД для нейро-нечеткого классификатора	
57	
2.2.4 Нейро-нечеткие модели оценки принадлежности ЭНТД к	
отдельным рубрикам.....	58
2.2.5 Модель для выбора рубрики, в наибольшей степени	
соответствующей ЭНТД.....	60
2.2.6 Процедура использования нейро-нечеткого классификатора для	
рубрицирования коротких ЭНТД.....	61
2.3 Модель анализа электронных неструктурированных текстовых	
документов на основе нечеткого дерева решений	62

2.4	Метод мониторинга и изменения рубрик электронных неструктурированных текстовых документов на основе их нечеткой динамической кластеризации	67
2.5	Выводы по главе	77
3	РАЗРАБОТКА АЛГОРИТМОВ АНАЛИЗА НЕСТРУКТУРИРОВАННЫХ ЭЛЕКТРОННЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ И МОНИТОРИНГА РУБРИЧНОГО ПОЛЯ.....	79
3.1	Алгоритмы реализации мультимодельного метода рубрицирования ЭНТД..	79
3.2	Алгоритмы для анализа коротких электронных неструктурированных текстовых документов на основе нейро-нечеткого классификатора с использованием весовых коэффициентов.....	92
3.3	Алгоритмы для анализа коротких неструктурированных электронных текстовых документов на основе нечетких деревьев решений.....	98
3.4	Выводы по главе	102
4	РЕЗУЛЬТАТЫ ПРАКТИЧЕСКОГО ИСПОЛЬЗОВАНИЯ АЛГОРИТМОВ АНАЛИЗА (РУБРИЦИРОВАНИЯ) НЕСТРУКТУРИРОВАННЫХ ЭЛЕКТРОННЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ.....	103
4.1	Структура средств информационной системы автоматизированного анализа электронных неструктурированных текстовых документов.....	103
4.2	Оценка точности рубрицирования электронных текстовых документов с использованием разработанных алгоритмов и средств	108
4.3	Результаты практического использования разработанных алгоритмов рубрицирования неструктурированных электронных текстовых документов в Администрации Смоленской области.....	112
4.4	Выводы по главе.....	123
	ЗАКЛЮЧЕНИЕ	124
	СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ	126
	ГЛОСАРИЙ.....	138
	ПРИЛОЖЕНИЕ 1 Результаты тестирования разработанных алгоритмов автоматизированного рубрицирования ЭНТД.....	142

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

NLP – natural language processing

TDM – text data mining

VSM – vector space model

XML – eXtensible Markup Language

АЛ – алгебраическая лингвистика

АОТ – автоматическая обработка текста

ВЛ – вычислительная лингвистика

ЕЯ – естественный язык

ЗС – значащие слова

ЗСХ – значение семантической характеристики

КЛ – компьютерная лингвистика

КЛ-1 – теоретическая компьютерная лингвистика

КЛ-2 – инженерная компьютерная лингвистика

КС – ключевое слово

КЭНТД – короткий электронный неструктурированный текстовый документ

ЛП – лингвистический процессор

МПО – модель предметной области

ПЛ – прикладная лингвистика

ПТР – плоский текстовый рубрикатор

СПТ – система понимания текстов

СХ – семантические характеристики

ТД – текстовый документ

ТЕЯ – текстовый документ, написанный на естественном языке

ЭНТД – электронный неструктурированный текстовый документ

ВВЕДЕНИЕ

Актуальность темы работы. В настоящее время одним из основных направлений государственной политики в Российской Федерации является повышение степени открытости органов государственной и муниципальной власти различных уровней, в том числе на основе организации их виртуального взаимодействия с населением. В результате происходит процесс постоянного совершенствования интернет-порталов органов исполнительной и законодательной власти, с использованием которых каждый гражданин или организация могут в электронном виде направить сообщение (жалобу, обращение, предложение и т.д.). Число подобных электронных контактов непрерывно растет. Например, за 2016 год в Администрации Санкт-Петербурга и Смоленской области поступило около 38 000 и 10 000 электронных сообщений, соответственно. С учетом жестко регламентированных сроков подготовки ответа возникает необходимость обеспечения автоматизированной обработки указанных сообщений с целью их рубрицирования (классификации) для повышения оперативности взаимодействия с профильными структурными подразделениями администраций. Решение данной задачи непосредственно связано с использованием процедур извлечения данных из текстовой информации на основе применения методов анализа электронных текстовых документов.

Электронные сообщения с точки зрения возможности их автоматизированной обработки обладают рядом специфических особенностей:

- в значительной части случаев небольшой размер, что затрудняет его статистический анализ;
- отсутствие структуризации (специальной разметки и полей для компьютерной обработки), что усложняет процедуры извлечения информации;
- наличие большого количества грамматических и синтаксических ошибок приводит к необходимости реализации нескольких дополнительных этапов обработки;
- нестационарность тезауруса (состава и важности слов), который зависит от выхода новых нормативных документов, выступлений должностных

лиц и политических деятелей и т.д., что приводит к необходимости использования процедур динамической кластеризации рубрик.

Целесообразность динамического мониторинга рубричного поля (состава и характеристик рубрик) также определяется необходимостью адаптации процедур реакции на поступающие сообщения к изменяющимся внешним и внутренним факторам (например, изменениям в организационной структуре органов власти).

Очевидно, что указанные особенности рассматриваемых текстовых документов (которые можно отнести к неструктурированным электронным текстовым документам – далее ЭНТД), накладывают определенные ограничения на алгоритмы применения морфологического, синтаксического и семантического анализов, а также на соответствующие им процедуры формализации информации для автоматизированной обработки текстов, в том числе в рамках виртуальных систем информационного обеспечения различных региональных социально-экономических процессов. В то же время, известные методы, модели и алгоритмы извлечения знаний и данных из текстовой информации не учитывают в требуемой степени необходимость непрерывного исследования динамики рубрик для неструктурированных с точки зрения отсутствия специальной разметки для машинной обработки электронных текстовых документов с последующим учетом выявленных изменений при их разделении на рубрики (рубрицировании). Следует также отметить, что небольшие размеры анализируемых электронных документов определяет целесообразность использования мульти-модельного подхода к их анализу и последующему рубрицированию на основе комплексного использования имеющейся статистической и экспертной информации.

Данная ситуация обуславливает противоречие между необходимостью повышения эффективности процедур автоматизированного анализа электронных неструктурированных текстовых документов в условиях изменения рубрик и несовершенством используемых в настоящее время методов и алгоритмов анализа текста на естественном языке с точки зрения результативности реше-

ния данной задачи. Указанное противоречие определяет актуальность темы научного исследования, которая связана с разработкой и практическим применением нового научно-методического и алгоритмического обеспечения информационных систем органов государственного управления различного уровня, осуществляющих автоматизированный анализ и рубрицирование (классификацию) ЭНТД.

В итоге можно констатировать, что разработка и совершенствование нейронно-нечетких методов и алгоритмов автоматизированного анализа электронных неструктурированных текстовых документов в условиях изменения рубрик является актуальной научно-технической задачей, которая имеет существенное значение для развития теоретических основ информатики в части совершенствования алгоритмов анализа текста и методов извлечения данных из текстов на естественном языке.

Степень разработанности темы. Разработке методов и алгоритмов автоматизированного анализа текстовой информации посвящены труды таких ведущих отечественных и зарубежных ученых, как Бочаров И.А., Виньков М.М., Заболеева-Зотова А.В., Орлова Ю.А., Попов Э.В., Розалиев В.Л., Фальк В. Н., Фоминых И. Б., Харин Н. П., Шаграев А. Г., Berger A., Bevainyte A., Chi Wang, Frank E., Lewis D.D., Manning C., Mitchell T.M., Wang Hong-bin, Quinlan J.R., Raghavan P., Ramage D., Rocchio J.J., Schutze H., Sebastiani F., Witten I.H., Yang Y., а также защищенные диссертационные работы таких авторов, как Александров М.Ю., Бойцов Л.М., Головкин Н.В., Гулин В.В., Епрев А.С., Мокроусов М. Н., Сидорова Е.А., Толчеев В.О., Тревгода С.А., Чугреев В.Л., Шабанов В.И., Шелманов А.О., Шмудевич М.М. В работах данных авторов обоснованы основные подходы к морфологическому, синтаксическому и семантическому анализу электронных текстовых документов.

Вопросы использования интеллектуальных методов в системах автоматизированного анализа и рубрицирования электронных текстовых документов нашли отражение в публикациях таких авторов, как Андреев А.М., Березкин Д.В., Ермаков А.Е., Мешкова Е.В., Морозов В.В., Симаков К.В., Цыганов И.Г.,

Шеменков П.С., а также в защищенных диссертационных работах Коржа В.В., Мешковой Е.В., Николаевой И.В., Полякова Д.В., Шеменкова П.С. Представленные в указанных трудах научные результаты демонстрируют возможность комплексного использования статистических данных и экспертных оценок для более полного извлечения информации из текстовых документов различных видов.

Однако, несмотря на значительное число научных работ по проблемам применения методов автоматизированного анализа и разделение по рубрикам текстовой информации в электронной форме, указанные выше особенности электронных сообщений, представляющих собой в общем случае ЭНТД, в достаточной степени отражения не нашли.

Целью исследования является снижение числа ошибок рубрицирования электронных неструктурированных текстовых документов в условиях изменения состава и характеристик рубрик на основе создаваемых нейро-нечетких методов и алгоритмов анализа этих документов, а также мониторинга и изменения рубрик.

Научная задача диссертации заключается в разработке и исследовании нейро-нечетких методов и алгоритмов анализа электронных неструктурированных текстовых документов.

Для реализации этой цели и решения научной задачи поставлены и выполнены следующие задачи диссертационного исследования:

1. Анализ задач и современных методов автоматизированного рубрицирования текстов и оценка перспектив их использования для анализа электронных неструктурированных текстовых документов с учетом особенностей электронных сообщений граждан в органы государственного и муниципального управления.

2. Разработка мультимодельного метода и алгоритмов анализа электронных неструктурированных текстовых документов с комбинированным использованием нечетко-логических, нейро-нечетких и вероятностных моделей.

3. Создание метода и алгоритмов мониторинга и изменения рубрик элек-

тронных неструктурированных текстовых документов на основе их нечеткой динамической кластеризации.

4. Разработка каскадной нейро-нечеткой модели и модели на основе нечеткого дерева решений для анализа и рубрицирования электронных неструктурированных текстовых документов, а также реализующих их алгоритмов.

5. Оценка точности рубрицирования электронных неструктурированных документов с использованием разработанных методов, моделей, алгоритмов и средств с использованием вычислительных экспериментов. Практическое использование разработанных алгоритмов и программных средств для автоматизированного анализа электронных неструктурированных текстовых документов в Администрации Смоленской области, а также в учебном процессе филиала НИУ «МЭИ» в г. Смоленске.

Объектом исследования являются теоретические основы автоматизированного анализа электронных неструктурированных текстовых документов в информационных системах.

Предметом исследования являются интеллектуальные методы и алгоритмы анализа электронных неструктурированных текстовых документов, а также мониторинга и изменения рубрик.

Соответствие паспорту специальности. Диссертационное исследование соответствует пунктам паспорта специальности ВАК 05.13.17– «Теоретические основы информатики»:

п. 5. «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях, разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений»;

п. 6. «Разработка методов, языков и моделей человеко-машинного общения; разработка методов и моделей распознавания, понимания и синтеза речи, принципов и методов извлечения данных из текстов на естественном языке».

Методологической базой исследования являются: теоретические основы информатики; системный анализ информационных процессов; методы теорий нечеткой логики и искусственных нейронных сетей; научные положения и вы-

воды, сформулированные в трудах отечественных и зарубежных авторов по вопросам автоматизированного анализа текстов на естественном языке.

Научная новизна работы заключается в разработке новых нейро-нечетких методов, комплекса моделей и алгоритмов автоматизированного анализа электронных неструктурированных текстовых документов, а также мониторинга и изменения рубрик этих документов.

Наиболее существенные научные результаты, полученные лично автором и выносимые на защиту, заключаются в следующем:

1. Предложены мультимодельный метод и алгоритмы анализа ЭНТД, отличающиеся комбинированным использованием нечетко-логических, нейро-нечетких и вероятностных моделей, а также представленным в виде системы нечетких продукционных правил набором условий целесообразности их применения с учетом характера динамики рубрик, позволяющие повысить точность выделения рубрик и отнесения к конкретным рубрикам текстовых документов в условиях взаимозависимости рубрик и различного объема статистических данных.

2. Разработаны метод и алгоритмы мониторинга и изменения рубрик (слияния, разделения, появления новых и ликвидации рубрик) для ЭНТД, отличающиеся использованием процедур нечеткой динамической кластеризации этих документов с учетом синтаксических ролей слов, а также числа и характеристик рубрик, что позволяет обеспечить адаптивную актуализацию рубрик в зависимости от структуры и показателей текстовых документов в условиях нестационарности состава тезауруса и важности ключевых слов рубрик.

3. Разработаны каскадная нейро-нечеткая модель и алгоритмы анализа ЭНТД, применяющие экспертную информацию для определения значимости ключевых слов при формализации и последующем рубрицировании текстовых документов на основе нейро-нечеткого классификатора, что позволяет анализировать документы небольшого размера на основе их унифицированного представления.

4. Разработаны нечетко-логическая модель и алгоритмы анализа ЭНТД документов, отличающиеся использованием синтаксических связей и ролей слов, а также нечеткой оценкой различий между документами в n -мерном пространстве признаков текстов при построении нечеткого дерева решений для отнесения документа к конкретной рубрике, что позволяет автоматизировать процедуру анализа с учетом степеней принадлежности документов к различным рубрикам в условиях взаимозависимости их тезаурусов, а также недостатка статистической информации при формировании новых рубрик.

Теоретическая и практическая значимость исследования состоит в развитии научных основ применения современных информационных интеллектуальных технологий для автоматизированного анализа и рубрицирования ЭНТД с использованием средств вычислительной техники для повышения эффективности информационных систем органов государственного и муниципального управления.

Практическая значимость основных положений диссертации также подтверждается результатами использования разработанных программных средств информационной системы автоматизированного анализа электронных неструктурированных текстовых документов в Администрации Смоленской области и учебном процессе филиала НИУ «МЭИ» в г. Смоленске.

Апробация работы. Основные положения и выводы диссертационной работы докладывались на таких научных мероприятиях как: IV Международная научно-техническая конференция «Энергетика, информатика, инновации» (Смоленск, 2013), V Международная научно-техническая конференция «Энергетика, информатика, инновации» (Смоленск, 2014), XII Международная научно-техническая конференция «Информационные технологии, энергетика и экономика» (Смоленск, 2015), V Международная научно-практическая конференция «Математическое моделирование, информатика, экономика» (Смоленск, 2015), XIII международная научно-техническая конференция «Информационные технологии, энергетика и экономика» (Смоленск 2015), VI Международная научно-техническая конференция «Энергетика, информатика, инновации»

(Смоленск, 2016), XIV Международная научно-техническая конференция «Интеллектуальные информационные технологии, энергетика и экономика» (Смоленск, 2017).

Публикации. По теме диссертации опубликовано 11 работ общим объемом 3 п.л., в том числе 3 статьи в научных журналах, рекомендованных ВАК РФ. Авторский вклад – 2,3 п.л.

Структура и объем работы. Диссертационная работа состоит из введения, четырёх глав, заключения, списка литературы, включающего 123 наименования, и одного приложения. Диссертация содержит 148 страниц машинописного текста, 64 рисунка и 12 таблиц.

1 АНАЛИЗ СОВРЕМЕННЫХ ПОДХОДОВ К АВТОМАТИЗИРОВАННОМУ АНАЛИЗУ ТЕКСТОВЫХ ДОКУМЕНТОВ

1.1 Общие процедуры и основные задачи автоматизированного анализа текстовых документов

В соответствии с перечнем основных задач государственной программы «Цифровая экономика Российской Федерации», утвержденной распоряжением Председателя Правительства РФ от 28.07.17 №1632-р, особое внимание необходимо уделять оптимизации систем обработки и обмена информацией [1]. Известно, что значительная часть существующих и перспективных систем данного типа осуществляет информационный обмен с использованием электронных неструктурированных текстовых документов (ЭНТД), написанных на естественном языке (ЕЯ), где под ЕЯ понимается сформировавшийся способ обмена информацией в рамках речевой коммуникации.

В настоящее время мировая информационная среда и хранилища информации содержат очень большое количество ЭНТД различного типа, написанных на ЕЯ, которые являются источниками знаний [2] и данных в различных областях человеческой деятельности. При этом количество подобных ЭНТД с каждым днём возрастает, что определяет необходимость ускоренного развития информационных систем автоматизированного анализа указанных документов (ИСАА ЭНТД).

В то же время, функционал ИСАА ЭНТД часто ограничен отдельными предметными областями: системы работают с определенной группой понятий и являются с этой точки зрения «закрытыми» системами, в которые очень трудно внести какие либо изменения (число рубрик, состав тезауруса, важность слов).

Вопросами разработки алгоритмического обеспечения ИСАА ЭНТД занимаются такие науки, как компьютерная лингвистика (КЛ), вычислительная лингвистика (ВЛ) [3], алгебраическая лингвистика (АЛ), которые тесно связаны с более общей дисциплиной – прикладной лингвистикой (ПЛ).

Компьютерная лингвистика как наука включает следующие основные разделы [4]:

- теоретическая компьютерная лингвистика (КЛ-1) – содержит весь перечень задач лингвистики и обеспечивает формирование требований к степени формализации текстовых документов;
- инженерная компьютерная лингвистика (КЛ-2) – область знаний, связанная с инструментарием обработки, изучения и решения специфических задач анализа текстовых документов на ЕЯ. По составу используемых источников данных и используемым методам анализа КЛ-2 выходит за пределы КЛ-1, но существенно основывается на её основных моделях;
- инструментальная компьютерная лингвистика - является результатом интеграции методического обеспечения КЛ-1 и КЛ-2 с целью его реализации в рамках систем ИСАА ЭНТД с использованием новых вспомогательных элементов анализа: корпусов, парсеров, лингвистических ресурсов и т.д.

КЛ связана с такими областями наук, как:

- фонология – исследует правила формирования и соединения звуков в словах;
- морфология – исследует внутреннюю структуру речи, а также категории слов [5];
- синтаксис – исследует внутреннюю структуру предложений, правила сочетаемости, а также порядок следования слов в предложениях;
- семантика и прагматика – семантика занимается анализом смысловой нагрузки слов, предложений и других единиц речи, а прагматика – исследует особенности выражения смысла в связи с конкретными целями общения на ЕЯ [6];
- лексикография – исследует лексикон конкретного естественного языка – грамматические свойства отдельных слов и методы создания словарей [4, 7].

Указанные области КЛ соответствуют основным этапам анализа ЭНТД:

- лексический анализ текста – выделение слов, знаков препинания, цифр, и прочих текстовых единиц;
- морфологический анализ – определение грамматических характеристик лексем, а так же основных словоформ;
- синтаксический анализ – установление структуры предложения – системы связей между словами;
- семантический анализ – построение структуры, ассоциированной непосредственно с передаваемым значением в границах используемого языка ЭНТД [8, 9];
- прагматический анализ – интерпретация семантической структуры в контексте модели текста и знаний о предметной области.

Данные этапы анализа используются практически в любых алгоритмах, реализуемых в ИСАА ЭНТД [10-12]. Обобщенная процедура анализа ЭНТД приведена на рис. 1.1. В основу данной процедуры положен так называемый треугольник анализа [13], который модифицирован с учетом важности одного из результатов анализа – рубрицирования ЭНТД.

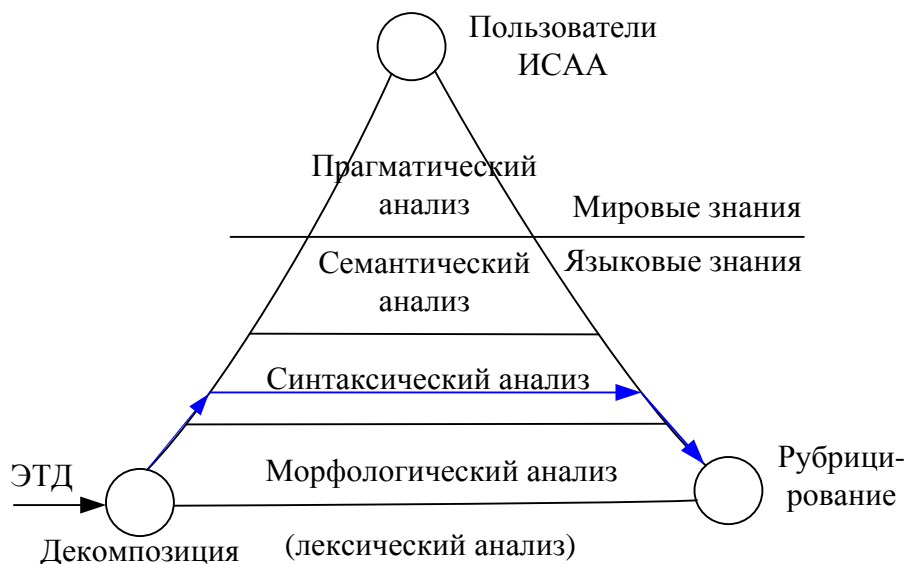


Рисунок 1.1 – Процедура анализа ЭНТД

Как видно из рисунка 1.1, перед началом анализа необходимо провести декомпозицию исследуемого ЭНТД на элементы с присвоением им некоторого

грамматического значения. Несмотря на то, что часто в качестве элементов ЭНТД рассматривают слова, можно оперировать и более детализированными элементарными единицами – знаками препинания, цифрами и т.п. Среди специфических проблем реализации этапа декомпозиции ЭНТД можно выделить следующие:

- в некоторых языках отсутствуют пробелы между словами (например, в китайском языке);
- написать правила отбора слов и словосочетаний в ряде случаев достаточно сложно;
- в тексте встречается большое количество символов и элементов, которые затрудняют использование лингвистического методического аппарата – телефонные номера, электронные адреса, ссылки на электронные ресурсы, формулы, смайлики, элементы таблиц и т.п.

Следующим этапом анализа ЭНТД является морфологический анализ – после выделения грамматических элементов необходимо определить для них статус в системе языка. Обычно для каждого слова находится морфема (т.е. форма, от которой произошло конкретное слово), которой приписывается грамматические характеристики (падеж, род, число и т.д.). Например, морфемой для существительного является именительный падеж, единственное число рассматриваемого слова.

Этап синтаксического анализа описывает связи слов в предложении, а также их синтаксические роли. В разных языках система синтаксических отношений создается разными средствами – вспомогательными словами, знаками препинания, пунктуацией или порядком слов.

Этап семантического анализа предполагает переход от непосредственно выделения синтаксических связей к их смысловой интерпретации, представленной некоторой семантической структурой. Обычно это формализованное представление ЭНТД соответствует информации из толкового словаря. Семан-

тический анализ применяется для более глубокого понимания ЭНТД и повышения точности методов их анализа, описанных в работе [13].

Этап прагматического анализа представляет собой этап интерпретации результатов «языкового» анализа применительно к практической деятельности пользователей ИСАА ЭНТД в контексте конкретной ситуации [14].

Очевидно, что с развитием средств вычислительной техники значительно расширяется спектр задач, решаемых ИСАА ЭНТД, которые можно условно разделить на четыре класса [14].

Первый класс задач включает задачи сбора информации и организации хранения большого числа полных текстовых документов в оригинальной форме.

Второй класс предполагает информационный поиск нужного ЭНТД в распределенных базах их хранения [15, 16].

Третий класс задач при использовании ИСАА ЭНТД связан, в том числе, с применением методов искусственного интеллекта, которые позволяют в результате обучения на некотором наборе ЭНТД генерировать решения по отношению документа к конкретному классу (рубрике) [17].

Четвертый класс связан с генерацией выходного текстового документа по результатам анализа (например, в рамках систем машинного перевода).

Для реализации перечисленных классов задач, связанных с анализом текстов, ИСАА ЭНТД должны реализовывать следующие функции [18, 19]:

- хранение объемных текстовых документов и реализация их «интеллектуального» поиска;
- автоматическое индексирование, рубрицирование и кластеризацию текстов по содержанию, установление сходства текстов;
- автоматическое аннотирование и реферирование ЭНТД;
- машинный перевод текста и речи [20];
- распознавание текста и речи;
- организация взаимодействия пользователя с компьютером ЕЯ;
- проверка правописания, грамматики и стиля;

- распознавание типа текстов (печатный, рукописный);
- поиск нужного документа по запросу (в т.ч. в сети интернет);
- автофильтрация (определение нежелательных документов: спам и т.п.);
- работа с электронными словарями;
- реализация вопросно-ответных процедур и процедур логического вывода;
- извлечение знаний (Text Mining, Information Retrieval), мнений (Opinion Mining, Sentiment Analysis) [14].

Автоматическое индексирование заключается в определении терминов, употребляемых в текстовом документе, нахождении их вариантов и родственных слов – их совокупность называется ключевыми словами или дескрипторами текстового документа. В отличие от ключевых слов, к одному дескриптору могут относиться слова, не являющиеся полностью синонимами, поэтому их не требуется различать в контексте выбранной предметной области. Последовательность сформированных дескрипторов определяет поисковый образ исходного ЭНТД, в котором отсутствуют семантические и синтаксические связи, поэтому данная структура достаточно плохо совместима с лингвистическими системами [21, 22].

Для получения более точной информации об ЭНТД в некоторых системах процедуру анализа дополняют поиском местоположения дескрипторов в предложениях, определением их ролей и весовых коэффициентов значимости. При этом иногда выделяют категории дескрипторов: агенты, качества, предметы, процессы [23].

Более сложные ИСАА ЭНТД анализируют семантические и синтаксические связи дескрипторов в предложениях, в том числе такие виды отношений между парами дескрипторов, как [20]:

- координативное – формальная связь;
- консекутивное – причинность или воздействие;
- ассоциативное – принадлежность в широком смысле;
- предикативное – отношение между предикатом и его актантами.

Отношения строятся также в зависимости от семантических категорий дескрипторов, которые они соединяют.

Достоинства автоматического реферирования заключаются в исключении субъективизма индексаторов; обеспечении стабильности результатов; упрощении обнаружения и исправления ошибок [20]. Автоматическое реферирование текстов приводит к формированию результирующих текстов рефератов трех типов: квазирефераты, рефераты-клише, рефераты. Под квазирефератами понимается последовательность отобранных из текста наиболее информативных предложений, которые часто представляют аннотацию или просто тематическое представление ЭНТД. Реферат-клише состоит тоже из наиболее информативных слов, которые подставляются в заранее созданные словесные шаблоны-клише. Создание обычного реферата является достаточно сложной задачей сжатия полного текста до его основного смыслового содержания. Данная процедура сжатия должна быть совместима со структурой документа и алгоритмом определения наиболее важных фрагментов текста [20].

С содержательной точки зрения значительная часть основных и вспомогательных задач, решаемых в ИСАА ЭНТД, могут быть отнесены к задачам классификации (рубрицирования) указанных документов. Задачу рубрицирования как разновидность задачи анализа ЭНТД можно сформулировать следующим образом [24].

Допустим, имеется множество объектов T и множество классов (рубрик ЭНТД) $C = \{c_i\}$ ($i = 1, \dots, N_c$), состоящее из N_c классов объектов. Каждый класс c_i представлен некоторым описанием F_i , имеющим определенную внутреннюю структуру. Процедура рубрицирования f заключается в выполнении преобразований над объектами $t \in T$, после которых либо делается вывод о принадлежности t к классу c_i , либо вывод о невозможности классификации t в существующих условиях.

В общем виде модель текстового рубрикатора можно представить в виде алгебраической системы вида:

$$R = \langle T, C, F, R_c, f \rangle,$$

где T – множество текстовых документов, C – множество рубрик или классов, F – множество описаний, R_c – отношение на множествах C и F , f – операция рубрицирования [25].

В общем случае решение задачи рубрицирования ЭНТД с использованием алгоритмов интеллектуального анализа данных предполагает обучение модели рубрикатора, которая подразумевает частичное или полное формирование C , F , R_c и f на основе некоторых априорных данных с заранее известными классами или рубриками [25].

1.2 Анализ современных методов автоматизированного рубрицирования текстовых документов

Учитывая важность и широкую распространенность задач рубрицирования в процедурах автоматизированного анализа ЭНТД, в настоящее время в ИСАА ЭНТД может использоваться достаточно развитый методический аппарат теории классификации объектов различного типа, при этом документ рассматривается как объект классификации (рубрицирования) [26].

На рисунке 1.2 приведены основные группы современных методов, которые могут использоваться в ИСАА ЭНТД для решения задач рубрицирования документов. На рисунке 1.2 с целью исключения повторений в изложении материала также выделены наиболее перспективные алгоритмы рубрицирования ЭНТД с учетом особенностей обращений граждан и организаций в органы исполнительной и законодательной власти различного уровня (обоснование указанного предложения будет приведено ниже).

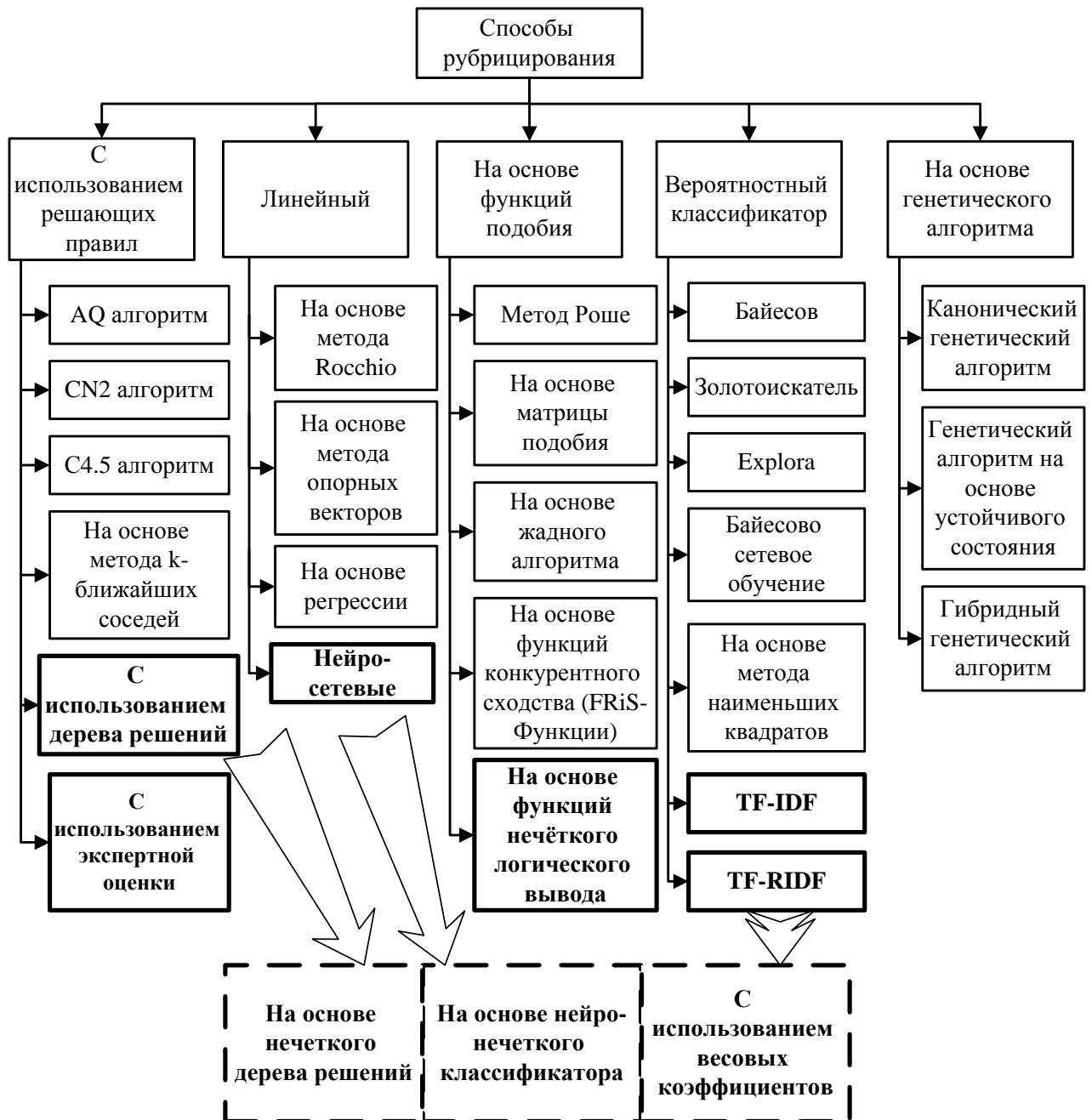


Рисунок 1.2 – Современные методы анализа ЭНТД

С точки зрения наличия этапа обучения модели для рубрицирования ЭНТД, все методы и алгоритмы можно условно разделить на две группы: с обучением и без обучения [27, 28].

Методы и алгоритмы рубрицирования ЭНТД с учителем предполагают наличие полной информации о рубричном поле (число и характеристики рубрик). В этом случае в качестве «учителя» используется некоторая подготовленная выборка ЭНТД (обучающее множество) с известными степенями принад-

лежности к той или иной рубрике. К подобным алгоритмам рубрицирования можно отнести следующие:

- алгоритм «наивной» байесовской классификации;
- алгоритм Роше;
- алгоритм k-ближайших соседей;
- алгоритм опорных векторов [29, 30];
- алгоритм деревьев принятия решений [31, 32];
- алгоритм наименьших квадратов.

В настоящее время методы и алгоритмы рубрицирования с учителем достаточно широко используются в ИСАА ЭНТД и подходят для решения большого количества возникающих задач. Однако необходимо учитывать, что данные методы требуют большого количества предварительной информации для обучения модели для рубрицирования.

Методы и алгоритмы классификации (рубрицирования) без учителя анализируют коллекцию ЭНТД с целью отнесения их к рубрикам таким образом, чтобы в каждую рубрику попали документы, наиболее близкие с точки зрения выбранной метрики. В данных методах и алгоритмах отсутствует необходимость в первоначальном обучении используемых моделей «с учителем», так как заранее неизвестны характеристики рубричного поля текстовых документов [33]. В общем случае алгоритмы рубрицирования без учителя (алгоритмы кластеризации) осуществляют группировку ЭНТД в пространстве признаков (параметров документа) с последующей интерпретацией результата. Классификация ЭНТД основывается на гипотезе, что близкие по смыслу текстовые документы релевантны одним и тем же запросам и выделенным рубрикам.

К алгоритмам кластеризации можно отнести следующие:

- алгоритм k-средних;
- плотностный алгоритм DBSCAN;
- нечёткий алгоритм c-средних;
- инкрементный алгоритм C2ICM;

- нейросетевой алгоритм SOM.

Для принятия решения о дальнейшем развитии существующих и разработке новых методов рубрицирования ЭНТД, особенности которых описаны во введении диссертации, представляется целесообразным более детальное рассмотрение отдельных наиболее часто используемых в настоящее время алгоритмов.

Наиболее популярными с точки зрения применения в ИСАА ЭНТД вероятностными алгоритмами рубрицирования являются Байесовы классификаторы [34], использующие процедуру f отнесения документа к рубрике на основе формулы Байеса для условной вероятности. Входной ЭНТД t представляется в виде последовательности терминов $\{w_k\}$, при этом каждая рубрика c_i характеризуется безусловной вероятностью отнесения к ней текстового документа $P(c_i)$, а так же условной вероятностью $P(w/c_i)$ встретить термин w в текстовом документе t при условии выбора рубрики c_i . В этом случае под вероятностью $P(t/c_i)$ понимается вероятность того, что ЭНТД будет классифицирован при условии выбора рубрики c_i .

В результате для оценки вероятности выбора рубрики c_i , при условии, что документ t пройдет успешное рубрицирование используется выражение вида [35]:

$$P(c_i | t) = \frac{P(c_i) \times P(t | c_i)}{\sum P(c_i) \times P(t | c_i)}. \quad (1.1)$$

Процедура рубрицирования ЭНТД f с использованием выражения (1.1) заключается в вычислении вероятностей $P(c_i/t)$ для всех рубрик c_i и выбора той рубрики, для которой данная вероятность максимальна. Обучение классификатора состоит в составлении словаря вероятностей всевозможных термов $\{w_n\}$ для каждой рубрики. Более подробно способы использования вероятностных алгоритмов классификации рассмотрены в работе [35].

Вероятностно-статистические классификаторы TF-IDF с выбором фраз рассмотрены в работе [36-38], в которой описаны проблемы кластеризации новостных статей, а также приведены результаты экспериментов [39].

Методам классификации на основе решающих правил посвящены работы [40].

Линейные классификаторы в большинстве случаев являются более предпочтительными в использовании благодаря своей компактности, высокой скорости обработки ЭНТД, возможности интерпретируемости результатов и т.д. Также для стандартных методов применения регрессионного анализа существуют возможности улучшения показателей качества с использованием трансдуктивного обучения [41]. Наиболее точным среди алгоритмов рубрицирования на основе регрессионных моделей являются модели на основе логистической регрессии. Оценка точности работы указанного алгоритма по сравнению с «наивным» байесовским алгоритмом приведена в работе [42].

В работах [43-46] рассмотрен широкий класс различных модификаций алгоритмов рубрицирования ЭНТД на основе решающих правил, применения искусственных нейронных сетей и вероятностно-статистических инструментов.

Анализ известных методов рубрицирования ЭНТД позволяет сделать вывод о существенной зависимости точности их применения от выбранного способа формализации указанных документов, которая в общем случае является важнейшей составляющей процедуры их предварительной обработки. На рисунке 1.3 представлены основные этапы предварительной обработки ЭНТД для дальнейшего рубрицирования, которые связаны с необходимыми способами формализованного представления документа [47-52].

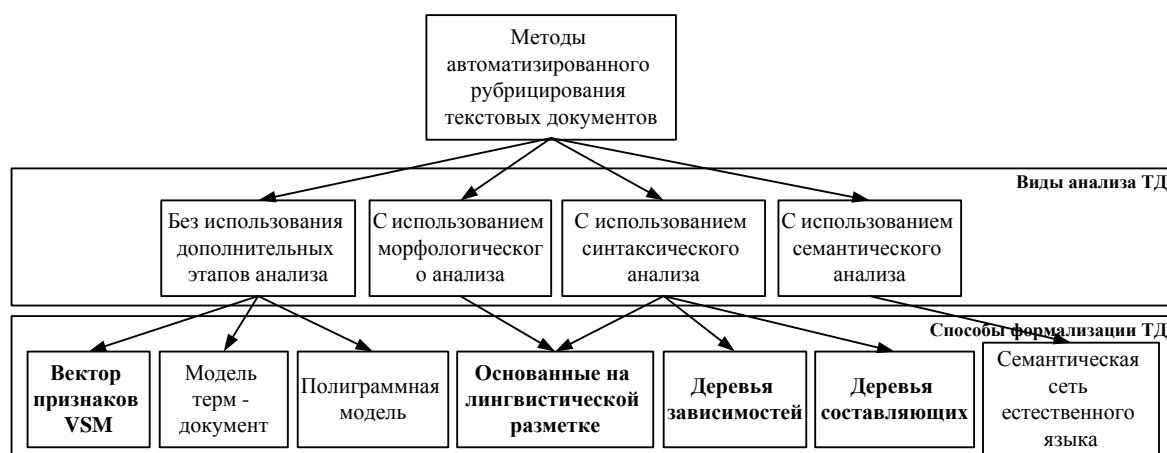


Рисунок 1.3 – Методы формализации ЭНТД для различных видов анализа текстов

Из рисунка 1.3 видно, что в методах, не использующих дополнительные этапы анализа, достаточно представить текстовый документ в виде моделей VSM, терм-документа или полиграммной структуры. При использовании морфологического или синтаксического анализов необходима лингвистическая разметка ЭНТД. Результаты этапа синтаксического анализа также можно представить в виде деревьев зависимостей и деревьев составляющих. Этап семантического анализа предполагает формирование семантической сети естественного языка. Более подробно способы представления текстовых документов с использованием процедур семантического анализа и поиска онтологий описаны в работах [53, 54]. В работах [55, 56] также рассмотрены способы описания структуры текстовых документов и выделение функциональных отношений между фрагментами [57].

Рассмотрим подробнее указанные способы формализации текстов с точки зрения возможности последующего автоматизированного ЭНТД указанного выше вида.

При использовании вектора признаков VSM (Vector Space Model) ЭНТД представляется в виде вектора, каждая координата которого соответствует частоте встречаемости одного из слов всей коллекции в этом тексте. Объединение всех таких векторов в единую таблицу приводит к формированию прямоугольной матрицы размером $n \times p$, где p – количество слов в коллекции (размерность пространства), а n – число документов.

Применение полиграммной модели со степенью n и основанием M предполагает представление ЭНТД в виде вектора $\{f_i\}$, $i=1, \dots, M^n$, где f_i – частота встречаемости i -й n -граммы в тексте, которая является последовательностью подряд идущих n – символов вида a_1, \dots, a_{n-1}, a_n , причем символы a_i принадлежат алфавиту, размер которого совпадает с M .

Модель терм-документ представляет модель, в рамках которой ЭНТД описывается лексическим вектором $\{\tau_i\}$ $i = 1, \dots, N_w$, где τ_i – важность (информационный вес) термина w_i в документе, N_w – полное количество терминов в

документной базе (словаре). Вес термина, отсутствующего в документе, принимается равным нулю.

Если в процессе рубрицирования применяются этапы синтаксического и семантического анализа, то для сохранения результатов предыдущих этапов, а также для записи новых характеристик лингвистических единиц необходимо представлять ЭНТД в более детальном виде. Так, для представления текстовых документов с синтаксическими характеристиками чаще всего используется лингвистическая разметка, которая предполагает задание информации о лингвистических единицах непосредственно в ЭНТД в форме разметки на специальном языке (например, SGML или XML).

Использование специальной разметки для представления лингвистической информации в ЭНТД является достаточно удобным для задач обработки текстов на естественном языке. Данный подход позволяет анализировать пользователем или разработчиком результаты рубрицирования ЭНТД с использованием стандартных программных инструментов. В то же время основной проблемой применения специальной разметки являются трудности при представлении сложных и пересекающихся структур ЭНТД, которые могут возникнуть вследствие неоднозначности анализа текста на одном из этапов обработки.

На рисунке 1.4 представлен пример грамматических зависимостей для предложения «Электрик разломал и без того плохо работающую розетку».

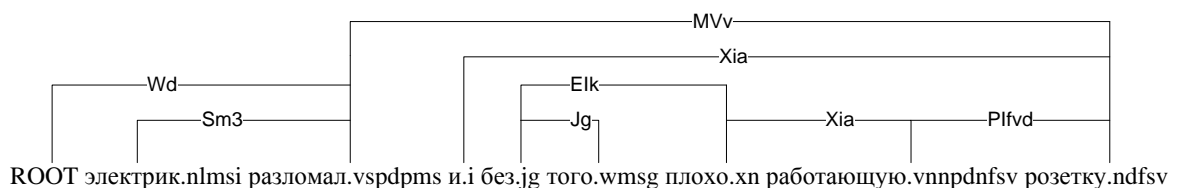


Рисунок 1.4 – Пример синтаксического дерева зависимостей для предложения

Данные зависимости предполагают возможность структурирования текста в виде древовидной структуры, в которые слова связаны ориентированными дугами, обозначающими синтаксическое подчинение между главным и зависимым словами.

Главной особенностью синтаксических деревьев зависимостей является отсутствие обозначающих составляющих нетерминальных вершин, при этом синтаксические связи имеют пометки, которые обозначают их тип. Типы связей определяют грамматические функции слов в предложении или общие семантические отношения между словами.

Пример дерева составляющих для предложения «Электрик не приходит на вызовы» приведен на рисунке 1.5.

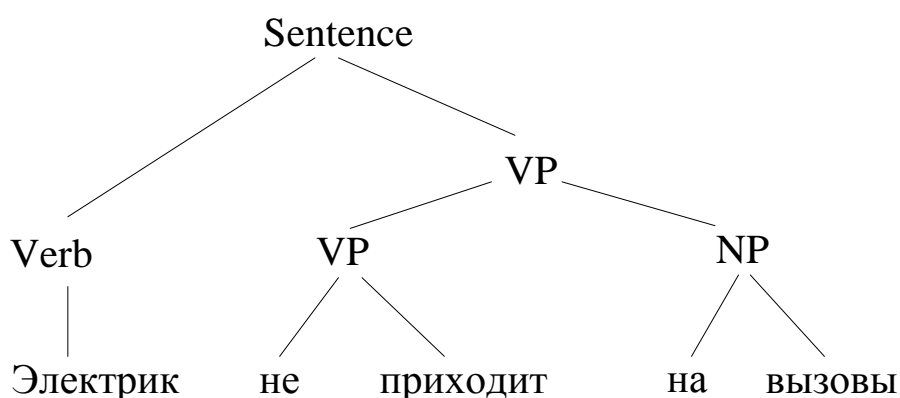


Рисунок 1.5 – Пример дерева составляющих для предложения

Дерево составляющих представляет собой модель формализации текста, в которой слова предложения на ЕЯ группируются в составляющие на основе лингвистических наблюдений того, что цепочки слов в предложении ЭНТД могут использоваться как единое целое и подчиняться единым грамматическим правилам. В этом случае составляющие можно перенести в середину или в конец предложения только целиком – их частичный перенос может привести к потере смысла.

1.3 Перспективы использования методов автоматизированного анализа текстов для рубрицирования электронных неструктурированных текстовых документов

Очевидно, что выбор того или иного алгоритма автоматизированного рубрицирования ЭНТД определяется типом данного документа. На рисунке 1.6

представлена возможность классификации рубрицируемых ЭНТД на основе четырех признаков: структурированности, объема, частоты встречаемости значимых слов и синтаксической связанности [58].

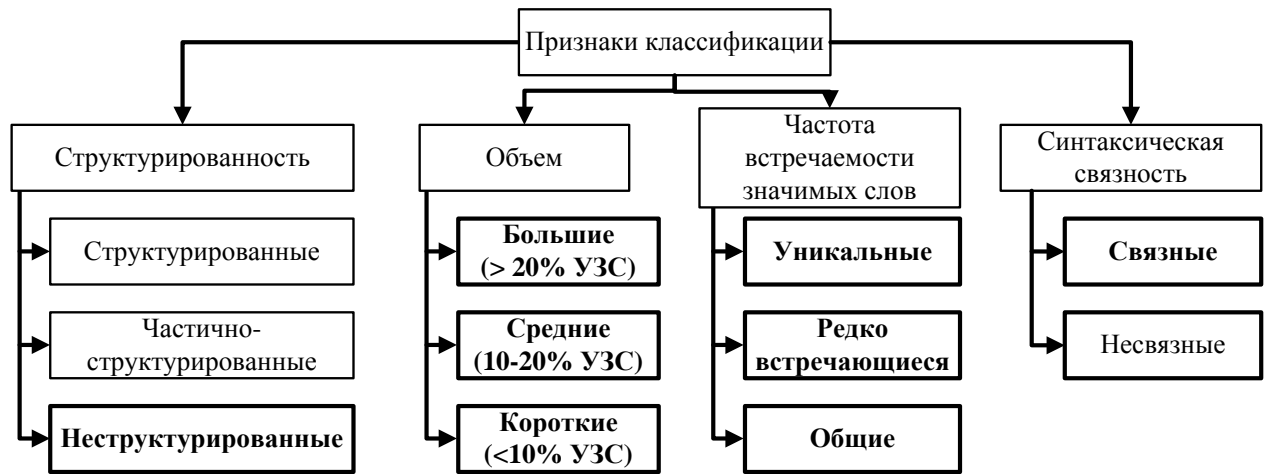


Рисунок 1.6 – Классификация ЭНТД

На рисунке 1.7 представлены примеры текстовых документов разной структурированности, а также характеристики темпов роста подобной информации [59].

Структурированные данные		Частично структурированные данные		Неструктурированные данные	
Реляционные БД		XML Docs	RSS Feeds	E-mail Web Logs	
Файлы табличных процессоров				Системы управления контентом Системы управления документами Таксономия Онтология	
Плоские файлы	Многомерные БД	Распознавание голоса			
Базы данных, доставшиеся в наследство: 1. иерархические БД; 2. БД на основе мэйнфреймов		Instant Messaging			
		Wikis			
Сокращается умеренно (с 15% до 46%)		Растет умеренно (с 18% до 47%)		Растет резко (с 61% до 81%)	

Рисунок 1.7 – Примеры текстовых документов различной степени структурированности

В настоящее время доля ЭНТД достаточно интенсивно увеличивается по отношению к частично структурированным и структурированным текстовым документам. К последним можно отнести ЭНТД в форматах xml, rdf и другим, подобных им. Эти форматы ориентированы, в основном, на описание данных.

К ЭНТД можно отнести документы в форматах html, doc, rtf и т.д. Обычно содержание подобных ЭНТД представляется пользователю для ознакомления с помощью обработки браузером Internet Explorer либо любыми текстовыми редакторами, такими как Microsoft Word и OpenOfficeWriter.

Признаки ЭНТД:

1. Текст написан на ЕЯ;
2. Отсутствует явно выделенное определение структуры;
3. Автоматическое выделение структур, как правило, не может быть выполнено однозначным образом.

Примерами типов неструктурированных текстовых документов могут являться: книги; журналы; метаданные; медицинские записи; аудио- и видеоматериалы; аналоговые данные; изображения; файлы, имеющие основой неструктурированный текст (сообщения электронной почты, веб-страницы, документы, созданные с помощью текстовых процессоров).

Говоря о слабоструктурированных ТД, можно выделить следующие их признаки и свойства:

- текстовый документ содержит некоторую разметку;
- содержимое текстового документа разбито внутренним форматированием;
- текстовый документ состоит из фрагментов, которые представляют собой либо некоторое значение, либо атрибут данных;
- во внутренней разметке документов нет формальных признаков, указывающих на то, что есть значение данных, а что есть атрибут данных.

Для такого рода документов будет наиболее эффективным их представление в виде совокупности объектов с последующей возможной идентификацией текстовых фрагментов как атрибутов и значений данных.

Примерами слабоструктурированных текстовых документов являются: анкеты, счета, налоговые декларации, страховые формы, прайс-листы, контракты, транспортные накладные, типовые договоры.

С точки зрения объема ЭНТД можно достаточно условно с учетом возможности применения тех или иных методов анализа разделить на очень короткие, короткие и длинные.

Коротким текстовым документом можно назвать ЭНТД документ, написанный на ЕЯ и содержащий информацию в лингвистической или цифровой форме, объем которого не позволяет применять известные процедуры статистического анализа текстов, но допускает использование для его анализа экспертной информации, полученной в результате комплексирования знаний лингвистов и специалистов в рассматриваемых предметных областях.

Очень короткие документы – это ЭНТД, объем содержащейся информации в которых не позволяет обеспечить автоматизацию процесса рубрицирования даже с использованием экспертных методов анализа.

Длинные документы – это ЭНТД, допускающие их автоматизированное рубрицирование с использованием вероятностно-статистических методов.

Если существует заранее определенный набор предметных областей, по которым будет осуществляться рубрицирование, а также тестовый набор размеченных документов для обучения алгоритмов, то можно выделить следующие три типа слов, встречающихся в ЭНТД: уникальные; редкие; общие.

Уникальные слова – слова, встречающиеся только в текстовых документах одной предметной области. *Редкие слова* – слова, встречающиеся в текстовых документах некоторой группы предметных областей, занимающихся смежной отраслью. *Общие слова* – слова, которые употребляются во всех или почти во всех предметных областях, и которые не несут в явном виде никакого признака предметной области.

В зависимости от строгости синтаксического построения ЭНТД разделяют на связные и несвязные документы.

На рисунке 1.6 показано (выделено дополнительно), что с точки зрения указанных выше особенностей, обращения граждан и организаций в электронном виде обычно являются неструктурированными и связными ЭНТД. С точки зрения объема и частоты встречаемости в тезаурусах рубрик ЭНТД указанные ЭНТД могут относиться к любой из выделенных групп.

На рисунке 1.8 представлены дополнительные признаки классификации, которые влияют на возможность применения различных моделей рубрицирования.

Для выбора модели рубрицирования целесообразно учитывать еще три дополнительных признака: характеристика рубричного поля, объем накопленной статистической информации и характеристика тезауруса рубрик.

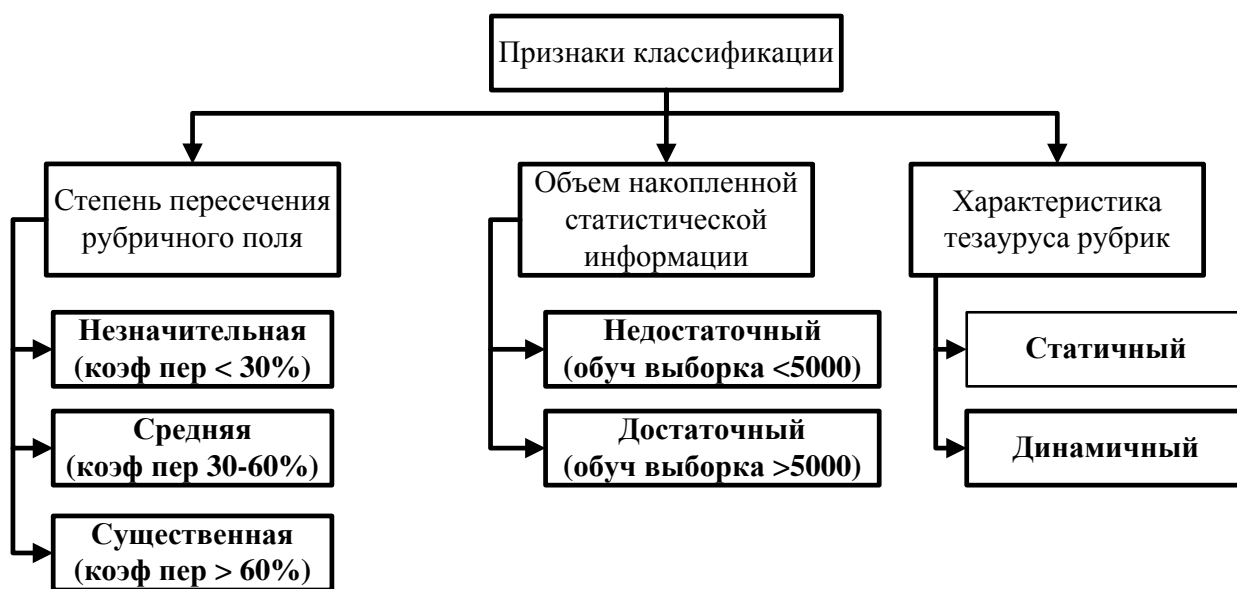


Рисунок 1.8 – Дополнительные признаки для рубрицирования ЭНТД

Характеристика рубричного поля показывает степень взаимосвязи их тезаурусов. Наиболее трудным условием рубрицирования является наличие взаимосвязанных рубрик.

Объем накопленной статистической информации определяет возможность построения тех или иных моделей рубрицирования. Так, при возможно-

сти корректного использования вероятностно-статистических методов объем ЭНТД можно считать «достаточным».

Характеристика тезауруса рубрик определяет, изменяется ли состав и степень влияния ЗС их словарей. Так, некоторые модели рубрицирования не могут применяться в условиях динамически изменяющихся тезаурусов из-за сложности переобучения и перестроения структуры модели рубрицирования [60-62].

В связи с особенностью рассматриваемых ЭНТД, а именно жалоб, обращений, предложений и т.п., поступающих на интернет-порталы органов исполнительной и законодательной власти (которые относятся к выделенным на рисунке 1.6 типам) и необходимостью их рубрицирования в особых условиях (выделены на рисунке 1.8), при разработке алгоритмического и программного обеспечения ИСАА ЭНТД целесообразно использовать несколько моделей рубрицирования в зависимости от характеристик конкретных ЭНТД. Так, для рубрицирования неструктурированных ЭНТД подходят следующие модели: вероятностные, на основе деревьев решений, на основе нейро-нечетких классификаторов и т.п. Объемные ЭНТД лучше всего анализировать с помощью вероятностных моделей, а короткие и очень короткие - нейро-нечетких классификаторов (рубрикаторов) или деревьев решений [63-67]. В условиях наличия взаимосвязанных рубрик лучше всего подходит модель рубрицирования на основе деревьев решений, а для невзаимосвязанных – нейро-нечеткий классификатор и вероятностные модели [68, 69]. Если объем накопленной статистической информации достаточный, то рациональней использовать вероятностные модели рубрицирования. В противном случае необходимо дополнять обучающую выборку экспертной информацией [70]. В условиях динамического тезауруса рубрик необходимы дополнительные процедуры изменения выбранных моделей рубрицирования, а также сам механизм идентификации предпосылок изменения рубричного поля [71].

Известные модели типа дерева решений предполагают многоуровневую структуру, включающую узлы (вершины), листья (самые нижние вершины) и

связи между ними. Узлы – некоторые атрибуты, связи – значения указанных атрибутов, листья – классы, на которые необходимо произвести разделение. Пример дерева решений, позволяющего ответить на вопрос: «хорошо ли отремонтирован фасад дома?», представлен на рисунке 1.9. Данный классификатор может разделить качество ремонта на два класса: П (положительный), О (отрицательный).

Процесс классификации начинается с корневого узла и движется к листьям, проверяя значения атрибутов вершин.

Как представляется, для применения модели с использованием дерева решений для рубрицирования ЭНТД рассматриваемых типов необходимо увеличить количество уровней для повышения точности рубрицирования, а также предложить нечеткие правила перехода и разработать способ формализации документов для адаптивного учета изменения рубричного поля [72-77].

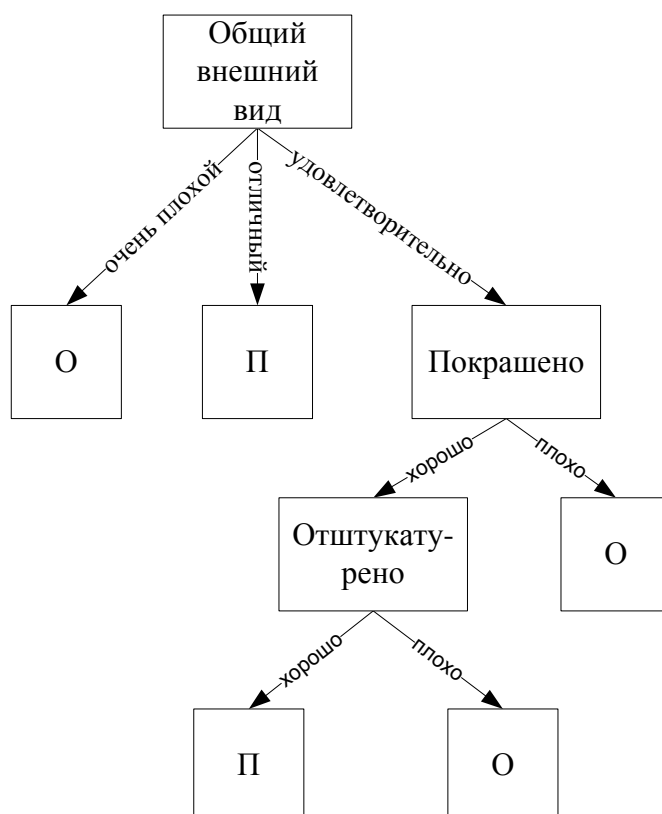


Рисунок 1.9 – Пример дерева решений, позволяющего ответить на вопрос: «хорошо ли отремонтирован фасад дома?»

Анализ показывает, что в условиях рубрицирования коротких ЭНТД при наличии достаточного объема статистической информации и незначительной степени пересечения тезаурусов рубрик целесообразно использовать нейросетевые алгоритмы, относящиеся в данном случае к области исследований, связанной с текстомайнингом (Text Mining) [78-80].

Несмотря на то, что эту область исследований часто называют текстовым дейтамайнингом (Text Data Mining), между ними есть принципиальное отличие – они работают с информацией разной степени структурированности. Text Data Mining позволяет извлекать новые знания (скрытые закономерности, факты, неизвестные взаимосвязи и т.п.) из больших объемов структурированной информации (хранимой в базах данных) [81], а текстомайнинг – находить новые знания в неструктурированных текстовых массивах [82].

Использование знаний и управление ими рассмотрено в работе [83, 84], в которой описаны не только основные определения и функции систем управления знаниями, но и перечислены реально существующие системы, позволяющие управлять знаниями [85, 86].

В этом смысле текстомайнинг добавляет к технологии дейтамайнинга дополнительный этап преобразования неструктурированных текстовых документов в структурированные массивы информации, что позволяет применять стандартные методы дейтамайнинга для дальнейшей обработки.

Наиболее простой задачей является текстомайнинг слабоструктурированных узкоспециализированных текстовых массивов (различные отчеты о поломках, результаты опросов и т.п.). В больших массивах текстовых документов, в которых набор лексики ограничен, новая информация достаточно точно извлекается на основе использования статистики по ключевым словам. В работе [87] рассмотрен метод на основе анализа ключевых слов для кластеризации текстовых документов. Если речь идёт о неструктурированных текстовых документах, то необходимо реализовывать процедуры «понимания» произвольных текстов на естественном языке. Даная задача является одной из «старейших» задач искусственного интеллекта (ИИ) [88-90], которая может быть решена несколькими

способами. Примерами указанных способов могут являться методы обработки данных на естественном языке — NLP (Natural Language Processing), нейросетевые методы и т.д. В работах [91-93] рассмотрены модели нейросетевых классификаторов со способами формализации текстовых документов, а также представления результатов работы классификатора в виде семантических образов.

Например, для рубрицирования ЭНТД может быть использован классификатор Гроссберга (ART). Сеть ART [94] состоит из двух слоев нейронов: сравнивающего и распознающего. В общем случае между слоями существуют прямые связи с весами w_{ij} от i -го нейрона входного слоя к j -му нейрону распознающего слоя, обратные связи с весами v_{ij} — от i -го нейрона распознающего слоя к j -му нейрону входного слоя. Также существуют тормозящие связи между нейронами распознающего слоя. Каждый нейрон распознающего слоя отвечает за один класс объектов. Веса w_{ij} используются на первом шаге классификации для выявления наиболее подходящего нейрона — класса, веса обратных связей v_{ij} хранят типичные образы соответствующих классов и используются для непосредственного сопоставления с входным вектором.

Процедура классификации может быть представлена следующими этапами:

Этап 1. Вектор X подается на вход сети и для каждого нейрона распознающего слоя определяется взвешенная сумма его входов по формуле вида:

$$|W| \times X \rightarrow Y.$$

Этап 2. За счет латеральных тормозящих связей распознающего слоя на его выходах устанавливается единственный сигнал с наибольшим значением, который определяется по формуле вида:

$$y_i = \max(Y),$$

а остальные выходы становятся близкими к 0.

Этап 3. Входной вектор X сравнивается с прототипом класса V_i :

$$S_i(X, V_i) \geq p \begin{cases} да \rightarrow c_{out} = c_i \\ нет \rightarrow y_i = 0 \end{cases},$$

где первое условие означает, что если результат сравнения превышает порог p ,

то делается вывод о том, что входной вектор принадлежит классу c_i , в противном случае выход данного нейрона обнуляется и повторяется шаг 2, в котором за счет обнуления самого активного нейрона происходит выбор нового.

Этап 4. Этапы 2 и 3 повторяются до тех пор, пока не будет получен класс c_{out} , либо пока не будут принудительно заблокированы все нейроны распознающего слоя.

При использовании нейросетевых классификаторов ЭНТД представляется в виде огромного массива бинарных значений, каждое из которых соответствует наличию или отсутствию всех слов из тезауруса всего рубричного поля [53]. Такое представление ЭНТД делает нерациональным применения данной модели рубрицирования в условиях динамически изменяющихся тезаурусов рубрик [95], в связи со сложностью перестроения нейро-нечеткой сети и способа формализации ЭНТД каждый раз при изменении состава рубрик. Следовательно, необходимо разработать способ формализации ЭНТД, который позволит использовать нейро-нечеткий классификатор в условиях динамического тезауруса рубрик, а также необходимо саму модель представить в виде каскада для удобного перестроения всей модели рубрицирования при изменении рубричного поля [96-102].

Огромное количество информации скапливается в многочисленных текстовых базах [103], хранящихся в личных ПК, локальных и глобальных сетях. И объем этой информации стремительно увеличивается. Чтение объемных текстов и поиск в гигантских массивах текстовых данных малоэффективны, поэтому становятся все более востребованными решения текстомайнинга. Поиску информации на основе нечетких методов кластеризации посвящена работа [104, 105].

В процессе использования известных программных продуктов для проведения дополнительных этапов анализа придётся столкнуться с проблемой разнообразия лингвистических разметок. Например, большинство синтаксических парсеров на выходе представляет каждое предложение текста в виде деревьев зависимостей, которые описывают лингвистической разметкой. Лингвистиче-

скую разметку для дальнейшей классификации и назначения весовых коэффициентов необходимо модифицировать, тем самым увеличивая размерность метрики [106, 107].

Например, при использовании синтаксического парсера LinkGrammar при анализе предложения: «Состояние труб водоснабжения очень плохое» получаем лингвистическую разметку вида:

«("LEFT-WALL" RW:6:RIGHT-WALL Wd:1:состояние.ndnsi)(Wd:0:LEFT-WALL "состояние.ndnsi" Mg:2:труб.ndfpg)(Mg:1:состояние.ndnsi "труб.ndfpg" Mg:3:водоснабжения.ndnsg)(Mg:2:труб.ndfpg "водоснабжения.ndnsg")("[очень]")("[плохое"])(RW:0:LEFT-WALL "RIGHT-WALL"))».

Дерево зависимостей данного предложения выглядит следующим образом:

```
+-----Wd-----+-----Mg-----+-----Mg-----+
|           | | |
LEFT-WALL состояние.ndnsi труб.ndfpg водоснабжения.ndnsg      [очень]
[плохое]
```

Синтаксический парсер MalpParser вместо подобной XML разметки использует построчное разбиение предложений на слова, которым приписываются характеристики в порядке, описанном в XML файле конфигурации, который выглядит следующим образом:

```
<?xml version="1.0" encoding="UTF-8"?>
<dataformat name="conllx">
  <column name="ID" category="INPUT" type="INTEGER"/>
  <column name="FORM" category="INPUT" type="STRING"/>
  <column name="LEMMA" category="INPUT" type="STRING"/>
  <column name="CPOSTAG" category="INPUT" type="STRING"/>
  <column name="POSTAG" category="INPUT" type="STRING"/>
  <column name="FEATS" category="INPUT" type="STRING"/>
  <column name="HEAD" category="HEAD" type="INTEGER"/>
  <column name="DEPREL" category="DEPENDENCY_EDGE_LABEL"
type="STRING"/>
```

```
<column name="PHEAD" category="IGNORE" type="INTEGER" default="_"/>
```

```
<column name="PDEPREL" category="IGNORE" type="STRING" default="_"/>
```

```
</dataformat>
```

В этом случае для использования сторонних систем анализа необходимо написать программы преобразования форматов представления текстовой информации, также модифицировать предложенные форматы для внесения дополнительных характеристик, необходимых для модифицируемых методов классификации текстовых документов. Например, для жалобы «Налоговая инспекция продолжает уже два месяца кошмарить нашу фирму» синтаксический парсер LinkGrammar представит его в лингвистической разметке вида:

```
(( "LEFT-WALL" RW:10:RIGHT-WALL Wd:3:продолжает.vnpdn3s )(
"налоговая.afsi" Afi:2:инспекция.ndfsi )(Afi:1:налоговая.afsi "инспекция.ndfsi"
Sf3:3:продолжает.vnpdn3s )(Wd:0:LEFT-WALL Sf3:2:инспекция.ndfsi "продол-
жает.vnpdn3s" I:7:кошмарить.vsndi E:4:уже.as )(E:3:продолжает.vnpdn3s
"уже.as" )( "два" IDBBT:6:месяца )(IDBBT:5:два "месяца" EI:7:кошмарить.vsndi
)(I:3:продолжает.vnpdn3s EI:6:месяца "кошмарить.vsndi" MVv:9:фирму.ndfsv )(
"нашу.wfsv" Afv:9:фирму.ndfsv )(MVv:7:кошмарить.vsndi Afv:8:нашу.wfsv
"фирму.ndfsv" )(RW:0:LEFT-WALL "RIGHT-WALL" )).
```

Из представленных на рисунке 1.2 типов методов для анализа коротких неструктурированных документов при наличии динамического изменения характеристик рубрик в наибольшей степени представляются перспективными методы на основе нечетких нейросетей, нечетких деревьев решений, вероятностные методы с использованием экспертных оценок.

1.4 Выводы по главе

Исследование общих процедур и основных задач анализа текстовых документов позволяет сделать выводы, что электронные сообщения граждан (жа-

лобы, обращения, предложения и т.д.) с точки зрения возможности их автоматизированной обработки обладают рядом специфических особенностей: в значительной части случаев небольшой объем документа, что затрудняет его статистический анализ; отсутствие структуризации, что усложняет процедуры извлечения информации; наличие большого количества грамматических и синтаксических ошибок приводит к необходимости реализации нескольких дополнительных этапов обработки; нестационарность тезауруса (состава и важности слов), который зависит от выхода новых нормативных документов, выступлений должностных лиц и политических деятелей и т.д., приводит к необходимости использования процедур динамической кластеризации рубрик.

Анализ методов рубрицирования текстовых документов показал, что особенности текстовых документов накладывают определенные ограничения на алгоритмы применения морфологического, синтаксического и семантического анализов [108-110], а также на соответствующие им процедуры формализации для автоматизированной обработки текстов, в том числе в рамках виртуальных систем информационного обеспечения различных региональных социально-экономических процессов. Результаты указанного анализа позволили выделить группы методов рубрицирования электронных текстов, относящихся к различным классам.

Анализ перспектив использования методов автоматизированного анализа текстов для рубрицирования электронных неструктурированных документов показал, что для решения указанной задачи целесообразно реализовывать мультимодельный подход, обеспечивающий комплексную реализацию алгоритмов динамической кластеризации для мониторинга состава и характеристик рубрик, с последующим рубрицированием документов с помощью алгоритмов интеллектуального анализа информации на основе инструментов теории нечетких множеств и искусственных нейронных сетей.

2 РАЗРАБОТКА МЕТОДОВ И МОДЕЛЕЙ АНАЛИЗА ЭЛЕКТРОННЫХ НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ И МОНИТОРИНГА РУБРИК

2.1 Мультимодельный метод анализа и рубрицирования электронных неструктурированных текстовых документов

С учетом сформулированных в разделе 1.2 особенностей неструктурированных текстовых сообщений, поступающих в электронном виде в органы законодательной и исполнительной власти субъектов РФ, определены следующие признаки, на основе которых выделены типовые ситуации рубрицирования этих ЭНТД: размер ЭНТД, степень пересечения тезаурусов рубрик, достаточность статистических данных для эффективного использования вероятностных моделей анализа ЭНТД.

В работе предлагается выделить три типоразмера ЭНТД: короткие (содержат менее 10% УЗС), средние (10-20% УЗС), большие (более 20% УЗС). Из числа значимых слов для рубрицирования ЭНТД исключены различные числовые значения, союзы, даты, URL ссылки и т.п. Конкретные значения размеров ЭНТД уточняются экспертами при настройке системы автоматизированного рубрицирования ЭНТД.

Степень пересечения тезаурусов рубрик K_{nep} зависит от числа уникальных слов для всех рубрик и определяется следующим образом:

$$K_{nep} = \frac{1}{J} \cdot \sum_{j=1}^J \sum_{i=1}^J \frac{Count(R_i \cap R_j)}{M_j},$$

где J – общее количество рубрик, $Count(R_i \cap R_j)$ – количество совпадающих значимых слов в тезаурусах рубрик R_i и R_j , M_j – общее число значимых слов в тезаурусе рубрики R_j .

По результатам проведенных исследований приняты следующие критериальные уровни пересечения тезаурусов: $K_{пер} < 0,15$ – незначительный уровень; $0,15 \leq K_{пер} < 0,4$ – средний уровень; $K_{пер} > 0,4$ – существенный уровень.

Процедура определения достаточности статистических данных для эффективного использования вероятностных моделей анализа ЭНТД приведена в подразделе 4.3.

В соответствии с рассмотренными признаками ЭНТД выделены описанные ниже типовые ситуации их рубрицирования.

Типовая ситуация 1. Выполняется анализ коротких или средних по размеру ЭНТД, в условиях средней или существенной степени пересечения рубрик, а также при недостаточности статистических данных для эффективного использования традиционных математических моделей для рубрицирования ЭНТД. При этом даже использование дополнительной экспертной информации не позволяет обеспечить требуемую точность рубрицирования. В этом случае ЭНТД должен быть передан специалистам для его ручного рубрицирования.

Типовая ситуация 2. Выполняется анализ средних по размеру ЭНТД, в условиях незначительной или средней степени пересечения рубрик, а также при недостаточности статистических данных для эффективного использования традиционных математических моделей для рубрицирования ЭНТД. В этом случае целесообразно применять модель с использованием весовых коэффициентов, позволяющую учитывать кроме статистических характеристик еще и экспертную информацию в процессе рубрицирования ЭНТД. Указанная модель, в отличие от известных [11], должна учитывать степень значимости слов в ЭНТД в зависимости от появления новых значимых событий, оказывающих влияние на тезаурусы рубрик.

Типовая ситуация 3. Выполняется анализ средних по размеру ЭНТД, в условиях средней или значительной степени пересечения рубрик, а также при недостаточности статистических данных для эффективного использования традиционных математических моделей для рубрицирования ЭНТД. Эта ситуация,

которая определяет целесообразность применения модели выбора рубрики на основе нечеткого дерева решений [111], что при использовании дополнительной экспертной информации позволит повысить точность и оперативность рубрицирования. Данная модель, в отличие от существующих моделей данного типа [34, 111], должна использовать синтаксические связи и роли слов, а также нечеткую оценку различий между документами в n -мерном пространстве признаков текстов при построении и использовании нечеткого дерева решений для отнесения документа к конкретным рубрикам в условиях взаимозависимости их тезаурусов.

Типовая ситуация 4. Выполняется анализ коротких ЭНТД при незначительной степени пересечения рубрик, в условиях достаточного объема и качества статистических данных о документах данного типа (для обучения гибридных нечетких моделей [92]). Для этой ситуации целесообразно использовать нейро-нечеткий классификатор на основе известного подхода [25, 91], но с использованием экспертной информации для определения значимости ключевых слов при формализации и последующего рубрицирования ЭНТД.

Типовая ситуация 5 характеризуется наличием больших рубрицируемых ЭНТД, достаточного объема статистических данных о документах данного типа и незначительной степенью пересечения рубрик. В этой ситуации целесообразно использовать известные вероятностные методы анализа текстовой информации [11, 56].

Типовая ситуация 6. Выполняется анализ больших ЭНТД при средней или существенной степени пересечения рубрик, в условиях достаточного объема и качества статистических данных о документах указанного типа. В этом случае целесообразно использовать метод голосования с учетом всех указанных выше моделей [112], что позволит снизить вероятность ошибок при рубрикации ЭНТД.

В таблице 2.1 представлено краткое описание типовых ситуаций рубрицирования ЭНТД, выделенных с учетом размера ЭНТД, степени пересечения рубрик, достаточности статистических данных для эффективного использования вероятностных моделей анализа ЭНТД, а также указано соответствие этих ситуаций предлагаемым моделям для эффективного рубрицирования.

Таблица 2.1 – Соответствие типовых ситуаций и предлагаемых моделей для эффективного рубрицирования ЭНТД

Типовая ситуация	Размер ЭНТД	Степень пересечения рубрик	Достаточность статистических данных	Тип модели для рубрицирования ЭНТД
1	короткий, средний	средняя, существенная	недостаточно	ручное рубрицирование
2	средний	незначительная, средняя	недостаточно	модель с использованием весовых коэффициентов
3	средний	средняя, существенная	недостаточно	нечеткое дерево решений
4	короткий	незначительная	достаточно	нейро-нечеткий классификатор
5	большой	незначительная	достаточно	вероятностный классификатор
6	большой	средняя, существенная	достаточно	метод голосования

Анализ этих типовых ситуаций показывает невозможность применения для автоматизированного рубрицирования ЭНТД единой модели и позволяет обосновать целесообразность создания мультимодельного метода для решения данной задачи. Выбор же конкретной модели при рубрицировании ЭНТД должен осуществляться по результатам идентификации соответствующей типовой ситуации. Алгоритмы построения и использования этих моделей подробно рассмотрены в подразделах 2.3, 2.4, 3.1–3.3.

На рисунке 2.1 приведена схема мультимодельного метода анализа и рубрицирования ЭНТД на основе комбинированного использования нечетко-логических, нейро-нечетких и вероятностных моделей. При этом выбор конкретной модели или их ансамбля осуществляется с помощью базы нечетких продукционных правил.

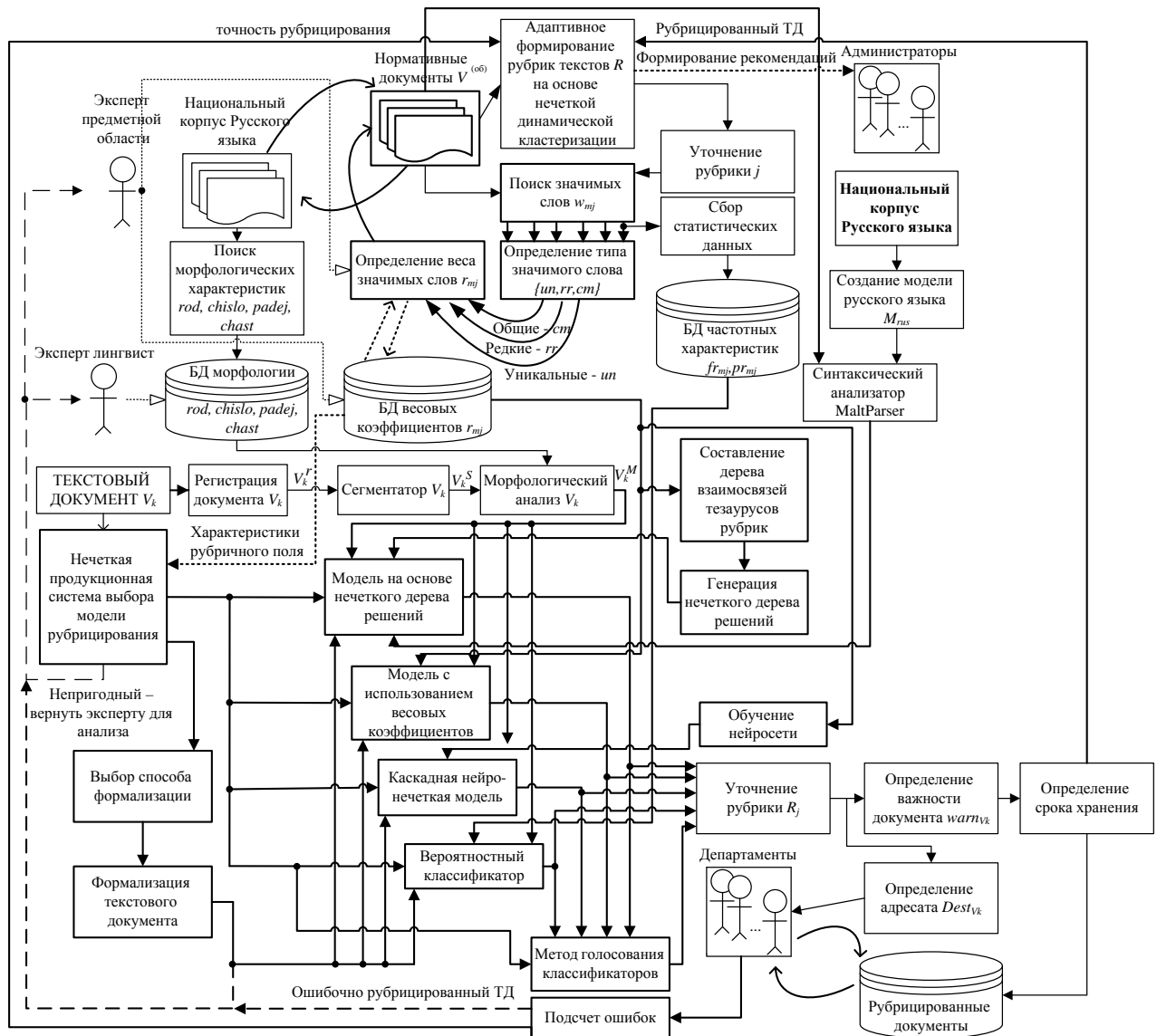


Рисунок 2.1 – Схема мультимодельного метода анализа и рубрицирования ЭНТД

В рамках реализации указанного метода предполагается анализ множества электронных неструктурированных текстовых документов:

$$V = \{V_1, \dots, V_k, \dots, V_K\}, \quad (2.1)$$

в котором каждый документ V_k представляется в виде набора значимых слов $V_k = \{v_1^{(k)}, \dots, v_{l_k}^{(k)}, \dots, v_{L_k}^{(k)}\}, k = 1, \dots, K$, $v_{l_k}^{(k)}$ – слово ЭНТД, $l_k = 1, \dots, L_k$, L_k – количество слов в k -м ЭНТД.

На этапе предварительного анализа каждый документ V_k проходит регистрацию и представляется в виде лингвистической разметки XML с использо-

ванием совокупности тегов (специальных меток, дескрипторов [113]). В результате ЭНТД начинается и закрывается тегом $\langle \text{XML} \rangle$ и $\langle / \text{XML} \rangle$, в который заносится тег заголовка $\langle \text{HEAD} \rangle$ с параметром уникального номера текстового документа $\langle \text{number_doc} \rangle DN_k \langle / \text{number_doc} \rangle$ и временем поступления документа $\langle \text{datetimt_doc} \rangle DT_k \langle / \text{datetime_doc} \rangle$. Заголовочная часть заканчивается тегом $\langle / \text{HEAD} \rangle$, а вся текстовая часть – $\langle \text{TEXT} \rangle$ и $\langle / \text{TEXT} \rangle$. После регистрации анализируется модифицированный текстовый документ $V_k^r = V_k \cup h^{(k)}$, где $h^{(k)}$ – заголовочная информация k -го документа, описанная по изложенным выше правилам.

ЭНТД V_k^r анализируется с целью выделения в его текстовой части в форме лингвистической разметки слов, предложений, абзацев и т.д. (на рисунке 2.1. блок «Сегментатор») [20].

В результате на выходе блока «Сегментатор» формируется модифицированный ЭНТД вида:

$$V_k^S = V_k \cup h^{(k)} \cup Sen^{(k)} \cup Abz^{(k)},$$

где $Sen^{(k)}$ – множество предложений ЭНТД, $Abz^{(k)}$ – множество абзацев, при этом полученная информация сохраняется в том же текстовом документе, но под отдельными тегами XML лингвистической разметки.

Следует отметить сложность реализации процедуры сегментации, а также высокую степень влияния ошибок на итоговые результаты автоматизированного рубрицирования ЭНТД. В первую очередь это характерно для процедур рубрицирования коротких ЭНТД.

После лингвистической разметки осуществляется морфологический анализ ЭНТД V_k^S с выделением лексических характеристик слов и морфем (наименьших имеющих смысл языковых единиц [20]).

Процедура морфологического анализа ставит в соответствие размеченные тегом $\langle \text{WORD} \rangle$ слова и такие морфологические характеристики, как:

1. падеж– $\text{padej} \in \{\text{nom}, \text{gen}, \text{dat}, \text{dat2}, \text{acc}, \text{ins}, \text{loc}, \text{gen2}, \text{acc2}, \text{loc2}, \text{voc}, \text{adnum}\}$,

2. число – $chislo \in \{sg, pl\}$,

3. род – $rod \in \{m, f, m-f, n\}$,

4. часть речи – $chast \in \{S, A, NUM, ANUM, V, ADV, PRAEDIC, PARENTH, SPRO, APRO, ADVPRO, PRAEDICPRO, PR, CONJ, PART, INTJ\}$,
 где nom – именительный падеж (труба, подъезд, чиновник, электрик), gen – родительный падеж (трубы, подъезда, чиновника, электрика), dat – дательный падеж (трубе, подъезду, чиновнику, электрику), $dat2$ – дистрибутивный дательный), acc – винительный падеж (трубу, подъезд, чиновника, электрика), ins – творительный падеж (трубой, подъездом, чиновником, электриком), loc – предложный падеж (трубе, подъезде, чиновнике, электрике), $gen2$ – второй родительный падеж, $acc2$ – второй винительный падеж, $loc2$ – второй предложный падеж, voc – звательная форма, $adnum$ – счётная форма (два киловатта, шестьдесят литров воды), sg – единственное число (водопроводчик, закон), pl – множественное число (водопроводчики, законы), m – мужской род (работник, стол), f – женский род (монетизация, работница, вода), $m-f$ – «общий род» (пьяница), n – средний род (электричество, озеро), S – существительное (пострадавший, кооператив, корпус, подвал), A – прилагательное (протекающий, мокрый, забитый, гнилой), NUM – числительное (четыре, десять, много), $ANUM$ – числительное-прилагательное (первый, седьмой, восьмидесятый), V – глагол (пользоваться, обрабатывать), ADV – наречие (сгоряча, очень), $PRAEDIC$ – предикатив (жаль, хорошо, пора), $PARENTH$ – вводное слово (кстати, по-моему), $SPRO$ – местоимение-существительное (она, что), $APRO$ – местоимение-прилагательное (который, твой), $ADVPRO$ – местоименное наречие (где, вот), $PRAEDICPRO$ – местоимение-предикатив (некого, нечего), PR – предлог (под, напротив), $CONJ$ – союз (и, чтобы), $PART$ – частица (бы, же, пусть), $INTJ$ – междометие (увы, ба-тюшки) [114].

В результате морфологического анализа формируется модифицированный ЭНТД вида:

$$V_k^M = V_k \cup h^{(k)} \cup M^{(k)},$$

где $M^{(k)}$ – результаты разбиения k -го документа на слова, предложения, абзацы, выделение структурных единиц, а также приписанные словам морфологические характеристики по описанным выше правилам.

Стандартные морфологические характеристики по всему тезаурусу хранятся в морфологической базе, которая заполняется из нескольких источников, важнейшим из которых является набор текстовых документов СинТагРус (Национальный корпус русского языка, содержащий размеченные тексты), а также другие текстовые словари [114]. Информация в СинТагРус размечена с использованием формата XML с расширением .tgt, поэтому поиск и запись морфологий слов осуществляется путем поиска нужных тегов и их параметров: тег <W> обозначает слово и содержит в себе форму слова в конкретном предложении, данный тег содержит параметры FEAT – морфологические признаки слова и LEMMA – начальная форма слова, а также LINK – часть речи.

Результаты морфологического анализа ЭНТД также сохраняются в виде отдельной лингвистической группы в исходном документе. Это означает, что помимо морфологической информации, приписанной каждому слову текста, для каждого предложения задана и его синтаксическая структура.

Важнейшим этапом реализации рассматриваемого мультимодельного метода является формирование и мониторинг рубричного поля, который подробно описан в подразделе 2.2 и приводит к описанию элементов множества рубрик:

$$R = \{R_1, \dots, R_j, \dots, R_J\}, \quad (2.2)$$

где $R_j = \{(w_1^{(j)}, r_1^{(j)}, fr_1^{(j)}, pr_1^{(j)}), \dots, (w_{m_j}^{(j)}, r_{m_j}^{(j)}, fr_{m_j}^{(j)}, pr_{m_j}^{(j)}), \dots, (w_{M_j}^{(j)}, r_{M_j}^{(j)}, fr_{M_j}^{(j)}, pr_{M_j}^{(j)})\}$, $j = 1, \dots, J$, $w_{m_j}^{(j)}$ – m_j -е слово в рубрике R_j , $m_j = 1, \dots, M_j$, $r_{m_j}^{(j)} \in [0, 1]$ – степень соответствия m_j -го слова j -й рубрике, $fr_{m_j}^{(j)}$ – частота встречаемости m_j -го слова j -й рубрике, $pr_{m_j}^{(j)}$ – порог употребления m_j -го слова j -й рубрике.

После формирования множества рубрик R осуществляется выбор конкретной модели для рубрицирования ЭНТД на основе выявления типовых ситуаций, приведенных в таблице 2.1.

Учитывая, что используемые в данной таблице значения характеристик ЭНТД (размер k -го ЭНТД $L_k = \{L_k^s, L_k^m, L_k^b\}$ рубричного поля; степень пересечения тезаурусов рубрик $Knep = \{Knep^s, Knep^l\}$; объем накопленных статистических данных $Vstat = \{Vstat^s, Vstat^l\}$) определяются с использованием экспертной информации, для реализации процедуры выбора конкретной модели для автоматизированного рубрицирования из множества моделей M_r предлагается использовать базу нечетких продукционных правил вида:

$$\begin{aligned}
 &\text{П1: ЕСЛИ } (<Vstat> \text{ есть } <Vstat^s>) \text{ И } (<Knep> \text{ есть } <Knep^l>) \text{ И} \\
 &(<L_k> \text{ есть } <L_k^s>) \text{ ТОГДА } <Mr> = <M_1>, \\
 &\text{П2: ЕСЛИ } (<Vstat> \text{ есть } <Vstat^s>) \text{ И } (<Knep> \text{ есть } <Knep^s>) \text{ И} \\
 &(<L_k> \text{ есть } <L_k^m>) \text{ ТОГДА } <Mr> = <M_2>, \\
 &\text{П3: ЕСЛИ } (<Vstat> \text{ есть } <Vstat^s>) \text{ И } (<Knep> \text{ есть } <Knep^l>) \text{ И} \\
 &(<L_k> \text{ есть } <L_k^m>) \text{ ТОГДА } <Mr> = <M_3>, \\
 &\text{П4: ЕСЛИ } (<Vstat> \text{ есть } <Vstat^l>) \text{ И } (<Knep> \text{ есть } <Knep^s>) \text{ И} \\
 &(<L_k> \text{ есть } <L_k^s>) \text{ ТОГДА } <Mr> = <M_4>, \\
 &\text{П5: ЕСЛИ } (<Vstat> \text{ есть } <Vstat^l>) \text{ И } (<Knep> \text{ есть } <Knep^s>) \text{ И} \\
 &(<L_k> \text{ есть } <L_k^l>) \text{ ТОГДА } <Mr> = <M_5>, \\
 &\text{П6: ЕСЛИ } (<Vstat> \text{ есть } <Vstat^l>) \text{ И } (<Knep> \text{ есть } <Knep^l>) \text{ И} \\
 &(<L_k> \text{ есть } <L_k^l>) \text{ ТОГДА } <Mr> = <M_6>,
 \end{aligned} \tag{2.3}$$

где S означает терм анализируемой нечеткой переменной, соответствующий «небольшому», M – «среднему», а L и B – «большому» значению данной переменной; $Mr = \{M_1, M_2, M_3, M_4, M_5, M_6\}$: M_1 – ручное рубрицирование, M_2 – модель с использованием весовых коэффициентов, M_3 – нечеткое дерево решений, M_4 – нейро-нечеткий классификатор, M_5 – вероятностный классификатор, M_6 – метод голосования.

Очевидно, что для использования практически каждой из перечисленных моделей для автоматизированного рубрицирования ЭНТД требуется соответ-

ствующая подготовка исходной информации, процедуры которой описаны в подразделах 2.3, 2.4, 3.1.

Предлагаемая процедура реализации указанного метода включает в себя этап определения сроков хранения актуальных документов в общей базе системы, которые зависят от степени важности документа V_k :

$$warn_{V_k} \in \{vwd, swd, lwd\},$$

где vwd – высокая степень важности, swd – средняя степень важности, lwd – низкая степень важности.

Перед записью в базу проанализированных документов определяется срок хранения документа по его степени важности, граничные значения которой определяются экспертами в зависимости от типа документа и существующих административных регламентов документооборота (например, для vwd – 5 лет, swd – 1 год, lwd – 6 месяцев). По истечении срока хранения ЭНТД больше не используются при мониторинге рубричного поля и не учитываются при пересчете статистических характеристик.

Одновременно с записью ЭНТД в базу данных происходит определение его адресата из множества $Dest = [Dest_1, ..., Dest_g, ..., Dest_G]$, где $g \in [1, ..., G]$, G – количество лиц, ответственных за реакцию на ЭНТД.

Сказанное выше позволяет заключить, что реализация мультимодельного метода автоматизированного рубрицирования ЭНТД предполагает формирование следующих четырех баз данных:

БД 1 – упомянутая выше база данных морфологий, которая необходима для проведения морфологического анализа и содержит лингвистические характеристики слов и их начальные формы. Данная база заполняется на основе информации из известных морфологических баз, а также пополняется экспертами;

БД 2 – база данных весовых коэффициентов значимости слов для рубрик, которая необходима при использовании соответствующего метода рубрицирования;

БД 3 – база для данных частотных характеристик, которая необходима для работы вероятностного классификатора и пополняется с использованием информации из рубрицированных ЭНТД, нормативных документов, выступлений должностных лиц и т.п.;

БД 4 – база рубрицированных ранее документов, которые необходимы для обработки жалоб и предложений, а также для настройки системы.

На рисунке 3.1 приведен алгоритм обработки и анализа ЭНТД при использовании мультимодельного метода для его автоматизированного рубрицирования.

Предложенный мультимодельный метод анализа электронных неструктурированных текстовых документов позволяет повысить точность выделения рубрик и отнесения к конкретным рубрикам текстовых документов с учетом их специфики в условиях взаимозависимости рубрик и различного объема статистических данных об ЭНТД.

2.2 Каскадная нейро-нечеткая модель анализа коротких электронных неструктурированных текстовых документов с использованием экспертной информации

2.2.1 Структура каскадной нейро-нечеткой модели для рубрицирования коротких электронных неструктурированных текстовых документов

Для рубрицирования коротких ЭНТД (при возникновении типовой ситуации №3, см. таблицу 2.1) предлагается каскадная нейро-нечеткая модель, структура которой представлена на рисунке 2.2.

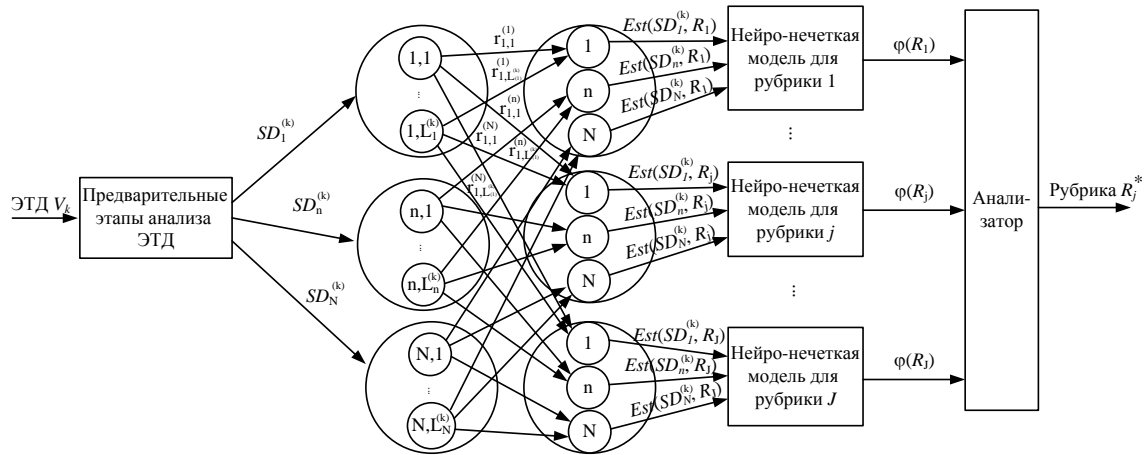


Рисунок 2.2 – Структура каскадного нейро-нечеткого классификатора для рубрицирования ЭНТД

Предлагаемая каскадная нейро-нечеткая модель включает в себя:

- модель для предварительного анализа ЭНТД с использованием синтаксического парсера, предназначена для формирования множеств ЗС ЭНТД, характеризующихся одинаковой синтаксической ролью в предложениях. На вход поступает ЭНТД V_k , а на выходе формируется множество вида:

$$SD_k = \{SD_1^{(k)}, ..., SD_n^{(k)}, ..., SD_N^{(k)}\}, k = 1, ..., K; \quad (2.4)$$
- модель формализации ЭНТД с использованием весовых коэффициентов, реализующая две процедуры: во-первых, сопоставление ЗС $v_p^{(k)}$ каждой синтаксической группы $SD_n^{(k)}$ с БД весовых коэффициентов, и на выходе слоя формируются степени влияния ЗС относительно каждой рубрики R_j ; во-вторых, аккумулярование и нормирование весовых коэффициентов. На выходе модели определяются оценки степеней принадлежности синтаксических групп $SD_n^{(k)}$ ко всем рубрикам R_j ;
- совокупность нейро-нечетких моделей оценки принадлежности (или отнесения) к отдельным рубрикам, каждая из которых предназначена для формирования степени принадлежности ЭНТД к отдельной рубрике R_j ;

- модель выбора рубрики, в наибольшей степени соответствующей анализируемому ЭНТД, предназначена для окончательного выбора рубрики, к которой относится ЭНТД.

2.2.2 Модель рубрицирования электронных неструктурированных текстовых документов с использованием весовых коэффициентов

Как показано в подразделе 2.1, при поступлении на анализ коротких ЭНТД в условиях достаточного объема статистической информации о документах данного типа (необходимой для построения и обучения нейро-нечетких моделей [92]) и незначительной степени пересечения рубрик имеет место типовая ситуация, при которой для рубрицирования ЭНТД целесообразно использовать нейро-нечеткий классификатор [92].

Для решения этой задачи, а также для непосредственного применения в ситуациях рубрицирования средних по размеру ЭНТД при отсутствии необходимого для использования нейро-сетевых моделей объема накопленной статистической информации о документах данного типа и слабой степени пересечения рубрик возникает типовая ситуация №2 (см. таблицу 2.1), которая определяет целесообразность применения модели агрегации и обработки ЭНТД с использованием весовых коэффициентов.

Процедура построения указанной модели с использованием весовых коэффициентов включает следующие шаги:

Шаг 1. Задается первоначальный набор (множество) рубрик по формуле (2.2).

Шаг 2. Выделяется набор ЭНТД с заранее определенными рубриками:

$$V^{(об)} = \{(V_1^{(об)}, RR_1), \dots, (V_b^{(об)}, RR_b), \dots, (V_B^{(об)}, RR_B)\}, \quad (2.5)$$

где $V^{(об)}$ – обучающая выборка, $V_b^{(об)}$ – b -й ЭНТД из обучающей выборки, $RR_b \in R, b = 1, \dots, B$ – рубрика, соответствующая b -му ЭНТД из обучающей вы-

борки. В этих ЭНТД производится поиск значимых слов $v_{l_b}^{(b)}$, длина которых превышает три символа.

В этом случае b -й ЭНТД $V_b^{(об)}$ можно представить в виде:

$$V_b^{(об)} = (v_1^{(b)}, \dots, v_{l_b}^{(b)}, \dots, v_{L_b}^{(b)}), b = 1, \dots, B,$$

где L_b – количество значимых слов b -го ЭНТД.

Шаг 3. Каждому l_b -му слову $v_{l_b}^{(b)}$ b -го ЭНТД назначается начальный весовой коэффициент (вес) $u_{l_b}^{(b)} = 0,5$, показывающий его степень соответствия j -й рубрике, к которой относится b -й ЭНТД. Таким образом, получаем множество пар следующего вида:

$$V_b^{(об)} = \{(v_1^{(b)}, u_1^{(b)}), \dots, (v_{l_b}^{(b)}, u_{l_b}^{(b)}), \dots, (v_{L_b}^{(b)}, u_{L_b}^{(b)})\}, b = 1, \dots, B.$$

Шаг 4. Проводится подстройка весовых коэффициентов модели с использованием элементов обучающей выборки $V^{(об)}$, в ходе которой веса $u_{l_b}^{(b)}$ значимых слов изменяются, исходя из их степени соответствия конкретной j -й рубрике с использованием алгоритма, блок-схема которого приведена на рисунке 3.13.

На выходе данного алгоритма формируются словари рубрик R_j вида:

$$R_j = \{(w_1^{(j)}, r_1^{(j)}), \dots, (w_{m_j}^{(j)}, r_{m_j}^{(j)}), \dots, (w_{M_j}^{(j)}, r_{M_j}^{(j)})\}, j = 1, \dots, J, \quad (2.6)$$

где $w_{m_j}^{(j)}$ – m_j -е слово в рубрике R_j , $m_j = 1, \dots, M_j$, M_j – общее количество значимых слов в j -й рубрике, $r_{m_j}^{(j)} \in [0, 1]$ – весовой коэффициент m_j -го слова для j -й рубрики.

Шаг 5. С учетом того, что описанный в подразделе 2.1 мультимодельный метод автоматизированного рубрицирования ЭНТД предполагает учет весовых коэффициентов для нейро-нечеткого классификатора в ситуации отсутствия большого объема обучающей выборки (вызванной, в том числе, динамикой рубричного поля), корректировка весовых коэффициентов $r_{m_j}^{(j)}$ на всех этапах работы с рассматриваемой моделью проводится с привлечением экспертов.

Процедура применения построенной модели при возникновении типовой ситуации №2 (см. табл. 2.1) проводится по следующему алгоритму:

Шаг 1. Совокупность V поступающих ЭНТД представляется по формуле (2.1).

Шаг 2. Определяется множество оценок $Est(V_k, R_j)$ следующего вида:

$$\forall j \in J : Est(V_k, R_j) = \{(v_{l_k}^{(k)}, u_{l_k}^{(k)})\}, \quad (2.7)$$

где для каждой пары $(v_{l_k}^{(k)}, u_{l_k}^{(k)}) : u_{l_k}^{(k)} = r_{m_j}^{(j)} \mid w_{m_j}^{(j)} = v_{l_k}^{(k)}, u_{l_k}^{(k)}$ – весовой коэффициент l_k -го значимого слова k -го ЭНТД для j -й рубрики.

Шаг 3. Рассчитывается показатель $\rho(V_k, R_j)$, характеризующий степень соответствия ЭНТД V_k рубрике R_j , вида:

$$\forall j \in J, \rho(V_k, R_j) = \frac{\sum_{l_k=1}^{L_k} u_{l_k}^{(k)}}{L_k}.$$

Шаг 4. Проводится непосредственная рубрикация ЭНТД V_k с учетом, что он в наибольшей степени относится к рубрике R_j^* , степень принадлежности к которой является максимальной:

$$R_j^* : \max_{j=1, \dots, J} \rho(V_k, R_j).$$

При построении модели рубрицирования с использованием весовых коэффициентов экспертные оценки должны определяться с учетом следующих обстоятельств:

- во-первых, в ЭНТД встречаются общие слова (употребляемые почти во всех рубриках), которые практически не несут информации о контексте конкретной рубрики. Следовательно, их веса $r_{m_j}^{(j)} = r_{cm}$ необходимо сделать намного меньше других;
- во-вторых, уникальные слова, встречающиеся только в рамках одной рубрики, являются самыми значимыми, поэтому их веса $r_{m_j}^{(j)} = r_{um}$ будут значительно больше других;

- в-третьих, имеются так же редкие слова, которые являются ни уникальными, ни общими, и встречаются только в некотором ограниченном количестве рубрик. Они несут определённую информацию о контексте рубрики, поэтому им назначаются промежуточные значения весовых коэффициентов $r_{m_j}^{(j)} = r_{rr}$.

С учетом сказанного, при определении весовых коэффициентов значимости слов проводится анализ тезаурусов рубрик с учетом их категории (уникальные, редкие и общие ЗС ЭНТД) [115]. Далее эксперт выбирает нужное соотношение и значение весовых коэффициентов (в рамках интервала $[0, 1]$) заданных трех типов значимых слов. Для разграничения редких и общих слов вводится порог встречаемости β . Если слово встретилось в ЭНТД только одной рубрики, то это уникальные слова, если не во всех и меньше порога, то редкие. Все остальные слова считаются общими.

В подразделе 4.3 приведены примеры значений весов указанных категорий слов ЭНТД, полученные в ходе проведения процедур экспертного оценивания и вычислительных экспериментов.

В связи с тем, что базу данных весовых коэффициентов редактирует группа экспертов, необходимо применять методы согласования групповых экспертных оценок с использованием известных процедур (простой ранжировки или предпочтения; задания весовых коэффициентов; парных сравнений; последовательных сравнений и т.п. [116]).

С учетом необходимости определения достаточно большого числа весовых коэффициентов представляется целесообразным применение метода простой ранжировки, в рамках которого каждый эксперт располагает признаками для каждой j -й рубрики в порядке предпочтения. В результате формируется J таблиц, в которых содержатся оценки N_e экспертов M_j слов, т.е. весовые коэффициенты $r_{m_j}^{(j)}, m_j = 1, \dots, M_j$.

Для оценки согласованности мнений экспертов рассчитывается коэффициент конкордации k_j , [116] по формуле вида:

$$k_j = \frac{12 \sum_{i=1}^{M_j} (r_{m_j}^{(j,e)})^2}{N_e^2 ((M_j)^3 - M_j) - N_e \sum_{e=1}^{N_e} T_e},$$

где T_e – коэффициент числа групп одинаковых рангов, назначенных экспертами; N_e – количество экспертов; e – номер эксперта.

Преимуществами модели рубрицирования ЭНТД с использованием весовых коэффициентов (при ее непосредственном применении) по сравнению с вероятностными моделями анализа текстов являются:

- возможность рубрицирования коротких ЭНТД, содержащих большое количество сокращений, цифр и минимум одно ключевое слово; из-за отсутствия явно выраженного порога частоты употребления слов;
- простота обновления базы знаний о рубриках, входящих в состав автоматизированных систем рубрицирования ЭНТД;
- учет информации, содержащейся в словах, не принадлежащих напрямую к рубрике.

В то же время процедурам непосредственного применения моделей рубрицирования с использованием весовых коэффициентов свойственны следующие недостатки:

- высокая степень влияния субъективизма при определении экспертных оценок;
- отсутствие порога распознавания предложения увеличивает вероятность, что анализируемый ЭНТД относится к предметной области, не описанной в рамках применяемой системы автоматизированного рубрицирования;
- появление уникального слова с большим весом в другой предметной области существенно повышает вероятность ошибочного рубрицирования ЭНТД.

Возникновение типовой ситуации №4 (см. таблицу 2.1) определяет возможность снижения влияния степени субъективности, присущей моделям, ис-

пользующим экспертные оценки, путем дополнительного этапа обработки информации, представленной в виде $Est(V_k, R_j)$, с использованием алгоритмов теории нечетких множеств.

2.2.3 Модель формализации электронных неструктурированных текстовых документов для нейро-нечеткого классификатора

При применении модели с использованием весовых коэффициентов для построения нейро-нечеткой модели рубрицирования предлагается следующая процедура.

Шаг 1. Проводится унификация набора параметров ЭНТД в виде:

$$S = \{s_1, \dots, s_n, \dots, s_N\}.$$

Как ранее отмечалось в подразделе 2.1, предлагается использовать следующие синтаксические характеристики слов анализируемых ЭНТД (например, выделяемые системой синтаксического анализатора текстов LinkGrammar [42]): s_1 – корневое слово или сказуемое; s_2 – подлежащее; s_3 – обстоятельство; s_4 – предмет, над которым совершается действие; s_5 – характеристика сказуемого [83].

С использованием указанных синтаксических параметров каждый ЭНТД V_k представляются в виде:

$$V_k = \{(v_1^{(k)}, h_1^{(k)}), \dots, (v_{l_k}^{(k)}, h_{l_k}^{(k)}), \dots, (v_{L_k}^{(k)}, h_{L_k}^{(k)})\},$$

где $v_{l_k}^{(k)}$ – l_k -е слово k -го ЭНТД, $h_{l_k}^{(k)}$ – синтаксический параметр, характеризующий l_k -е слово k -го ЭНТД такое, что $h_{l_k}^{(k)} = s_n \mid s_n \neq 0, n = 1, \dots, N, l_k = 1, \dots, L_k$.

Каждому ЭНТД V_k ставится в соответствие множество синтаксических групп SD_k по формуле 2.4, где $SD_k = \{SD_1^{(k)}, \dots, SD_n^{(k)}, \dots, SD_N^{(k)}\}, k = 1, \dots, K, SD_n^{(k)}$ – множество слов k -го ЭНТД соответствующих n -му синтаксическому параметру, $SD_n^{(k)} = \{v_p^{(k)} \mid h_p^{(k)} = s_n, \forall p\}, n = 1, \dots, N, p = 1, \dots, L_n^{(k)}, L_n^{(k)}$ – количество слов n -го множества k -го ЭНТД.

Шаг 2. Проводится сопоставление множества SD_k всем рубрикам R_j :

$$SD_k \leftrightarrow \{R_1, \dots, R_j, \dots, R_J\}.$$

Для этого вводится множество оценок следующего вида:

$$\forall j \in J : Est(SD_k, R_j) = \{Est(SD_n^{(k)}, R_j)\}, n = 1, \dots, N, \quad (2.8)$$

где $Est(SD_n^{(k)}, R_j) = \frac{1}{L_n^{(k)}} \cdot \sum_{p=1}^{L_n^{(k)}} u_p^{(k)}$, где для каждой пары $u_p^{(k)} = r_{m_j}^{(j)} \mid w_{m_j}^{(j)} = v_p^{(k)}$, $u_p^{(k)}$ –

весовой коэффициент p -го значимого слова k -го ЭНТД для j -й рубрики, а $r_{m_j}^{(j)}$ – весовые коэффициенты значимых слов рубрик, сконфигурированные для модели с использованием весовых коэффициентов.

В результате на вход нейро-нечеткого классификатора для j -й рубрики поступает множество $Est(SD_k, R_j)$.

Результативность предлагаемого способа формализации ЭНТД несущественно зависит от количества содержащихся в нем значимых слов, что позволяет использовать нейро-нечеткий классификатор для рубрицирования документов различного объема без изменения его структуры.

2.2.4 Нейро-нечеткие модели оценки принадлежности электронных неструктурированных текстовых документов к отдельным рубрикам

Детализированная структура нейро-нечеткой модели оценки принадлежности ЭНТД к отдельной рубрике приведена на рисунке 2.3.

На входы элементов 1-го слоя каждой такой нейро-нечеткой модели поступают значения параметров рубрицируемого ЭНТД в виде $Est(SD_n^{(k)}, R_j)$ – оценки степени соответствия слов n -й синтаксической характеристики k -го ЭНТД j -й рубрике, которые вычисляются по формуле 2.8.

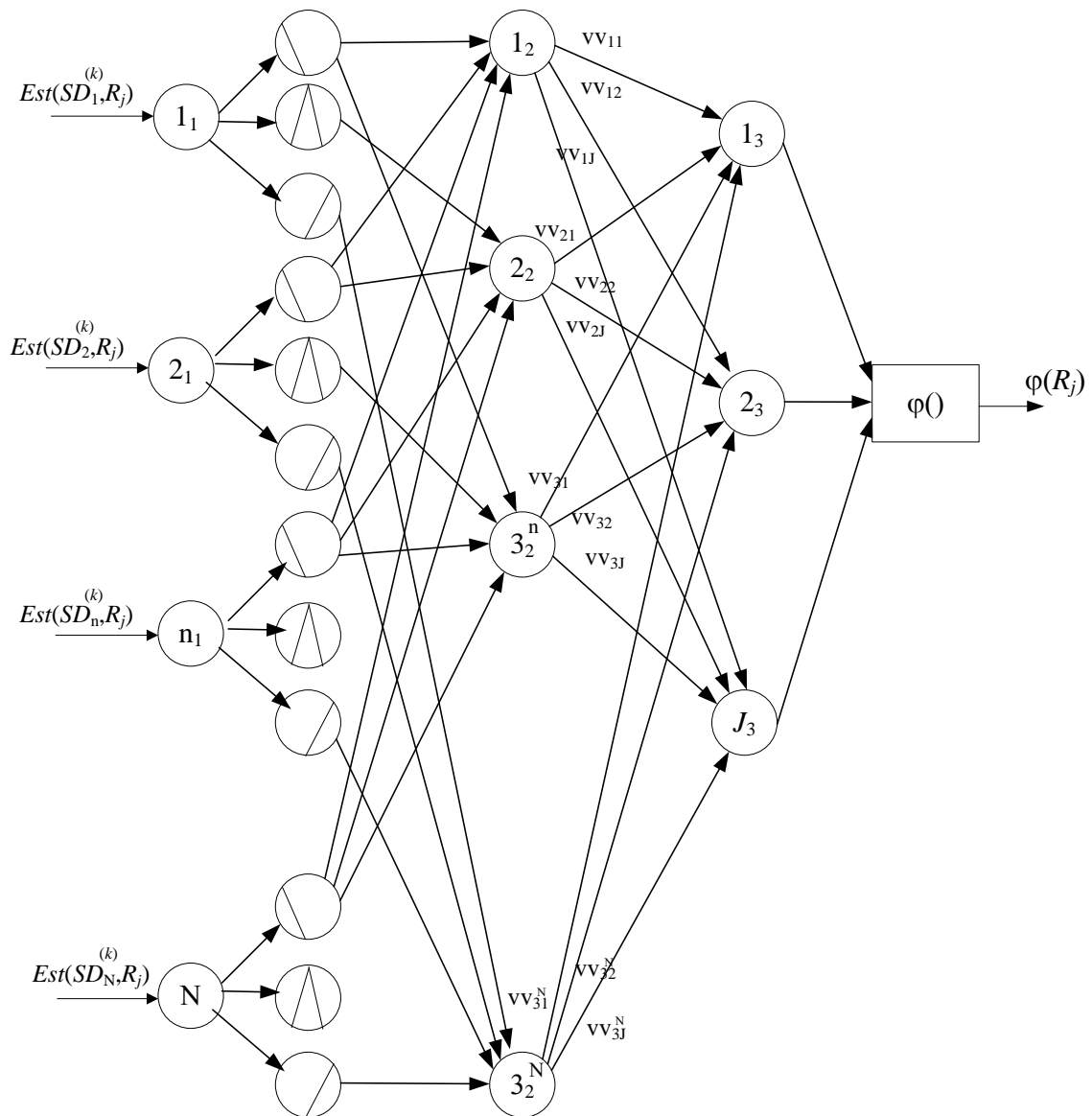


Рисунок 2.3 – Структура нейро-нечеткой модели оценки принадлежности ЭНТД к j -й рубрике

Элементы 2-го слоя модели реализуют нечеткие функции активации для правил вывода, которые оценивают влияние анализируемого слова на определение рубрики и представляют собой терм-множества, соответствующие значениям: «слабое», «среднее» и «высокое» влияние. Для снижения сложности использования указанных моделей представляется целесообразным выбор функций треугольного типа (рисунок 2.4).

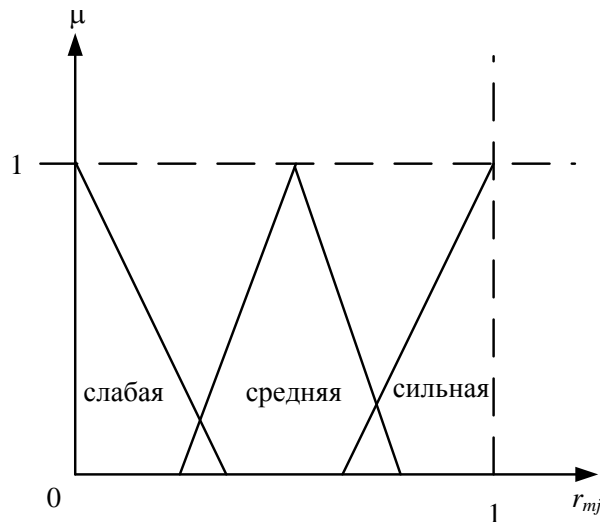


Рисунок 2.4 – Вид функций принадлежности к терм-множествам нечеткого множества «степень влияния параметра ЭНТД на определение рубрики»

Элементы 3-го слоя модели реализуют вычисление функции минимума по всем входным значениям, при этом количество нейронов данного слоя равно 3^N ; коэффициенты vv нейронов настраиваются при обучении. Четвертый слой состоит из J элементов, каждый из которых реализует функцию максимума.

В результате на выходе частной модели формируется степень принадлежности ЭНТД к j -й рубрике.

2.2.5 Модель для выбора рубрик, в наибольшей степени соответствующих электронном неструктурированном текстовым документам

На вход каскадной нейро-нечеткой модели анализа поступает множество ЭНТД $V = \{V_1, \dots, V_k, \dots, V_K\}$, которое модифицируется предварительными этапами анализа и синтаксическим парсером в множество из синтаксических подмножеств ЗС ЭНТД $SD = \{SD_1, \dots, SD_k, \dots, SD_K\}$. Относительно каждого синтаксического подмножества вычисляется оценка его степени близости к каждой рубрике $\forall j \in J : Est(SD_k, R_j) = \{Est(SD_n^{(k)}, R_j)\}, n = 1, \dots, N$. Оценки $Est(SD_n^{(k)}, R_j)$ подаются на входы нейро-нечеткой модели для j -й рубрики. Выходы со всех

нейро-нечетких моделей поступают на анализатор, который позволяет определить рубрику R_j^* , к которой относится ЭНТД по формуле:

$$R_l^* : \max_{j=1, \dots, J} \varphi(R_j).$$

2.2.6 Процедура использования нейро-нечеткого классификатора для рубрицирования коротких электронных неструктурированных текстовых документов

Процедура использования нейро-нечеткого классификатора для рубрицирования ЭНТД представлена на рисунке 2.5.

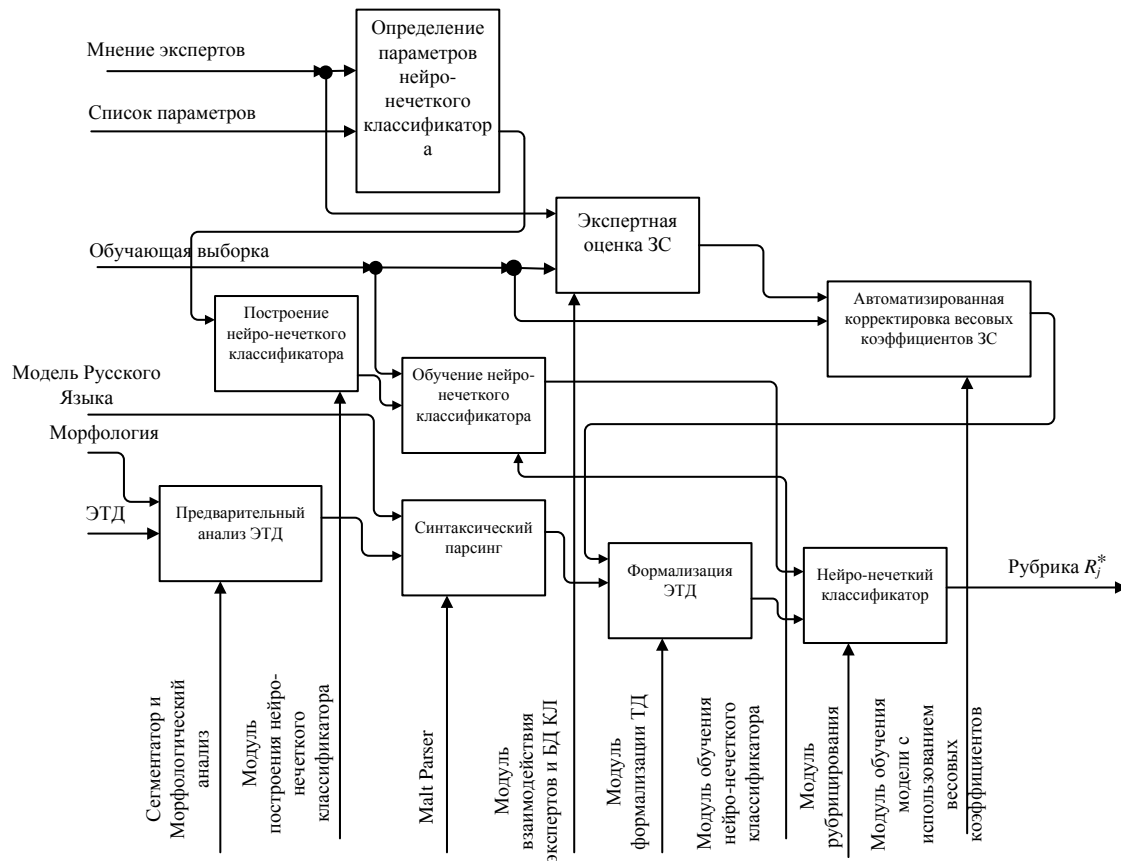


Рисунок 2.5 – Диаграмма процедуры обработки текстовых документов при рубрицировании текстов в автоматизированных системах органов государственного и муниципального управления

На диаграмме наглядно представлены этапы анализа и их участники: модули и подсистемы-помощники, регламентная документация, спецификации.

2.3 Модель анализа электронных неструктурированных текстовых документов на основе нечеткого дерева решений

Как показано в подразделе 2.1, при анализе ЭНТД среднего размера в условиях отсутствия накопленной статистической информации о документах данной рубрики и сильной степени пересечения рубрик имеет место типовая ситуация №3, которая определяет целесообразность использования модели выбора рубрики на основе нечеткого дерева решений (НДР) [111].

Предлагаемая процедура построения нечеткого дерева решений для анализа ЭНТД включает в себя следующие шаги.

Шаг 1. Для всех тезаурусов рубрик R формируются корневые вершины. Количество дочерних узлов d по умолчанию равно 2.

Шаг 2. Если выбранный узел является окончательной рубрикой (листом), то переход на шаг 6, иначе находятся d пересечений тезаурусов рубрик, которые поглощают все остальные словари с коэффициентом покрытия 0,9 (ищутся d множеств тезаурусов рубрик, максимально непохожих друг на друга, т.е. их словари не должны совпадать на 90%).

Шаг 3. В случае, если словари совпадают на 90% и более, необходимо оценить близость словарей данных рубрик (сравнить соответствующие весовые коэффициенты) в n -мерном пространстве, чтобы определить необходимость построения дополнительных уровней в дереве.

Так, если расстояние ρ между словарями меньше порогового значения $\rho_{пор}$, то создание дополнительных уровней в дереве не требуется, и выбранные рубрики становятся листьями, иначе для рубрик, расстояние между которыми больше $\rho_{пор}$, создается один узел дерева, а для которых расстояние больше $\rho_{пор}$, создается второй узел дерева.

Для каждой рубрики R_j вычисляются J координат по каждому w_{m_j} слову следующего вида:

$$KD^{(R_j)} = \{(w_{m_j}^{(j)}, u_{m_j}^{(j,1)}, u_{m_j}^{(j,2)}, \dots, u_{m_j}^{(j,j)}, \dots, u_{m_j}^{(j,J)})\}, j = 1, \dots, J, m_j = 1, \dots, M_j,$$

где $u_{m_j}^{(j,J)} = r_p^{(J)} \mid w_{m_j}^{(j)} = w_p^{(J)}$ – весовой коэффициент m_j -го значимого слова j -й рубрики в контексте J -й рубрики.

Для вычисления расстояния ρ между рубриками необходимо найти центры их кластерных полей:

$$C^{(R_j)} = \{U_1^{(j)}, \dots, U_j^{(j)}, \dots, U_J^{(j)}\},$$

где $U_1^{(j)} = \frac{\sum_{m_j=1}^{M_j} u_{m_j}^{(j,1)}}{M_j}, \dots, U_J^{(j)} = \frac{\sum_{m_j=1}^{M_j} u_{m_j}^{(j,J)}}{M_j}$ – координаты центра кластерного поля рубрики R_j .

Далее вычисляется ρ по следующей формуле:

$$\rho(R_j, R_i) = \rho(C^{R_j}, C^{R_i}) = 1 - \frac{1}{\sqrt{J}} \cdot \sqrt{\sum_{p=1}^J (U_p^{(j)} - U_p^{(i)})^2}.$$

Шаг 4. Если не удаётся найти d пересечений на шаге 3, количество пересечений для поиска увеличивается на 1 ($d=d+1$) и повторяется шаг 3, иначе – переход на шаг 5.

Шаг 5. Переходим к первому дочернему узлу и переходим на шаг 2.

Шаг 6. Если есть следующий дочерний узел, то выбираем его и переходим на шаг 2, иначе дерево построено.

Создаваемое НДР является бинарным, хотя допускается и большая арность дерева.

Целесообразность построения и использования НДР для рубрицирования ЭНТД обусловлена одновременным выполнением следующих условий:

- рубрики взаимосвязаны по тематике, т.е. они характеризуются большой степенью пересечения тезаурусов;
- число слоев НДР больше, чем количество рубрик;
- недостаточно данных для обучения вероятностной модели, нейро-нечеткого классификатора или модели на основе весовых коэффициентов.

Процедура рубрицирования ЭНТД при помощи модели на основе НДР включает следующие шаги.

Шаг 1. Входной документ, по аналогии с формулой 2.4, представляется в виде:

$$SD' = \{SD'_1, ..., SD'_k, ..., SD'_K\},$$

где $SD'_k = \{v_{l_k}^{(k)}\}$, $l_k = 1, ..., L_k$, L_k – количество слов k -го ЭНТД.

Шаг 2. Для того чтобы определить принадлежность ЭНТД V_k к рубрикам, требуется пройти нечеткое дерево решений от корневого узла до листа, сопоставляя множество SD'_k с узлами нечеткого дерева решений.

$$SD'_k \leftrightarrow \{R_{\sum_1}^{(h)}, ..., R_{\sum_g}^{(h)}, ..., R_{\sum_G}^{(h)}\},$$

где $R_{\sum_g}^{(h)}$ – сумма рубрик, относящихся к g -му узлу нечеткого дерева решений на h -м уровне, G – количество узлов на h -м уровне.

Для этого введем множество оценок, по аналогии с формулой 2.8, следующего вида:

$$\forall j \in J : Est(SD'_k, R_{\sum_g}^{(h)}) = \{Est(SD_n^{(k)'}, R_{\sum_g}^{(h)})\}, n = 1, ..., N,$$

где $Est(SD_n^{(k)'}, R_{\sum_g}^{(h)}) = \{(v_p^{(k)}, u_p^{(k)})\}$, а для каждой пары

$(v_p^{(k)}, u_p^{(k)}) : u_p^{(k)} = r_{m_{R_{\sum_g}^{(h)}}}^{(R_{\sum_g}^{(h)})} | w_{m_{R_{\sum_g}^{(h)}}}^{(R_{\sum_g}^{(h)})} = v_p^{(k)}$, $u_p^{(k)}$ – степень соответствия p -го слова k -го

ЭНТД сумме рубрик $R_{\sum_g}^{(h)}$, соответствующих g -му узлу на h -м слое нечеткого

дерева решений; $r_{m_{R_{\sum_g}^{(h)}}}^{(R_{\sum_g}^{(h)})}$ – среднее значение степени соответствия $v_p^{(k)}$ слова k -го

ЭНТД рубрикам $R_{\sum_g}^{(h)}$.

Сопоставление же ЭНТД V_k рубрикам $R_{\sum_g}^{(h)}$ выполняется с использованием следующих нечетких множеств:

$$\forall j \in J, FS(SD'_k, R_{\sum_g}^{(h)}) = \left\{ \left(\mu_{FS(SD'_k, R_{\sum_g}^{(h)})}(SD_n^{(k)'}) / s_n \right) \right\}, n = 1, ..., N,$$

где $\mu_{FS(SD_k', R_{\Sigma_g}^{(h)})}(SD_n^{(k)'})$ – степень принадлежности k -го ЭНТД $R_{\Sigma_g}^{(h)}$ рубрик по синтаксическому параметру s_n :

$$\mu_{FS(SD_k', R_{\Sigma_g}^{(h)})}(SD_n^{(k)'}) = \frac{1}{L_n^{(k)}} \sum_{p=1}^{L_n^{(k)}} u_p^{(k)}, n=1, \dots, N.$$

Введем показатель $\rho(SD_k', R_{\Sigma_g}^{(h)})$, характеризующий *степень соответствия* ЭНТД V_k к рубрикам $R_{\Sigma_g}^{(h)}$. Для определения этой степени соответствия могут быть применены различные способы [53]. Наиболее целесообразным для решения данной задачи представляется использование дополнения относительного евклидова расстояния между нечеткими множествами:

$$\forall j \in J, \rho(SD_k', R_{\Sigma_g}^{(h)}) = 1 - \frac{1}{\sqrt{N}} \sqrt{\sum_{n=1}^N \left(1 - \mu_{FS(SD_k', R_{\Sigma_g}^{(h)})}(SD_n^{(k)'})\right)^2},$$

Считается, что ЭНТД V_k в наибольшей степени относится к той рубрике $R_{\Sigma_l}^*$, степень принадлежности к которой является максимальной:

$$R_{\Sigma_l}^* : \max_{j=1, \dots, J} \rho(SD_k', R_{\Sigma_g}^{(h)}).$$

Таким образом, на каждом h -м слое для g -го узла для текстового документа вычисляются столько характеристик $\rho(SD_k', R_{\Sigma_g}^{(h)})$, сколько у данного узла есть дочерних узлов, среди них выбирается узел с максимальным значением характеристики принадлежности, и переходим на него для проверки следующих условий (должно быть пороговое значение или функция активации).

Шаг 3. Шаг 2 повторяется до тех пор, пока не достигнем самого нижнего слоя, который и определяет рубрику входного текстового документа.

Достоинства моделей анализа ЭНТД на основе НДР:

- более высокая точность рубрицирования ЭНТД при указанных выше условиях применения по сравнению с моделями, вследствие меньшей вероятности случайных ошибок рубрицирования на верхних ярусах НДР;
- меньшая трудоемкость процедуры рубрицирования ЭНТД вследствие направленного (а не переборного) анализа по отдельной ветви НДР.

Пример структуры НДР представлена на рисунке 2.6.

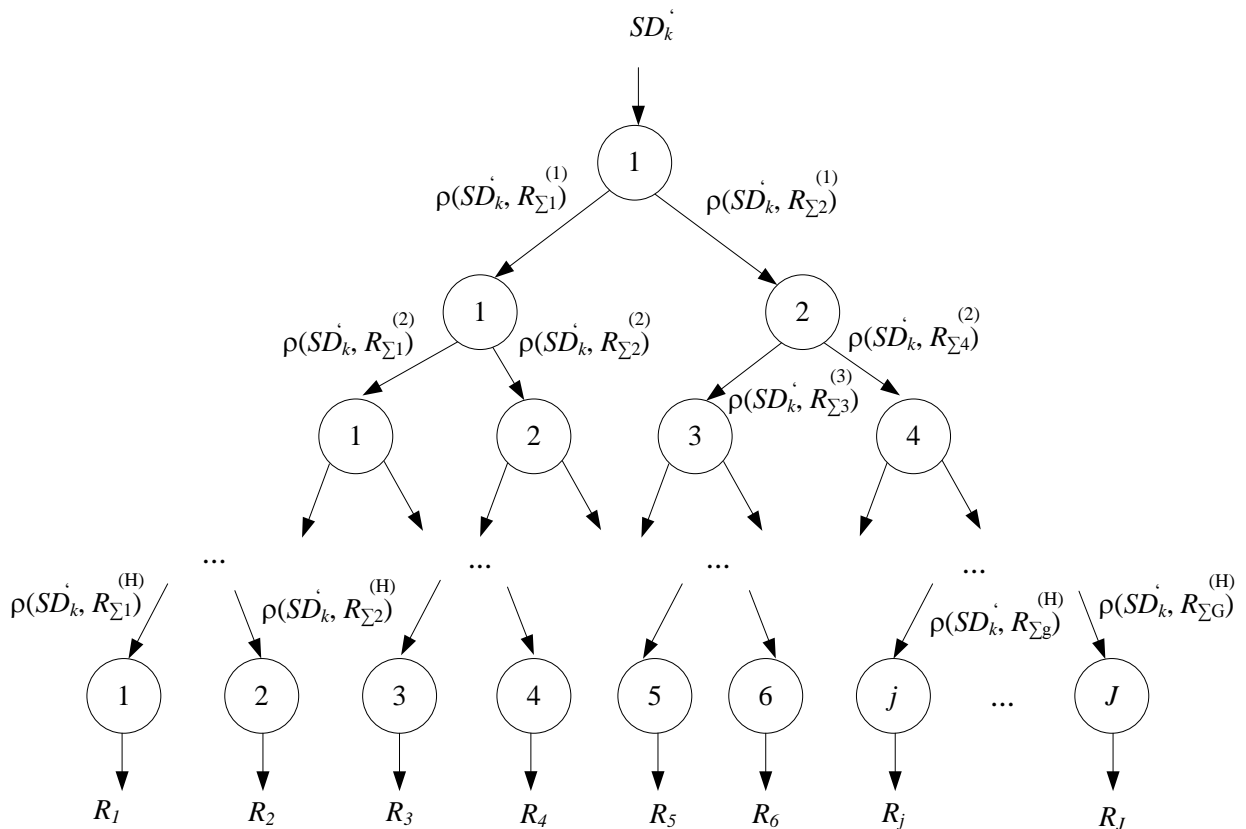


Рисунок 2.6 – Нечеткое дерево решений для k -го ЭНТД

На рисунке 2.7 процесс рубрицирования ЭНТД представлен в виде IDEF-диаграммы.

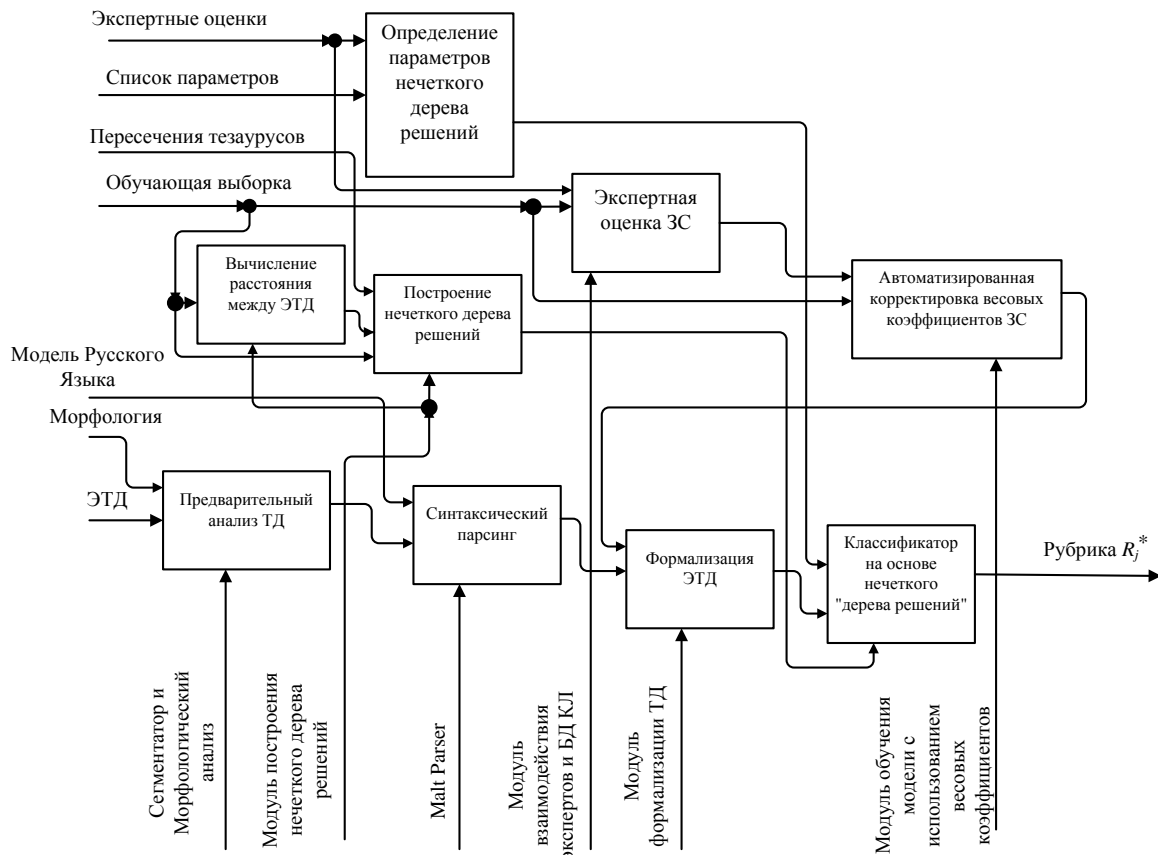


Рисунок 2.7 – Процесс рубрицирования с использованием модели на основе НДР

На диаграмме наглядно представлены этапы анализа и их участники: модули и подсистемы-помощники, регламентная документация, спецификации.

2.4 Метод мониторинга и изменения рубрик электронных неструктурированных текстовых документов на основе их нечеткой динамической кластеризации

Для мониторинга и изменения рубричного поля ЭНТД предлагается представленный ниже метод на основе их нечеткой динамической кластеризации, который включает ряд этапов.

Этап 1. Сопоставление ЭНТД рубрикам.

Для того чтобы определить принадлежность ЭНТД V_k к рубрикам, требуется сопоставить множество SD_k со всеми рубриками R_j .

$$SD_k \leftrightarrow \{R_1, \dots, R_j, \dots, R_J\},$$

Для этого введем множество оценок следующего вида:

$$\forall j \in J : Est(SD_k, R_j) = \{Est(SD_n^{(k)}, R_j)\}, n = 1, \dots, N,$$

где $Est(SD_n^{(k)}, R_j) = \{(v_p^{(k)}, u_p^{(k)})\}$, а для каждой пары $(v_p^{(k)}, u_p^{(k)}) : u_p^{(k)} = r_{m_j}^{(j)} \mid w_{m_j}^{(j)} = v_p^{(k)}$, $u_p^{(k)}$ – степень соответствия p -го слова k -го ЭНТД j -й рубрике.

Сопоставление же ЭНТД V_k рубрикам R_j выполняется с использованием следующих нечетких множеств:

$$\forall j \in J, FS(SD_k, R_j) = \left\{ \left(\mu_{FS(SD_k, R_j)}(SD_n^{(k)}) / s_n \right) \right\}, n = 1, \dots, N,$$

где $\mu_{R_{j_n}}(SD_n^{(k)})$ – степень нечеткой принадлежности k -го ЭНТД j -й рубрике по синтаксическому параметру s_n :

$$\mu_{FS(SD_k, R_j)}(SD_n^{(k)}) = \frac{1}{L_n^{(k)}} \sum_{p=1}^{L_n^{(k)}} u_p^{(k)}, n = 1, \dots, N.$$

Введем показатель $\rho(SD_k, R_j)$, характеризующий *степень соответствия* ЭНТД V_k рубрике R_j . Для определения этой степени соответствия могут быть применены различные способы [53]. Наиболее целесообразным для решения данной задачи представляется использование дополнения относительного евклидова расстояния между нечеткими множествами:

$$\forall j \in J, \rho(SD_k, R_j) = 1 - \frac{1}{\sqrt{N}} \sqrt{\sum_{n=1}^N \left(\mu_{R_j}(R_{j_n}) - \mu_{FS(SD_k, R_j)}(SD_n^{(k)}) \right)^2},$$

где $\tilde{R}_j = \left\{ \left(\mu_{R_j}(R_{j_n}) / s_n \right) \right\}$ – нечеткое множество, характеризующее «ядро» рубрики R_j . Для рассматриваемого случая $\tilde{R}_j = \{(1/s_1), (1/s_2), (1/s_3), (1/s_4), (1/s_5)\}$, т.е.

$$\forall j \in J, \rho_1(SD_k, R_j) = 1 - \frac{1}{\sqrt{N}} \sqrt{\sum_{n=1}^N \left(1 - \mu_{FS(SD_k, R_j)}(SD_n^{(k)}) \right)^2}.$$

Считается, что ЭНТД V_k в наибольшей степени относится к той рубрике R_l^* , степень принадлежности к которой является максимальной:

$$R_l^* : \max_{j=1, \dots, J} \rho_1(SD_k, R_j).$$

Этап 2. Проверка условий пересмотра состава и структуры рубричного поля.

Для повышения гибкости мониторинга рубрицирования и пересмотра рубричного поля помимо рассмотренного выше показателя $\rho_1(SD_k, R_j)$, характеризующего степень соответствия ЭНТД SD_k рубрике R_j , введем еще два дополнительных показателя:

$$\forall j \in J, \rho_{0,5}(SD_k, R_j) = 1 - \frac{1}{\sqrt{N}} \sqrt{\sum_{n=1}^N \left(0,5 - \mu_{FS(SD_k, R_j)}(SD_n^{(k)})\right)^2},$$

$$\forall j \in J, \rho_0(SD_k, R_j) = 1 - \frac{1}{\sqrt{N}} \sqrt{\sum_{n=1}^N \left(\mu_{FS(SD_k, R_j)}(SD_n^{(k)})\right)^2},$$

где $\rho_{0,5}(SD_k, R_j)$ характеризует *степень неопределенности отнесения ЭНТД SD_k к рубрике R_j* , а $\rho_0(SD_k, R_j)$ – *степень несоответствия ЭНТД SD_k рубрике R_j* .

Реализация данного этапа предлагает вычисление показателей $\rho_1(SD_k, R_j)$, $\rho_{0,5}(SD_k, R_j)$, $\rho_0(SD_k, R_j)$ для всех ЭНТД и их анализ, по результатам которого на основе приведенных ниже условий осуществляется пересмотр состава и структуры рубричного поля.

На рисунке 2.8 приведена графическая иллюстрация отнесения одного ЭНТД к 1-й рубрике, а другого ЭНТД – к 3-й рубрике (при наличии трех рубрик и двух синтаксических параметров).

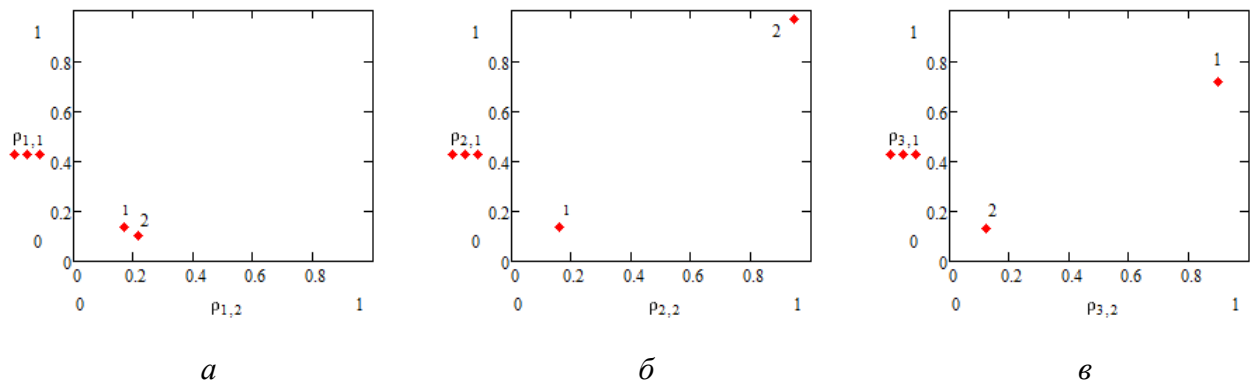


Рисунок 2.8 – Пример рубрицирования двух ЭНТД по двум синтаксическим параметрам при наличии трех рубрик

Этап 3. Идентификация типовой ситуации и изменение состава и структуры рубричного поля.

По результатам мониторинга рубрик в зависимости значений полученных показателей $\rho_1(SD_k, R_j)$, $\rho_{0,5}(SD_k, R_j)$, $\rho_0(SD_k, R_j)$, $k = 1, \dots, K$, $j = 1, \dots, J$ возможны следующие типовые ситуации изменения рубричного поля:

- выделение дополнительной рубрики на «стыке» существующих рубрик;
- разделение рубрики;
- формирование новой рубрики;
- исключение рубрики;
- объединение рубрик.

Рассмотрим предлагаемые условия выявления основных ситуаций и правила пересмотра состава и структуры рубричного поля.

Основанием для *выделения дополнительной рубрики на «стыке» уже существующих рубрик R_i и R_j* является поступление в систему автоматизированного рубрицирования значимого количества ЭНТД, для которых выполняется следующее условие:

$$\begin{aligned}
& \left(\alpha < \rho_1(SD_k, R_i) < \beta \text{ И } \alpha < \rho_1(SD_k, R_j) < \beta \right) \text{ И} \\
& \left(\rho_{0,5}(SD_k, R_i) < \alpha \text{ И } \rho_{0,5}(SD_k, R_j) < \alpha \right) \text{ И} \\
& \left(\alpha < \rho_0(SD_k, R_i) < \beta \text{ И } \alpha < \rho_0(SD_k, R_j) < \beta \right) \text{ И} \\
& \forall R_l \in R, l \neq i \neq j: \rho_1(SD_k, R_l) > \beta \text{ И } \rho_{0,5}(SD_k, R_l) > \alpha \text{ И } \rho_0(SD_k, R_l) < \alpha,
\end{aligned} \tag{2.9}$$

где α и β – нижнее и верхнее граничные значения, определяющие целесообразность пересмотра рубричного поля (обычно, $\alpha = 0,4$ и $\beta = 0,7$ [54]).

На рисунке 2.9 приведена иллюстрация ситуации целесообразности формирования дополнительной «стыковой» рубрики для двух синтаксических параметров.

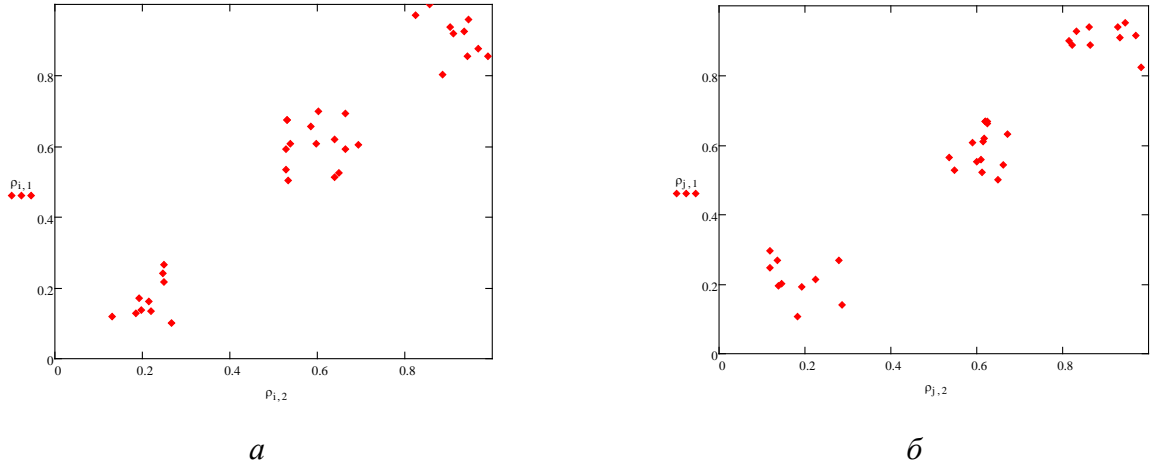


Рисунок 2.9 – Иллюстрация ситуации целесообразности формирования дополнительной «стыковой» рубрики

Количество ЭНТД, удовлетворяющих условиям (2.9), которые определяют целесообразность формирования дополнительной рубрики, может определяться на основе анализа допустимого уровня ошибок рубрицирования ЭНТД для случая их позиционирования на «стыках» рубрик. Под ошибками рубрицирования в данном контексте понимается неправильное отнесение ЭНТД к рубрике, а также рубрицирование, влекущее за собой дополнительные трудности подготовки ответа. Последнее обстоятельство может возникнуть вследствие недостаточной компетентности специалистов, которые в соответствии с регламентом закреплены за рубриками.

Это позволяет сформулировать следующее правило о формировании дополнительной рубрики:

$$K_1 C_1 + K_2 C_2 > (K_1 + K_2) C_3, \text{ при } t_{оме} < t_{дон}, \quad (2.10)$$

где K_1 – число неправильно рубрицированных ЭНТД на момент мониторинга рубричного поля; K_2 – число неправильно рубрицированных ЭНТД; C_1 – затраты, вызванные неправильным рубрицированием одного ЭНТД; C_2 – затраты, вызванные трудностью подготовки ответа на ЭНТД; C_3 – затраты, связанные с обработкой ЭНТД и с использованием дополнительной сформированной рубрики; $t_{оме}$ – отведенное время ответа на ЭНТД; $t_{дон}$ – максимально допустимое время ответа на ЭНТД.

Примечание. Данное правило может быть применено и для других ситуаций, требующих динамического изменения рубричного поля.

Основанием для *разделения рубрики* R_j является поступление в систему автоматизированного рубрицирования значимого (с точки зрения выполнения правила (2.10)) количества ЭНТД, для которых выполняется следующее условие:

$$\begin{aligned} &\alpha < \rho_1(SD_k, R_j) < \beta \text{ И } \rho_{0,5}(SD_k, R_j) < \alpha \text{ И } \alpha < \rho_0(SD_k, R_j) < \beta \text{ И} \\ &\forall R_l \in R, l \neq j: \rho_1(SD_k, R_l) > \beta \text{ И } \rho_{0,5}(SD_k, R_l) > \alpha \text{ И } \rho_0(SD_k, R_l) < \alpha, \end{aligned} \quad (2.11)$$

где j – номер разделяемой рубрики.

На рисунок 2.10 приведена иллюстрация ситуации целесообразности разделения рубрики для двух синтаксических параметров.

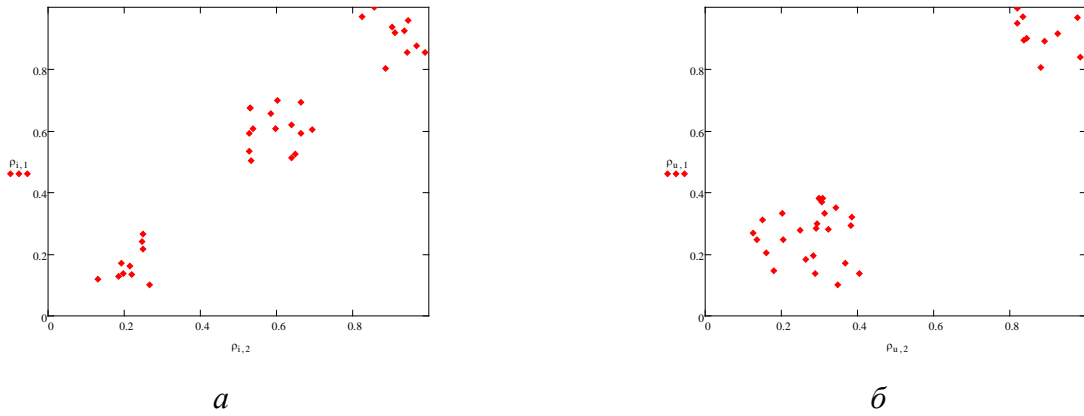


Рисунок 2.10 – Иллюстрация ситуации целесообразности разделения рубрики

На рисунок 2.11 показаны результаты эффективного рубрицирования ЭНТД после разделения рубрики.

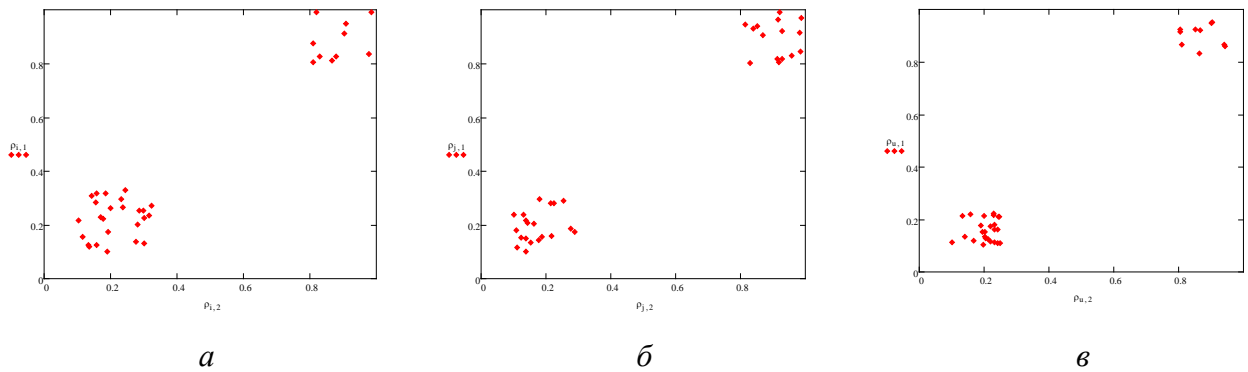


Рисунок 2.11 – Иллюстрация результатов рубрицирования ЭНТД
после разделения рубрики

Примечание. Представленные на рисунке 2.11 результаты рубрицирования ЭНТД являются целевыми и для других ситуаций, требующих динамического изменения рубричного поля.

Основанием для *формирования новой рубрики* является ситуация, которая проиллюстрирована на рисунке 2.12, возникающая при появлении достаточного количества ЭНТД, для которых выполняется следующее условие:

$$\forall R_l \in R : \rho_1(SD_k, R_l) > \beta \text{ И } \rho_{0,5}(SD_k, R_l) > \alpha \text{ И } \rho_0(SD_k, R_l) < \alpha. \quad (2.12)$$

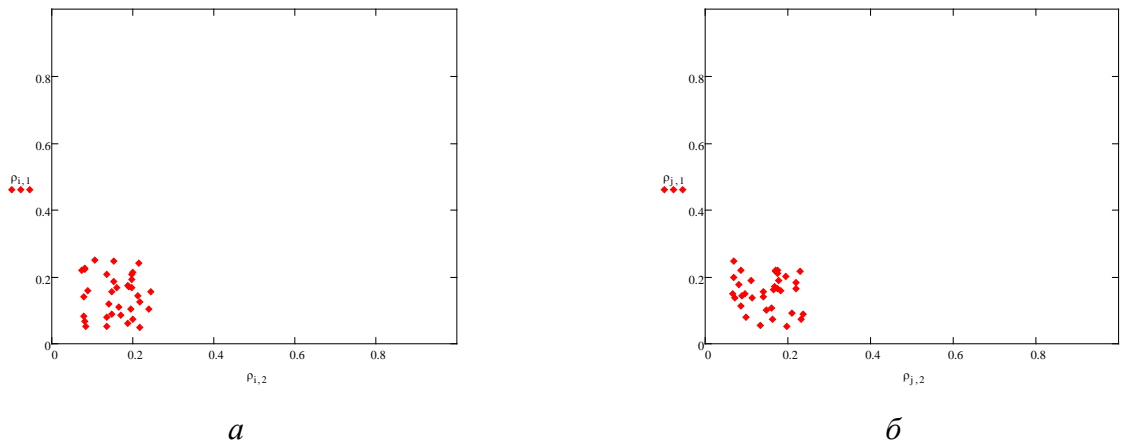


Рисунок 2.12 – Иллюстрация ситуации целесообразности формирования новой рубрики

Основанием для *исключения рубрики* является ситуация, возникающая при появлении подавляющего числа ЭНТД, для которых выполняется следующее условие:

$$\rho_1(SD_k, R_j) > \beta \text{ И } \rho_{0,5}(SD_k, R_j) > \alpha \text{ И } \rho_0(SD_k, R_j) < \alpha. \quad (2.13)$$

Если для рубрики практически все ЭНТД позиционируются так, как это показано на рисунке 2.13, то эту рубрику целесообразно исключить.

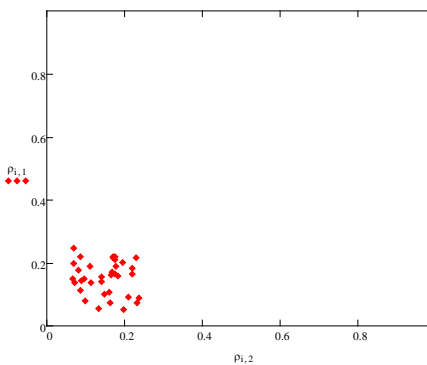


Рисунок 2.13 – Иллюстрация ситуации целесообразности исключения рубрики

Основанием для объединения рубрик R_i и R_j является ситуация, возникающая при появлении достаточного количества ЭНТД, для которых выполняется следующее условие:

$$\begin{aligned}
 & \left(\rho_1(SD_k, R_i) < \alpha \text{ И } \rho_1(SD_k, R_j) < \alpha \right) \text{ И} \\
 & \left(\rho_{0,5}(SD_k, R_i) > \alpha \text{ И } \rho_{0,5}(SD_k, R_j) > \alpha \right) \text{ И} \\
 & \left(\rho_0(SD_k, R_i) > \beta \text{ И } \rho_0(SD_k, R_j) > \beta \right) \text{ И} \\
 & \forall R_l \in R, l \neq i \neq j: \rho_1(SD_k, R_l) > \beta \text{ И } \rho_{0,5}(SD_k, R_l) > \alpha \text{ И } \rho_0(SD_k, R_l) < \alpha,
 \end{aligned} \tag{2.14}$$

где R_i и R_j – объединяемые рубрики.

Выполнение условия (2.14) для объединения рубрик R_i и R_j проиллюстрировано на рисунке 2.14.

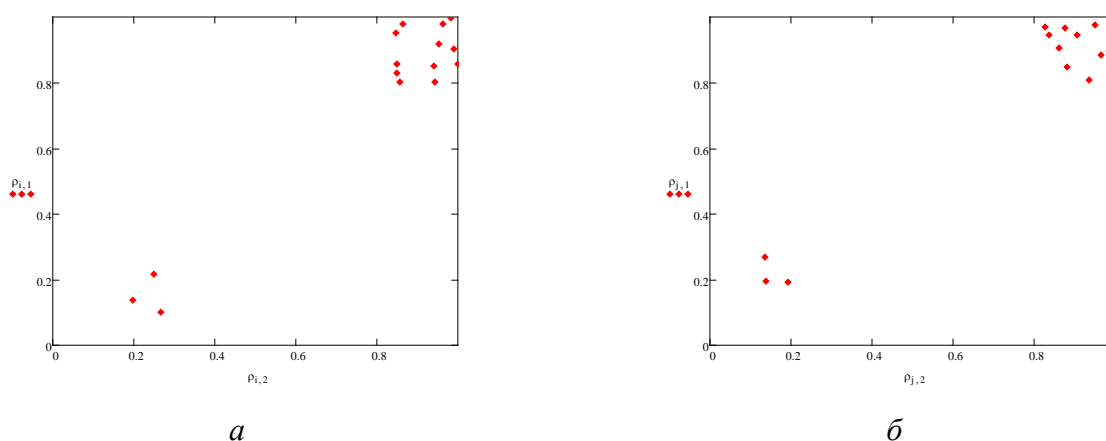


Рисунок 2.14 – Иллюстрация ситуации целесообразности объединения рубрик

Использование предлагаемого способа анализа и рубрицирования электронных текстовых документов позволяет повысить качество и оперативность процессов реагирования органов власти различного уровня на жалобы, предложения и обращения граждан и организаций, поступающие через сеть интернет на порталы и адреса электронной почты органов государственной власти.

При наличии указанных выше возможных изменениях рубричного поля необходимо корректировать параметры разработанных моделей рубрицирования

ЭНТД. В таблице 2.2 представлены типовые ситуации изменения структуры рубричного поля и соответствующие им изменения в моделях рубрицирования.

Таблица 2.2 – Изменения в моделях рубрицирования ЭНТД в зависимости от динамики рубричного поля

<div> <div>Модель</div> <div>Ситуация</div> </div>	Модель с использованием весовых коэффициентов	Модель на основе нейро-нечеткого классификатора	Модель на основе НДР
Выделение дополнительной рубрики на «стыке» существующих рубрик	Тезаурус новой рубрики заполняется словами ЭНТД, попадающими на стык рубрик; происходит перерасчет весовых коэффициентов имеющихся рубрик.	Для выделенной рубрики строится и обучается на всем наборе ЭНТД новая нейро-нечеткая сеть.	«Стыковые» рубрики в виде листов НДР объединяются в узел, и происходит повторное разбиения узла на листы.
Разделение рубрики	Появляется новая рубрика, тезаурус которой заполняется словами из ЭНТД, вызывающих разделение рубрики; происходит перерасчет весовых коэффициентов первоначальной рубрики.	Строится новая нейро-нечеткая сеть для выделенной рубрики и обучается на всем наборе ЭНТД.	Исходная рубрика в виде листа НДР становится узлом, из которого следуют два листа – рубрики (первоначальной и новой).
Формирование новой рубрики	Появляется новая рубрика, тезаурус которой заполняется словами из ЭНТД, не попадающих в имеющиеся рубрики; перерасчет весовых коэффициентов тезаурусов остальных рубрик не требуется.	Для сформированной рубрики строится и обучается на всем наборе ЭНТД новая нейро-нечеткая сеть.	Узлы, соответствующие сформированной рубрике и наиболее близкой к ней, объединяются. Полученный узел дробится на два листа.
Исключение рубрики	Из БД удаляется рубрика вместе с тезаурусом.	Удаляется нейро-нечеткая сеть для данной рубрики	Удаляется соответствующий лист в НДР. Если остаётся один соседний лист для некоторого узла, то данный узел становится листом

Продолжение таблицы 2.2 – Изменения в моделях рубрицирования ЭНТД в зависимости от динамики рубричного поля

Объединение рубрик	Тезаурусы рубрик объединяются, и происходит перерасчет весовых коэффициентов для новой рубрики.	Создается новая нейро-нечеткая сеть, обучаемая на всем наборе ЭНТД; обученные ранее нейро-нечеткие сети для объединяемых рубрик исключаются.	Листы НДР, которые соответствуют сливающимся рубрикам, объединяются в один лист. Если данный лист является единственным для некоторого узла, то данный узел становится листом.
--------------------	---	--	--

Из таблицы 2.2. видно, что с практической точки зрения указанные варианты изменения рубричного поля не приводят, в общем случае, к необходимости существенной модификации структуры и параметров предлагаемых моделей для рубрицирования ЭНТД.

2.5 Выводы по главе

Предложен мультимодельный метод анализа электронных неструктурированных текстовых документов, отличающийся комбинированным использованием нечетких, нейро-нечетких и вероятностных моделей. Представлена структура процедуры использования метода с описанием всех информационных взаимосвязей между различными моделями анализа текстов, а также общая схема выбора моделей для рубрицирования электронных сообщений в зависимости от объема содержащейся в них информации. Использование указанного метода позволяет повысить точность выделения рубрик и отнесения к конкретным рубрикам текстовых документов с учетом их специфики в условиях зависимости рубрик друг от друга и различного объема статистических данных.

Разработан метод мониторинга и изменения рубрик (слияния, разделения, появления новых и ликвидации) для электронных неструктурированных текстовых документов. Данный метод основан на использовании процедур динамической кластеризации этих документов с учетом синтаксических ролей слов, а

также числа и характеристик рубрик. Использование отдельного кластерного поля для каждой рубрики дает возможность обеспечить адаптивную актуализацию рубрик в зависимости от структуры и показателей текстовых документов в условиях нестационарности состава тезауруса и важности ключевых слов рубрик.

Предложена каскадная нейро-нечеткая модель рубрицирования коротких электронных неструктурированных текстовых документов с учетом определения значимости ключевых слов при их формализации для последующего анализа на основе нейро-нечеткого классификатора. Данная модель позволяет рубрицировать короткие электронные неструктурированные текстовые документы в условиях нехватки статистических данных для использования вероятностных классификаторов. Применение указанной модели позволяет использовать возможности нейро-нечетких сетей для комплексного анализа статистической и экспертной информации.

Разработана модель рубрицирования электронных неструктурированных текстовых документов с учетом синтаксических связей и ролей слов в предложениях на основе нечеткого дерева решений. Построение дерева решений основано на анализе степени пересечений словарей рубрик, а также расстояний между рубриками в n -мерном пространстве признаков. Данная модель позволяет более точно рубрицировать электронные неструктурированные текстовые документы в условиях взаимосвязанных рубрик, а также повысить оперативность обработки поступивших документов.

3 РАЗРАБОТКА АЛГОРИТМОВ АНАЛИЗА ЭЛЕКТРОННЫХ НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ И МОНИТОРИНГА РУБРИЧНОГО ПОЛЯ

3.1 Алгоритмы реализации мультимодельного метода рубрицирования электронных неструктурированных текстовых документов

Для практического использования предлагаемого во 2-й главе мультимодельного метода анализа электронных неструктурированных текстовых документов предлагается алгоритм его реализации, использующий подходы описанные в работах [117-119], схема которого представлена на рисунке 3.1.

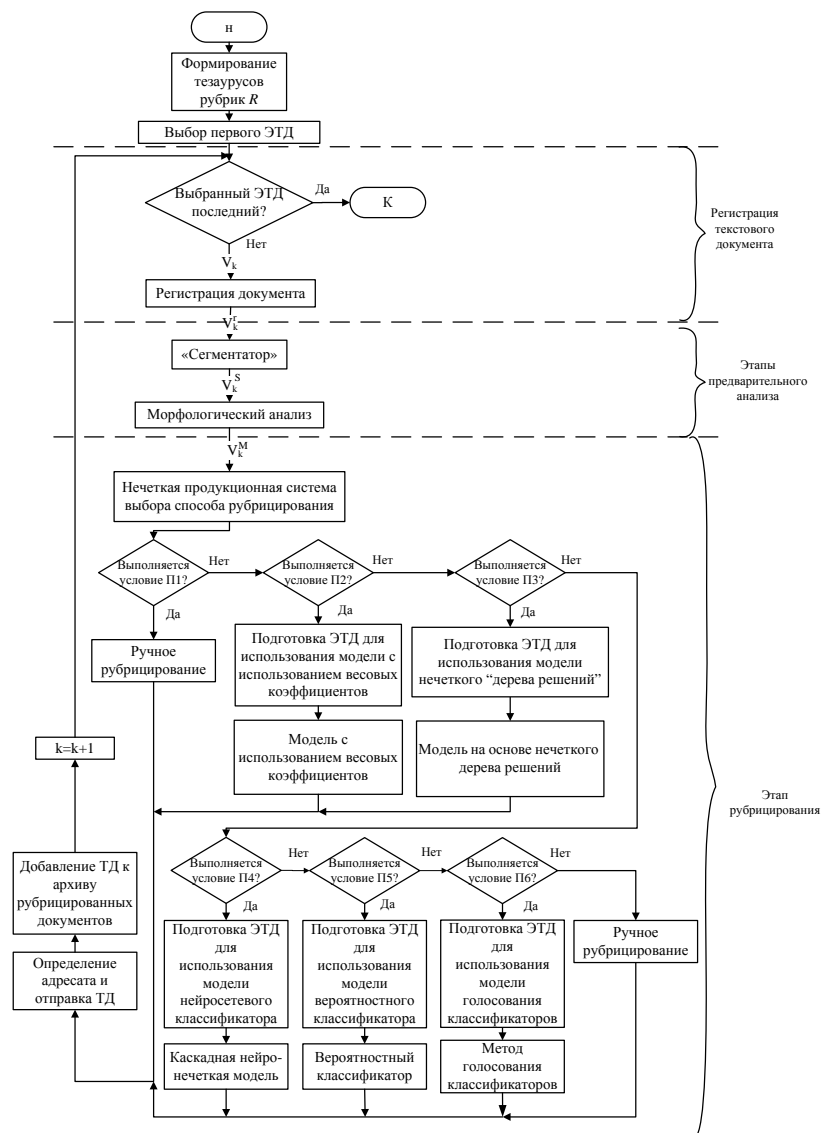


Рисунок 3.1 – Схема алгоритма реализации мультимодельного метода рубрицирования ЭНТД

Первым подготовительным этапом метода является этап сегментации (см. подраздел 2.1), схема алгоритма реализации которого представлена на рисунке 3.2. В результате сегментации из зарегистрированного ЭНТД V_k^r формируется сегментированный ЭНТД V_k^s .

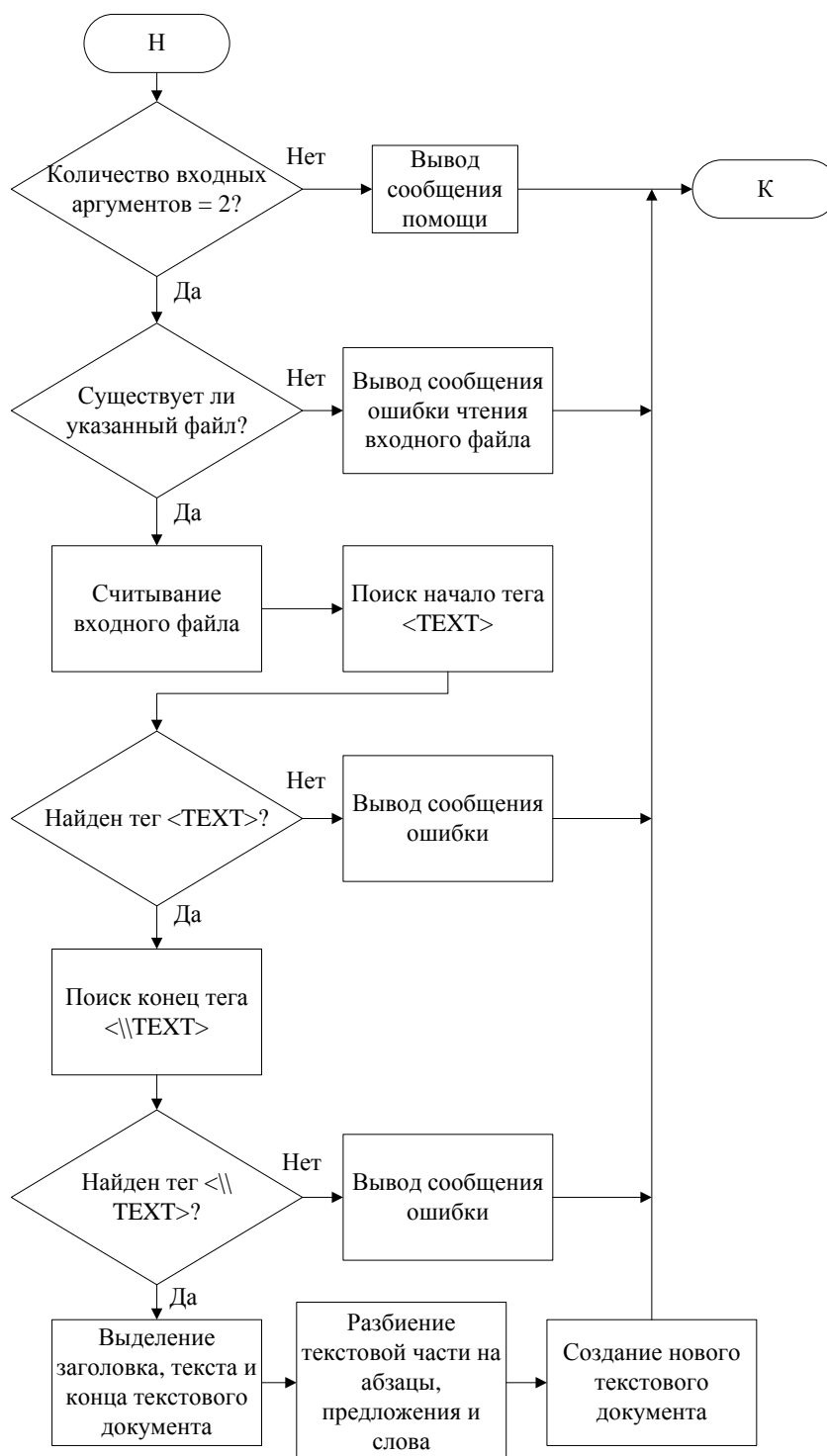


Рисунок 3.2 – Схема алгоритма сегментации ЭНТД

Как видно из рисунка 3.2, для построения и применения моделей рубрицирования ЭНТД (на основе нечеткого дерева решений, модели рубрицирования с использованием весовых коэффициентов, нейро-нечеткого классификатора, вероятностного классификатора) эти документы необходимо разбить на слова, абзацы и предложения (см. подраздел 2.1).

В свою очередь, указанный алгоритм сегментации включает в себя процедуру выделения абзацев в ЭНТД (рисунок 3.3). Данная процедура необходима для более детализированного разбиения ЭНТД и предполагает дальнейшее выделение предложений и слов, поиск URL ссылок, сокращений, дат и т.п., перевод их в удобную для дальнейшего анализа форму, «обрамление» тегами. В результате формируется набор последовательно расположенных абзацев ЭНТД.

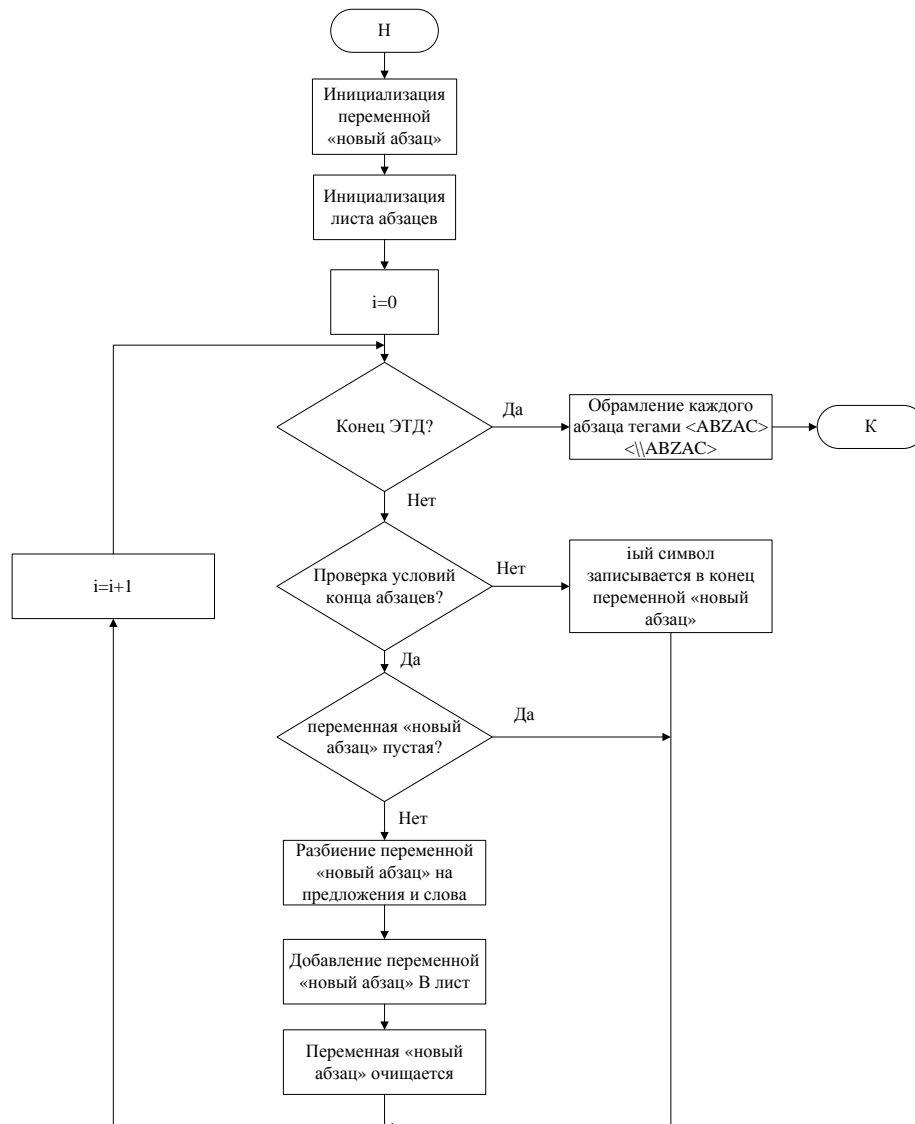


Рисунок 3.3 – Схема алгоритма разбиения ЭНТД на абзацы

После выделения абзацев ЭНТД на следующем этапе метода выполняется их разбиение на предложения (рисунки 3.4 и 3.5). Основными шагами данного алгоритма являются: проверка условий конца предложения, использование тегов вида <ABZAC> и <PREDLOJ> для выделения правильных текстовых структур, расстановка флагов конца предложений для корректной обработки слов.

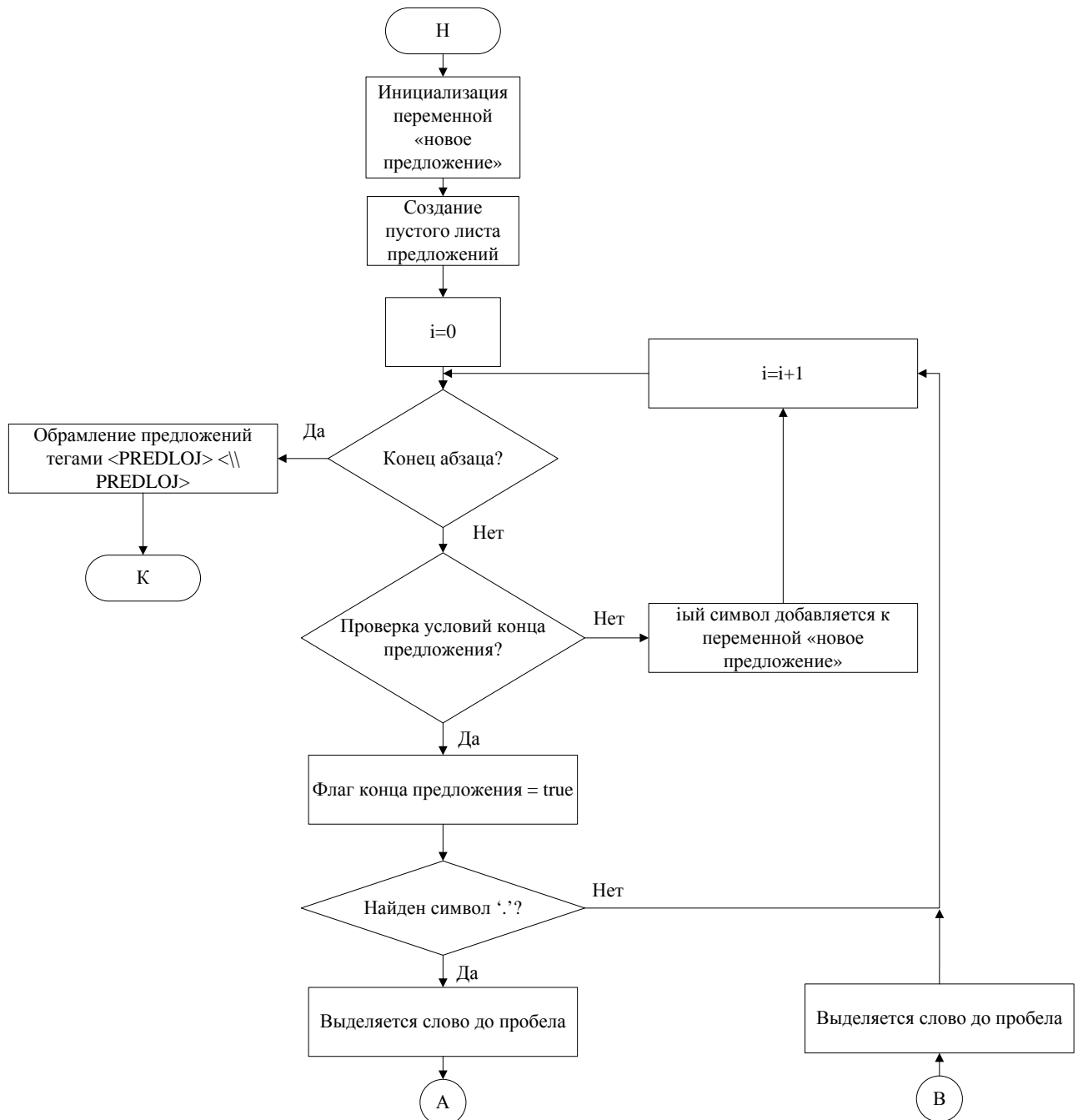


Рисунок 3.4 – Схема алгоритма разбиения абзаца ЭНТД на предложения

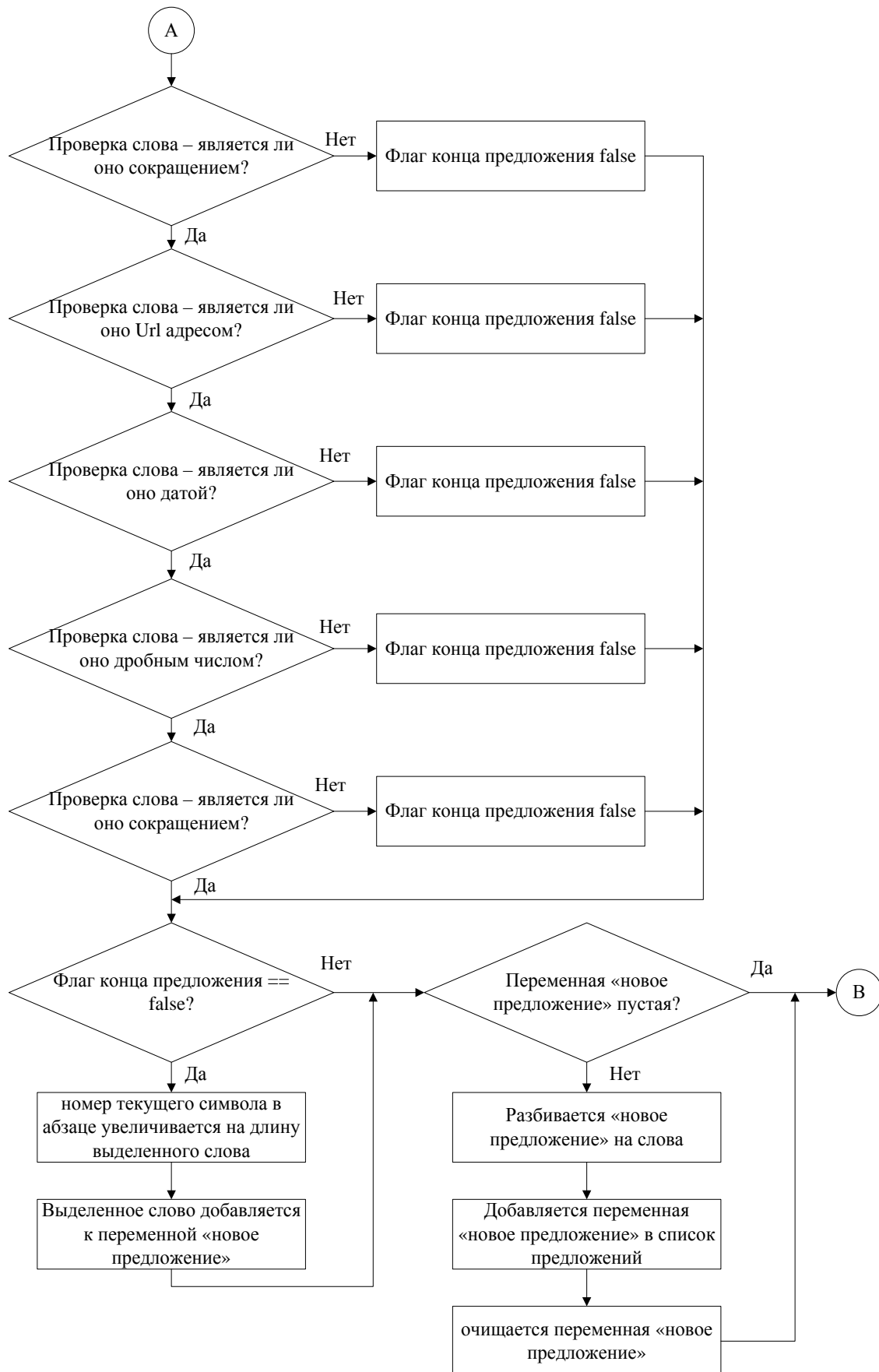


Рисунок 3.5 – Схема алгоритма разбиения абзаца ЭНТД на предложения
(продолжение)

В результате реализации данного алгоритма ЭНТД представляется в виде набора последовательно расположенных предложений ЭНТД.

Конечная процедура сегментации текстового документа заключается в выделении отдельных слов (рисунок 3.6.), что необходимо для выполнения следующих этапов метода – морфологического и синтаксического анализа.

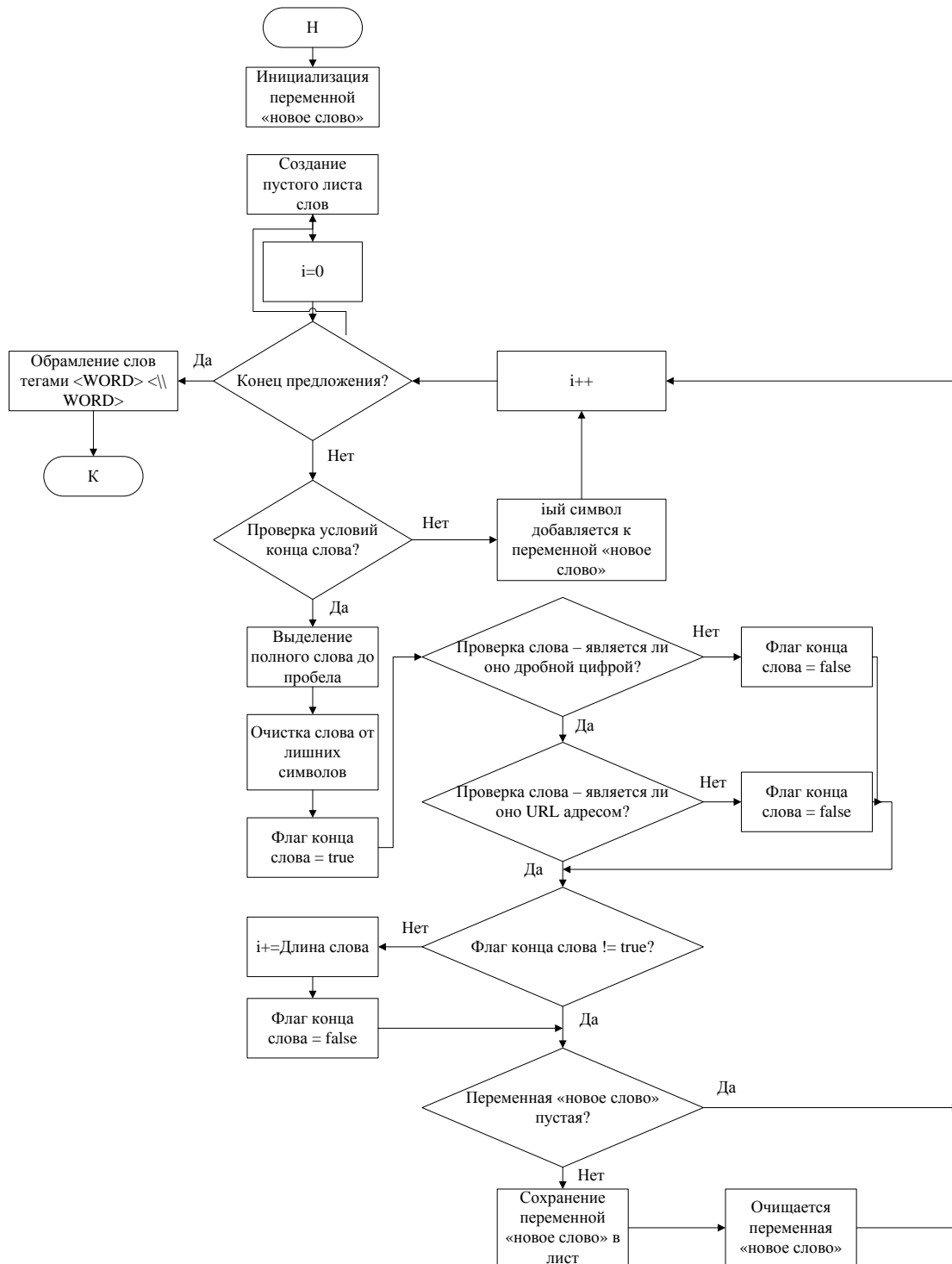


Рисунок 3.6 – Схема алгоритма разбиения предложения ЭНТД на слова

В результате реализации данного алгоритма ЭНТД представляется в виде набора последовательно расположенных слов ЭНТД.

Вторым подготовительным этапом метода является этап морфологического анализа слов ЭНТД. Входной информацией для этого этапа является сегментированный ЭНТД V_k^s , а выходной – ЭНТД, содержащий морфологические характеристики слов V_k^M . Схема алгоритма морфологического анализа слов ЭНТД представлена на рисунке 3.7.

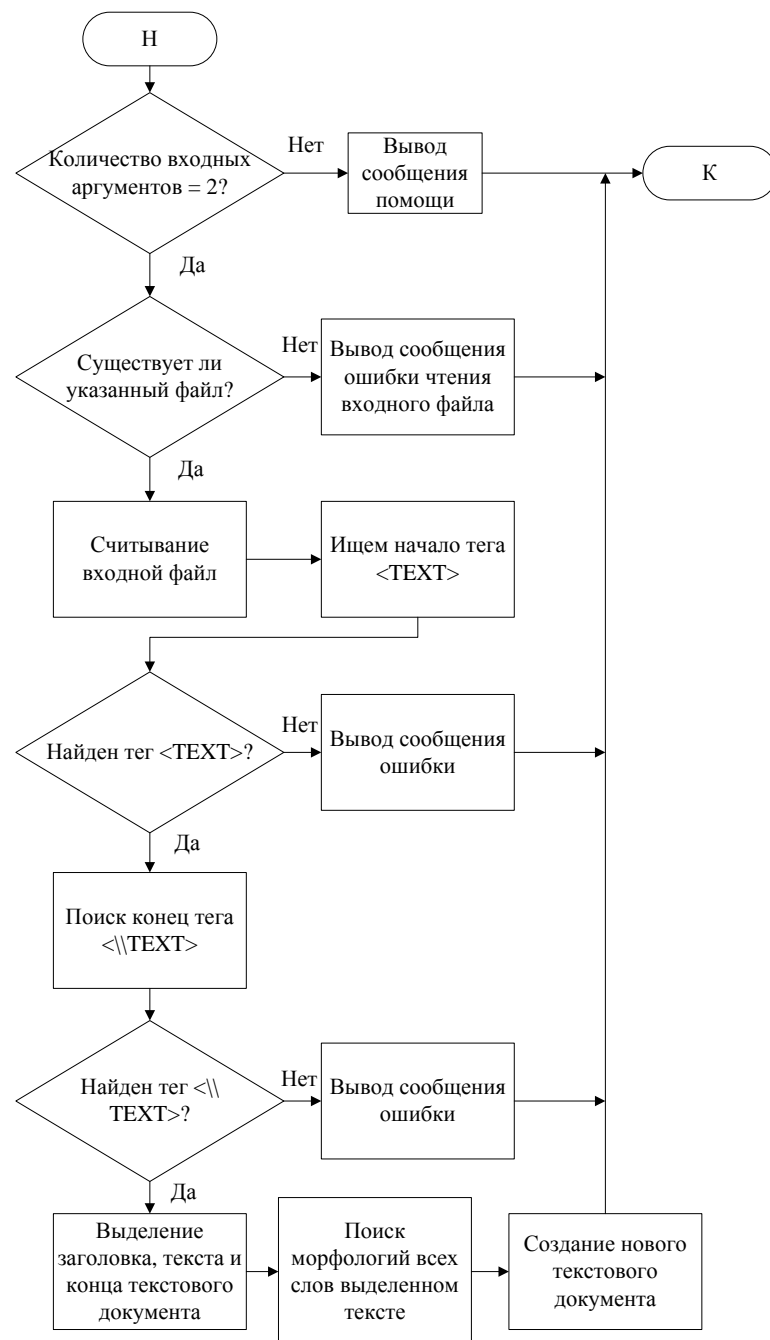


Рисунок 3.7 – Схема алгоритма морфологического анализа слов ЭНТД

В свою очередь, указанный алгоритм морфологического анализа слов включает в себя процедуру определения морфологических характеристик этих слов ЭНТД (рисунки 3.8 и 3.9). Данная процедура необходима для дальнейшего корректного вычисления статистических характеристик и весовых коэффициентов значимых слов, а также для синтаксического анализа слов ЭНТД.

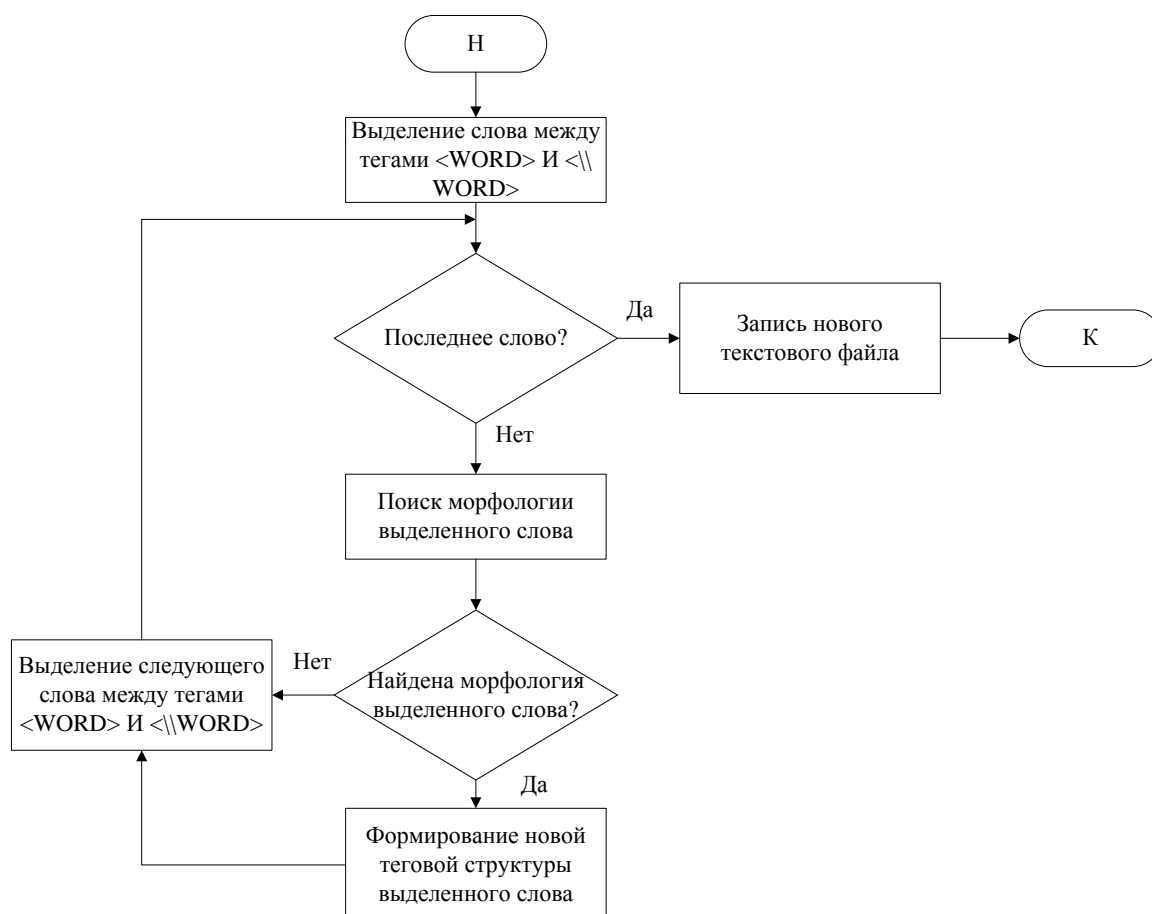


Рисунок 3.8 – Схема алгоритма определения морфологических характеристик слов ЭНТД

В результате формируется набор последовательно расположенных слов ЭНТД, с заполненными морфологическими характеристиками и обрамленные в специальные теги (см. подраздел 2.1).

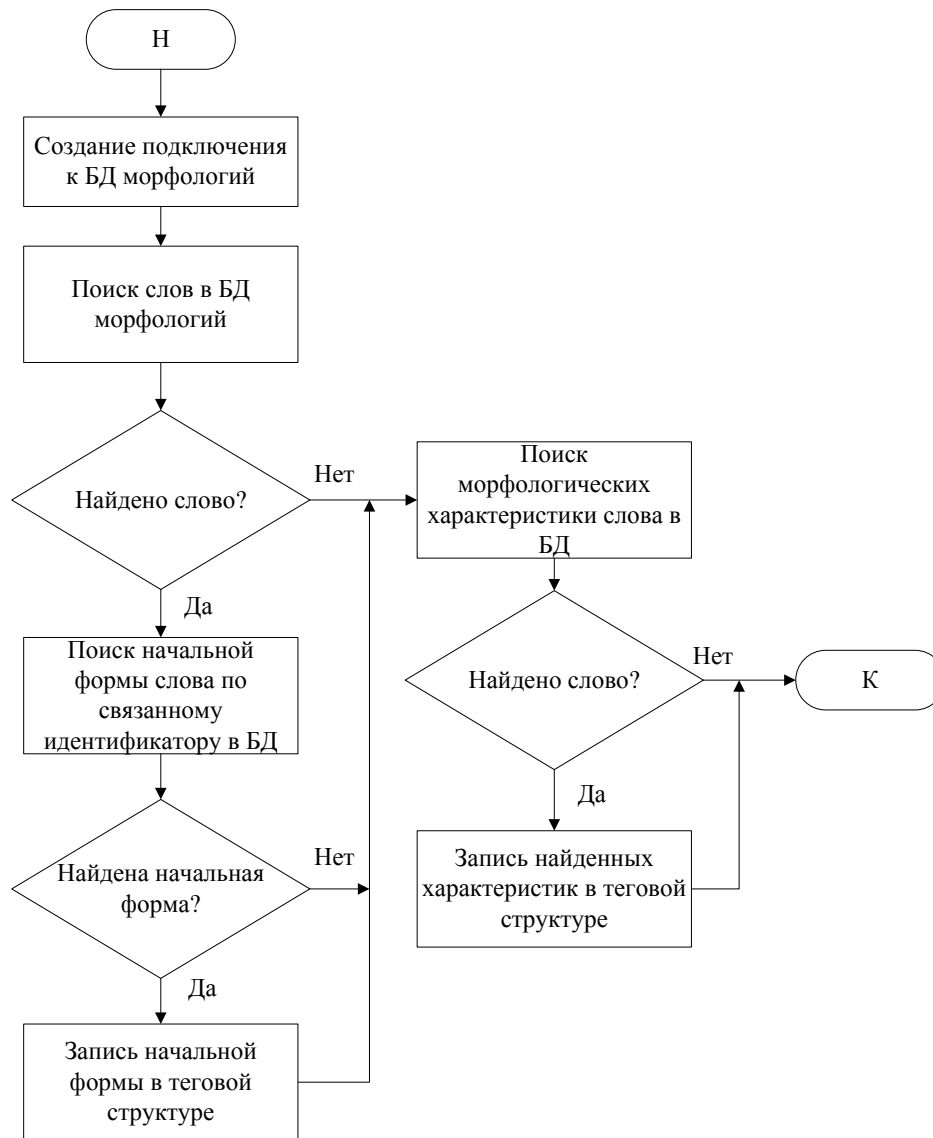


Рисунок 3.9 – Схема алгоритма поиска морфологических характеристик выделенного слова

Для реализации процедур морфологического анализа ЭНТД необходимо сформировать базы данных слов с их морфологическими и лингвистическими характеристиками (базы морфологий) с использованием следующих трёх источников:

во-первых, текстового файла (словаря) *terms.txt*, содержащего около 80 000 строк, каждая строка представляет из себя совокупность слов в форме разных падежей, склонений, числе и роде. Каждая строка имеет следующий вид:

- дворник дворника дворнику дворника дворником дворнике дворники дворников дворникам дворников дворниками дворниках;

- отопитель отопителя отопителю отопитель отопителем отопителе отопителем отопителей отопителям отопители отопителями отопителях;
- водоснабжение, водоснабженье водоснабжения, водоснабженье водоснабжению, водоснабженье водоснабжение водоснабжением, водоснабженьем водоснабжении, водоснабженье водоснабжения, водоснабженье водоснабжений, водоснабженье водоснабжениям, водоснабженьям водоснабжения водоснабжениями, водоснабженьями водоснабжениях, водоснабженьях;
- электроснабжение, электроснабженье электроснабжения, электроснабженье электроснабжению, электроснабженье электроснабжение электроснабжением, электроснабженьем электроснабжении, электроснабженье электроснабжения, электроснабженье электроснабжений, электроснабженье электроснабжениям, электроснабженьям электроснабжения электроснабжениями, электроснабженьями электроснабжениях, электроснабженьях;

во-вторых, текстового файла (словаря) morfSlovar.txt, который также содержит слова в различных формах и включает в себя в настоящее время 4197236 строк. Строка представляет из себя непосредственно слово, а далее идёт описание его морфологических характеристик и указатель на начальную форму. Каждая строка имеет следующий вид:

- дворники сущ мн им 1282392;
- отопителями сущ неод мн тв 1325956;
- электроснабжениях сущ неод мн пр 1574230;

в-третьих, национального корпуса русского языка СинТагРус, содержащего в настоящее время более 104000 словоформ вида:

- <W DOM="2" FEAT="V COB СТРАД ПРИЧ ПРОШ ЕД СРЕД ИМ" ID="1" LEMMA="СОГЛАСОВЫВАТЬ" LINK="опред">Согласованное</W>;

- <W DOM="17" FEAT="S ЕД СРЕД ИМ НЕОД" ID="2"
LEMMA="РЕШЕНИЕ" LINK="предик">решение</W>;
- <W DOM="5" FEAT="А ЕД ЖЕН РОД" ID="3"
LEMMA="РЕГИОНАЛЬНЫЙ" LINK="опред">Региональной</W>;
- <W DOM="5" FEAT="А ЕД ЖЕН РОД" ID="4"
LEMMA="ЭНЕРГЕТИЧЕСКИЙ" LINK="опред">энергетической</W>;
- <W DOM="2" FEAT="S ЕД ЖЕН РОД НЕОД" ID="5"
LEMMA="КОМИССИЯ" LINK="квазиагент">комиссии</W>.

В тэг <W> заключено слово с его морфологическими характеристиками, которые описаны в значении атрибута "FEAT", а начальная форма слова в значении атрибута "LEMMA".

Для хранения данных о морфологических характеристиках слов ЭНТД целесообразно использовать базу MorfAnalysDB2 с таблицей tbWords, поля которой представлены на рисунке 3.10.

MorfAnalysDB2

tbWords
id
feat
lemma
link
text

Рисунок 3.10 – Поля таблицы tbWords морфологической базы данных

Краткое описание полей tbWords представлено в таблице 3.1.

Таблица 3.1 – Поля электронной таблицы tbWords

Название	Тип	Описание
Id	uniqueidentifier	Первичный ключ
Feat	nvarchar(50)	Морфологические характеристики
Lemma	nvarchar(50)	Морфема слова
Link	nvarchar(50)	Часть речи
Text	nvarchar(50)	Само слово

Для хранения сведений о морфологических характеристиках слов ЭНТД целесообразно сформировать базу данных MorfAnalysDB с двумя связными таблицами tbWords и tbWordsParams, поля которых представлены на рисунке 3.11.

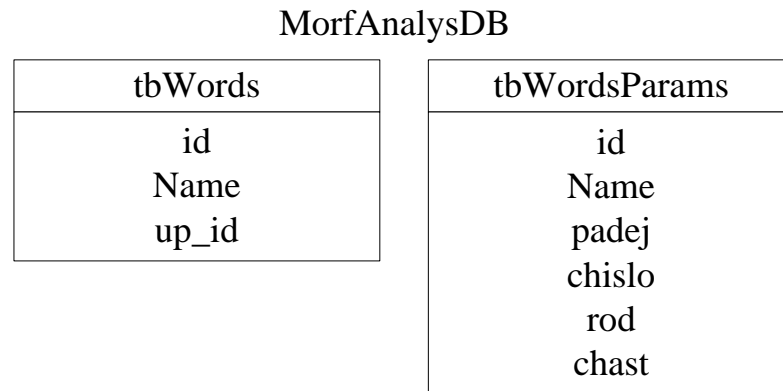


Рисунок 3.11 – Сущности tbWords и tbWordsParams, позволяющие хранить морфологические характеристики слов

Краткое описание полей tbWords и tbWordsParams представлено в таблицах 3.2 и 3.3, соответственно.

Таблица 3.2 – Поля таблицы tbWords

Название	Тип	Описание
Id	uniqueidentifier	Первичный ключ
Name	nvarchar(50)	Само слово
up_id	uniqueidentifier (50)	Идентификатор морфемы

Таблица 3.3 – Поля таблицы tbWordsParams

Название	Тип	Описание
Id	uniqueidentifier	Первичный ключ
Name	nvarchar(50)	Само слово
Padej	nvarchar(50)	Падеж
Chislo	nvarchar(50)	Число
Rod	nvarchar(50)	Род
Chast	nvarchar(50)	Часть речи

Для хранения данных о рубриках и весовых и статистических характеристиках значимых слов необходимо сформировать базу данных AnalysDB с двумя связными таблицами tbInfBlocks и tbFreqAndWeightOfWords, поля которых представлены на рисунке 3.12.

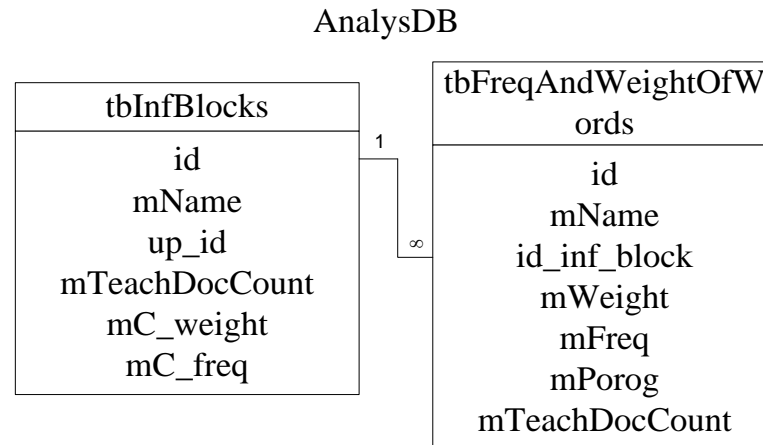


Рисунок 3.12 – Поля таблиц tbInfBlocks и tbFreqAndWeightOfWords

Краткое описание полей tbInfBlocks представлено в таблице 3.4.

Таблица 3.4 – Поля таблицы tbInfBlocks

Название	Тип	Описание
Id	uniqueidentifier	Первичный ключ
mName	nvarchar(50)	Название рубрики
Up_id	uniqueidentifier (50)	Идентификатор родительской рубрики
mTeachDoc-Count	nvarchar(50)	Количество документов для обучения по данной рубрики
mC_weight	nvarchar(50)	Количество известных весовых коэффициентов
mC_freq	nvarchar(50)	Количество известных частотных характеристик

Краткое описание полей tbFreqAndWeightOfWords представлено в таблице 3.5.

Таблица 3.5 – Поля таблицы tbFreqAndWeightOfWords

Название	Тип	Описание
Id	uniqueidentifier	первичный ключ
mName	nvarchar(50)	Само слово
Id_inf_block	nvarchar(50)	Идентификатор рубрики
mWeight	nvarchar(50)	Весовой коэффициент
mFreq	nvarchar(50)	Частотная характеристика
mPorog	nvarchar(50)	Пороговое значение частотной характеристики
mTeachDocCount	nvarchar(50)	Количество документов, в которых найдено данное слово по конкретной рубрике

Разработанные в данном подразделе алгоритмы позволяют осуществить необходимую подготовку ЭНТД для формализации и использования необходимой модели рубрицирования.

3.2 Алгоритмы для анализа коротких электронных неструктурированных текстовых документов на основе нейро-нечеткого классификатора с использованием весовых коэффициентов

Для построения и применения для анализа коротких ЭНТД нейро-нечеткого классификатора с использованием весовых коэффициентов (см. подраздел 2.2) разработаны рассмотренные ниже алгоритмы: вычисления весовых коэффициентов значимых слов тезаурусов рубрик; построения, обучения и использования нейро-нечеткого классификатора.

Схема алгоритма вычисления весовых коэффициентов значимых слов тезаурусов рубрик представлена на рисунке 3.13.

Входными данными для этого алгоритма является множество рубрик R и обучающая выборка документов $V^{(об)}$, а в результате работы алгоритма настраиваются весовые коэффициенты r_{mj} значимых слов тезаурусов рубрик, что обеспечивает корректное представление характеристик ЭНТД и работу нейро-нечеткого классификатора при анализе ЭНТД в целом.

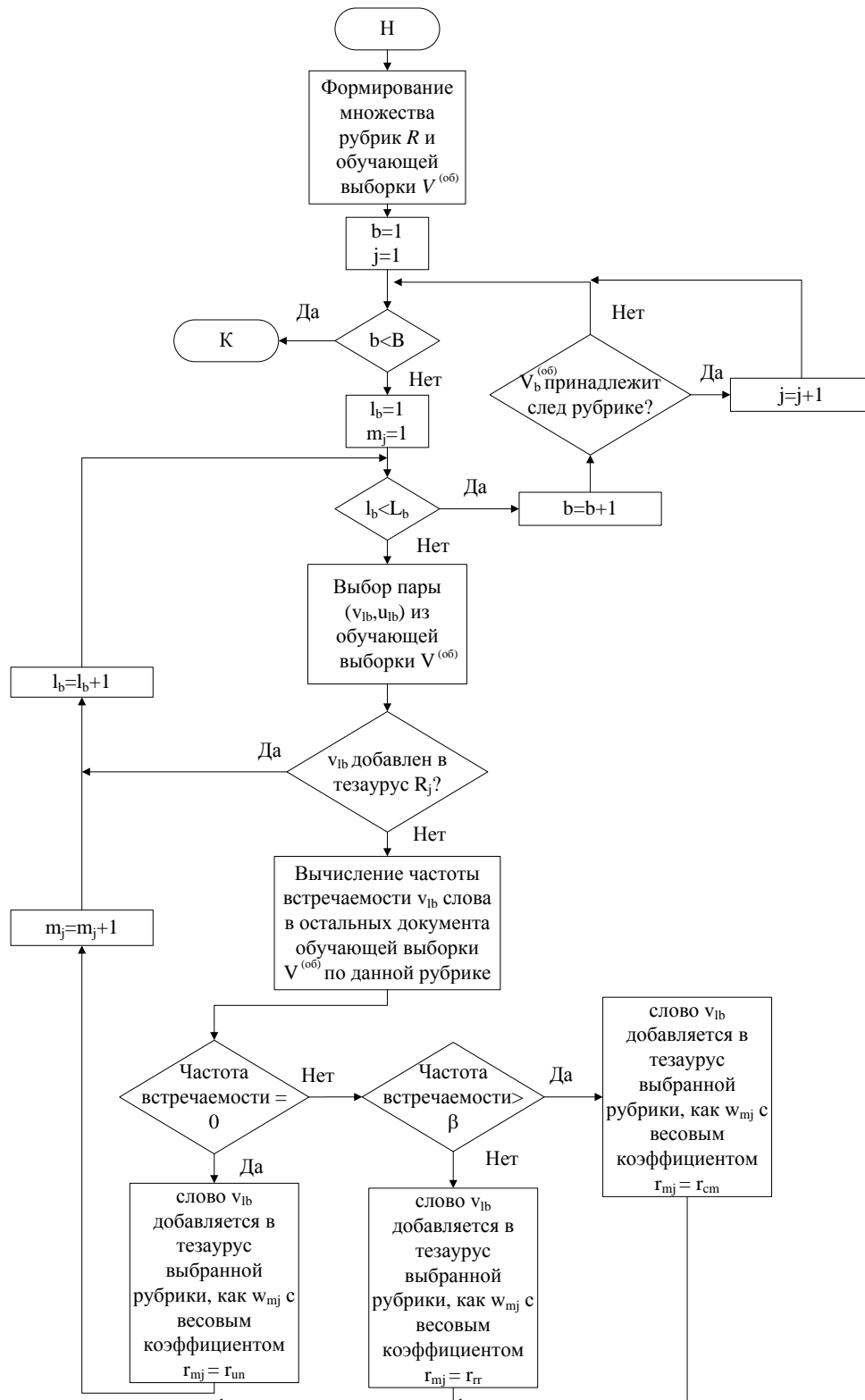


Рисунок 3.13 – Схема алгоритма вычисления весовых коэффициентов значимых слов рубрик

На рисунке 3.14 представлен алгоритм, реализующий модель рубрицирования с использованием весовых коэффициентов ЗС ЭНТД.

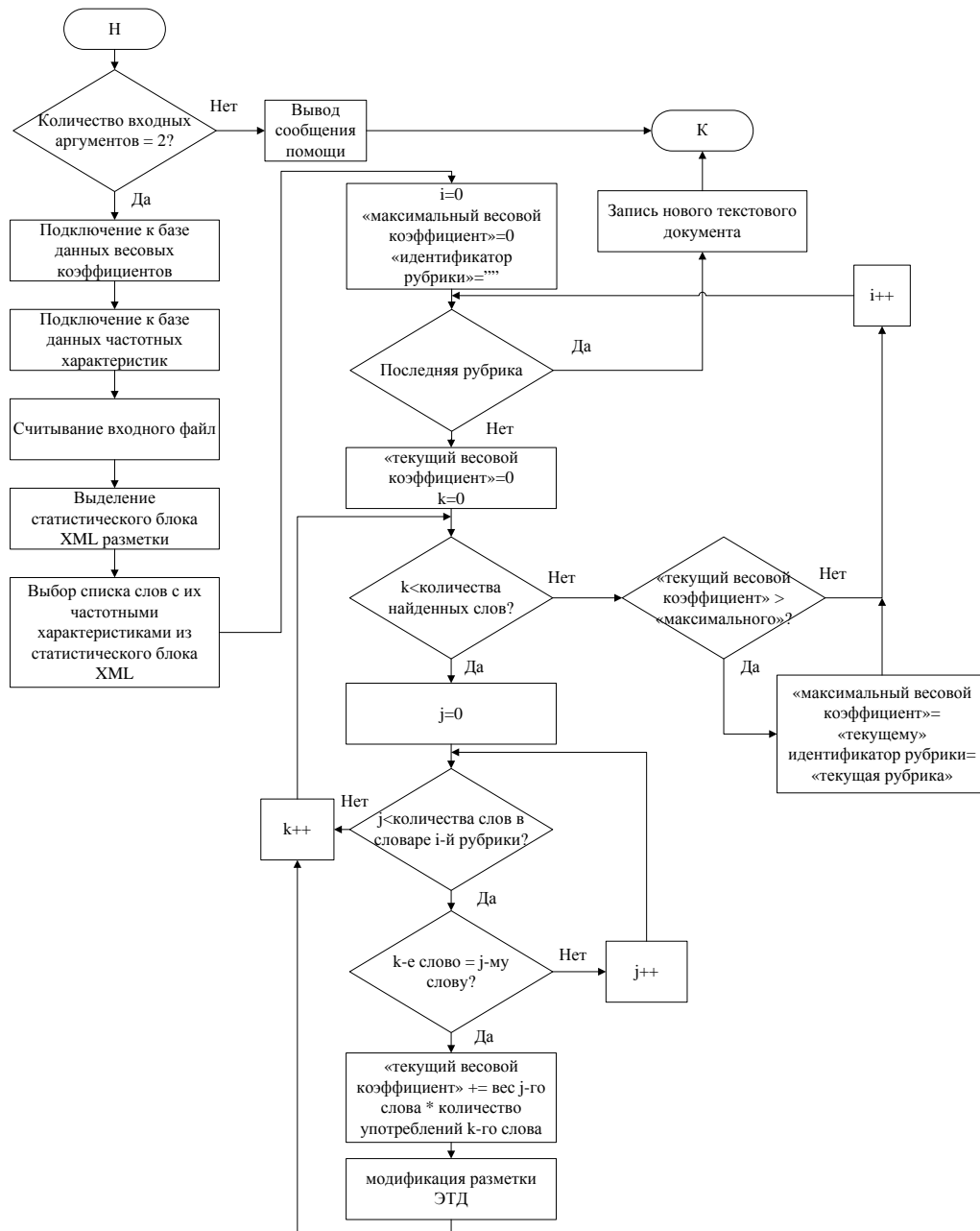


Рисунок 3.14 – Схема алгоритма, реализующего модель рубрицирования с использованием весовых коэффициентов ЗС ЭНТД

В результате данного алгоритма формируется степень принадлежности ЭНТД к наиболее близкой рубрике.

Формализация ЭНТД для нейро-нечеткой сети (см. подраздел 2.2) требует проведения синтаксического анализа. Для выполнения данного этапа подходит

синтаксический парсер MaltParser, который использует специальный формат текстовых документов, описанный следующим XML файлом:

```
<?xml version="1.0" encoding="UTF-8"?>
<dataformat name="conllx">
  <column name="ID" category="INPUT" type="INTEGER"/>
  <column name="FORM" category="INPUT" type="STRING"/>
  <column name="LEMMA" category="INPUT" type="STRING"/>
  <column name="CPOSTAG" category="INPUT" type="STRING"/>
  <column name="POSTAG" category="INPUT" type="STRING"/>
  <column name="FEATS" category="INPUT" type="STRING"/>
  <column name="HEAD" category="HEAD" type="INTEGER"/>
  <column name="DEPREL" category="DEPENDENCY_EDGE_LABEL" type="STRING"/>
  <column name="PHEAD" category="IGNORE" type="INTEGER" default="_"/>
  <column name="PDEPREL" category="IGNORE" type="STRING" default="_"/>
</dataformat>
```

Таблица 3.6 – Описание полей и свойств формата MaltParser

Атрибут	Значения атрибута	
category	Категория столбца, одно из следующих:	
	INPUT	Вводите данные как в режиме обучения, так и в парсере, например, в темах части речи или в словарных формах
	DEPENDENCY_EDGE_LABEL	Столбец с меткой зависимости. Если анализатор должен научиться создавать маркированные диаграммы зависимостей, они должны иметься в режиме обучения
category	Категория столбца, одно из следующих:	
	OUTPUT	Такой же столбец, как и DEPENDENCY_EDGE_LABEL, который использовался в MaltParser версий 1.0, 1.1
	PHRASE_STRUCTURE_EDGE_LABEL	Столбец, содержащий метку края фразовой структуры
	PHRASE_STRUCTURE_NODE_LABEL	Столбец с меткой категории фраз
	SECONDARY_EDGE_LABEL	Столбец, содержащий метку вторичного края

Продолжение таблицы 3.6 – Описание полей и свойств формата MaltParser

category	Категория столбца, одно из следующих:	
	HEAD	Столбец HEAD определяет немаркированную структуру графа зависимостей и также выводит данные анализатора в режиме синтаксического анализа
	IGNORE	Значение столбца будет проигнорировано и, следовательно, не будет присутствовать в выходном файле
type	Определяет тип данных столбца и/или его обработку во время обучения и разбора:	
	STRING	Значение столбца будет использоваться как строковое значение в модели признаков.
	INTEGER	Значение столбца будет использоваться в качестве целочисленного значения в модели функций.
	BOOLEAN	Значение столбца будет использоваться как логическое значение в модели функций.
	REAL	Значение столбца будет использоваться как реальное значение в модели функций.
default	The default output for columns that have the column type IGNORE.	

Пример проанализированного текстового предложения:

- 1 Покрашенная покрашенный P P P—nsna 3 опред __
- 2 недавно недавно A A Afpnsnf3 опред __
- 3 стена стена N N Ncnsnn 4 предик __
- 4 облезает облезть V V Vmip3s-a-e 0 ROOT __

На рисунке 3.15 представлена схема алгоритма обучения нейро-нечеткого классификатора, а на рисунке 3.16 – алгоритма рубрицирования.

В результате использования алгоритма, представленного на рисунке 3.16, определяется максимальная степень принадлежности ЭНТД к наиболее близкой рубрике. Данный алгоритм позволяет повысить качество анализа и точность рубрицирования коротких ЭНТД за счет использования предложенного нейро-нечеткого классификатора при условии незначительной степени пересечения рубрик и достаточного объема статистической информации о документах данного типа для обучения этого классификатора.

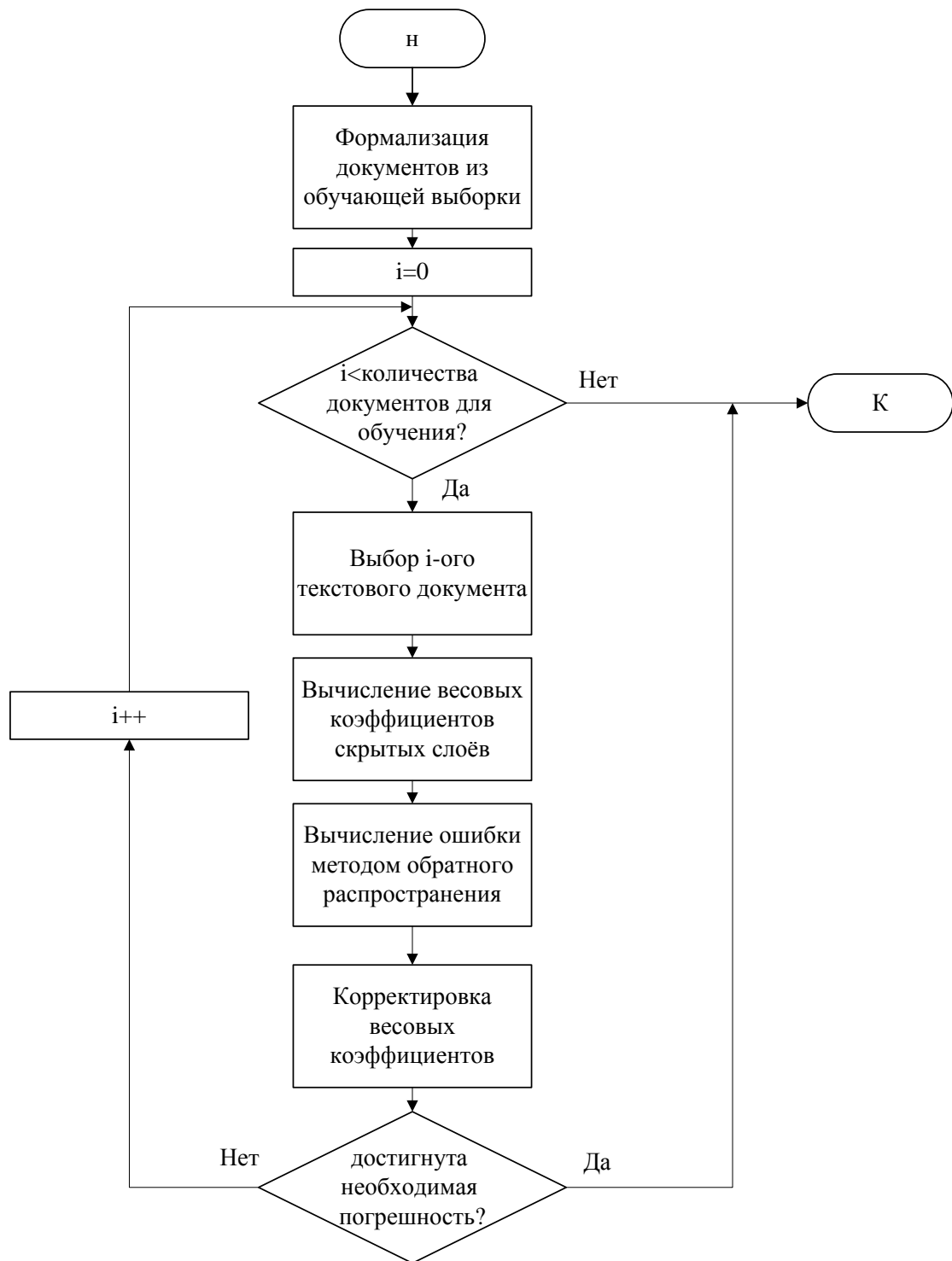


Рисунок 3.15 – Схема алгоритма обучения нейро-нечеткого классификатора



Рисунок 3.16 – Схема алгоритма рубрицирования на основе нейро-нечеткого классификатора

3.3 Алгоритмы для анализа коротких электронных неструктурированных текстовых документов на основе нечетких деревьев решений

Для построения модели анализа и рубрицирования ЭНТД на основе нечеткого дерева решений (см. подраздел 2.3) реализована схема, представленная на рисунке 3.17.

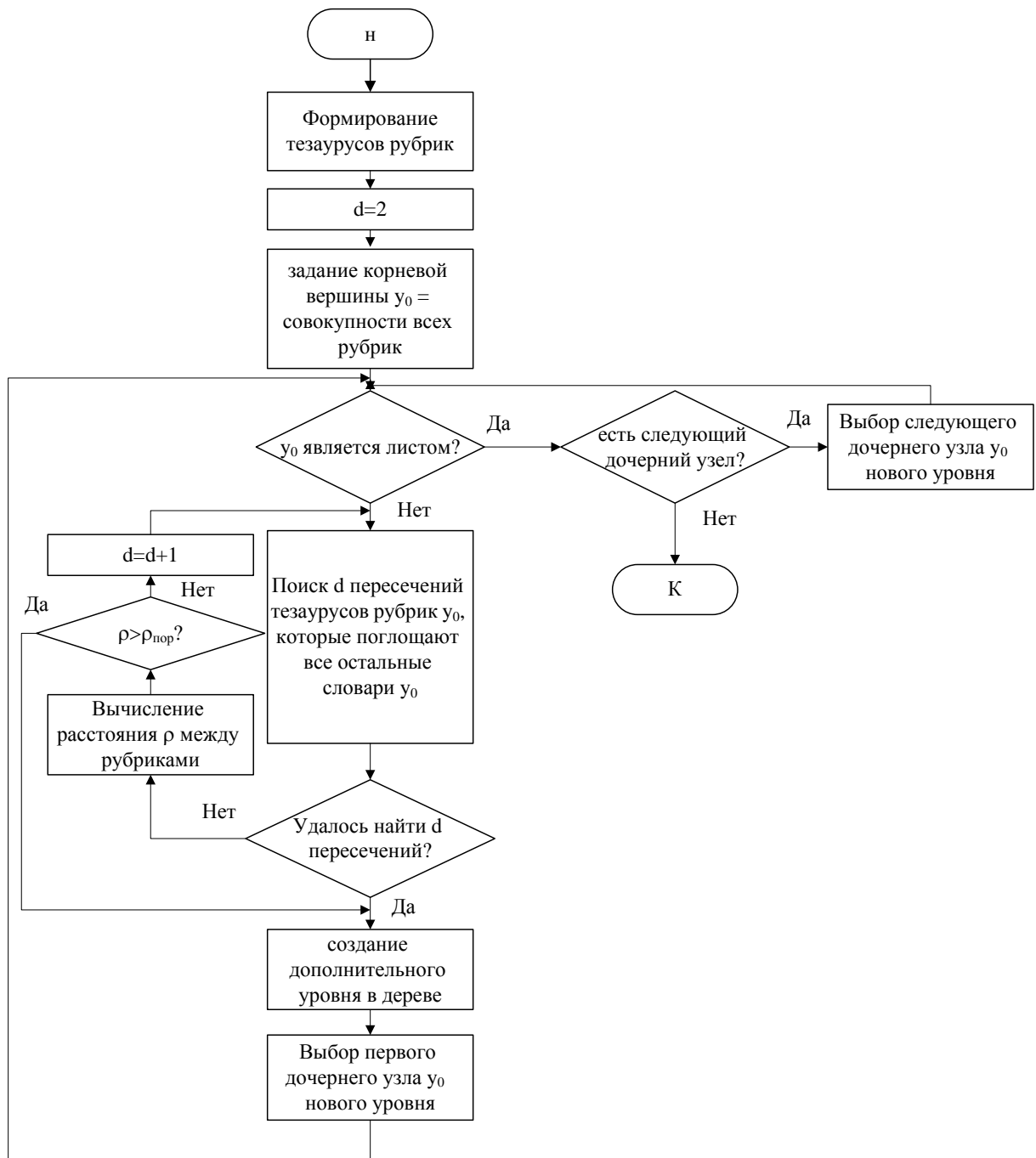


Рисунок 3.17 – Схема алгоритма построения нечеткого дерева решений

На вход данного алгоритма поступают тезаурусы рубрик, а результатом его работы является нечеткое дерево решений, позволяющее рубрицировать ЭНТД.

Процедура рубрицирования на основе построенного дерева решений показана на рисунке 3.18.

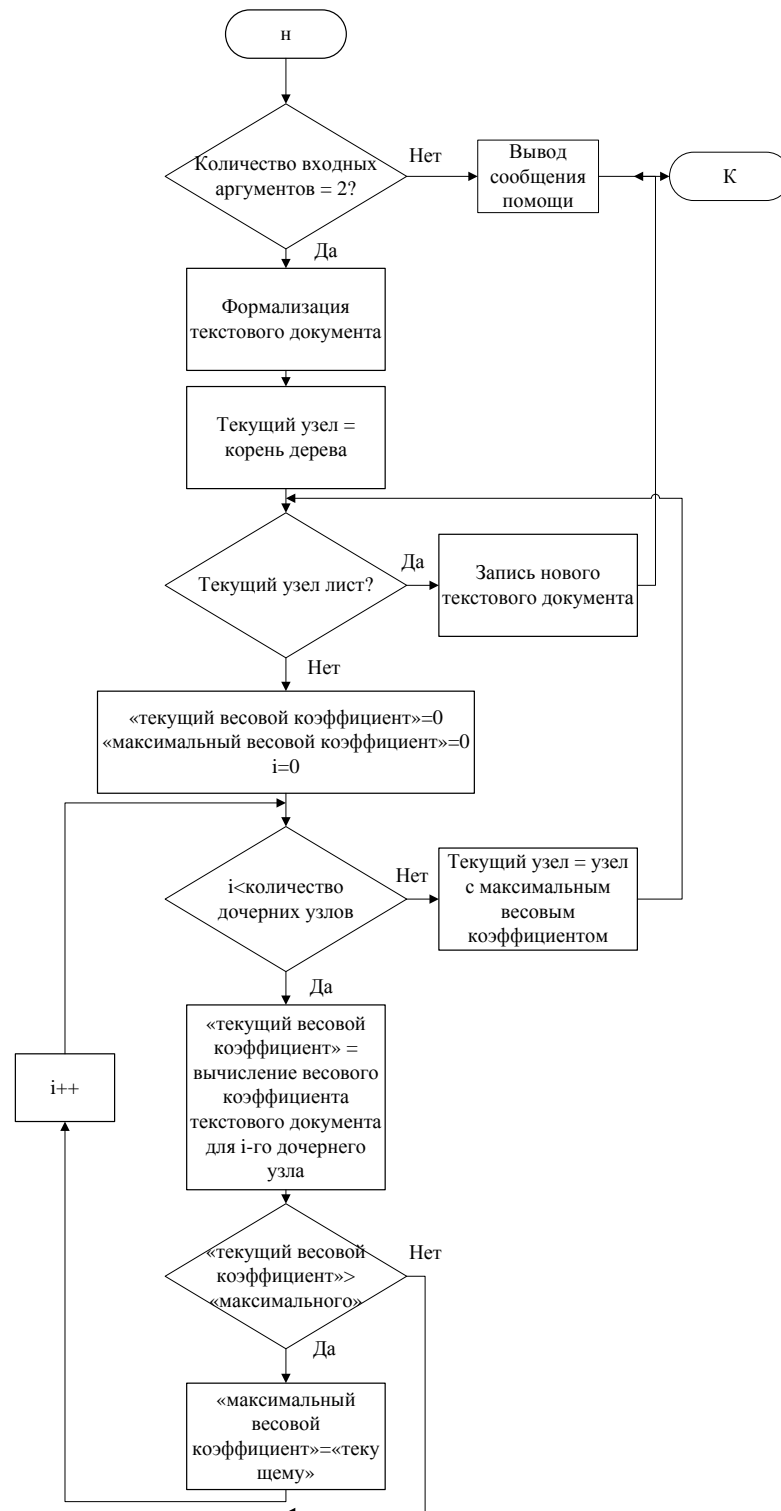


Рисунок 3.18 – Алгоритм рубрицирования ЭНТД на основе нечеткого дерева решений

В результате этого алгоритма определяется максимальная степень принадлежности ЭНТД к наиболее близкой рубрике.

Данный алгоритм позволяет повысить точность рубрицирования ЭНТД

среднего размера в условиях существенной степени пересечения рубрик и отсутствия статистической информации о рубрицируемых документах.

На рисунке 3.19 представлен алгоритм, реализующий метод изменения моделей рубрицирования ЭНТД.

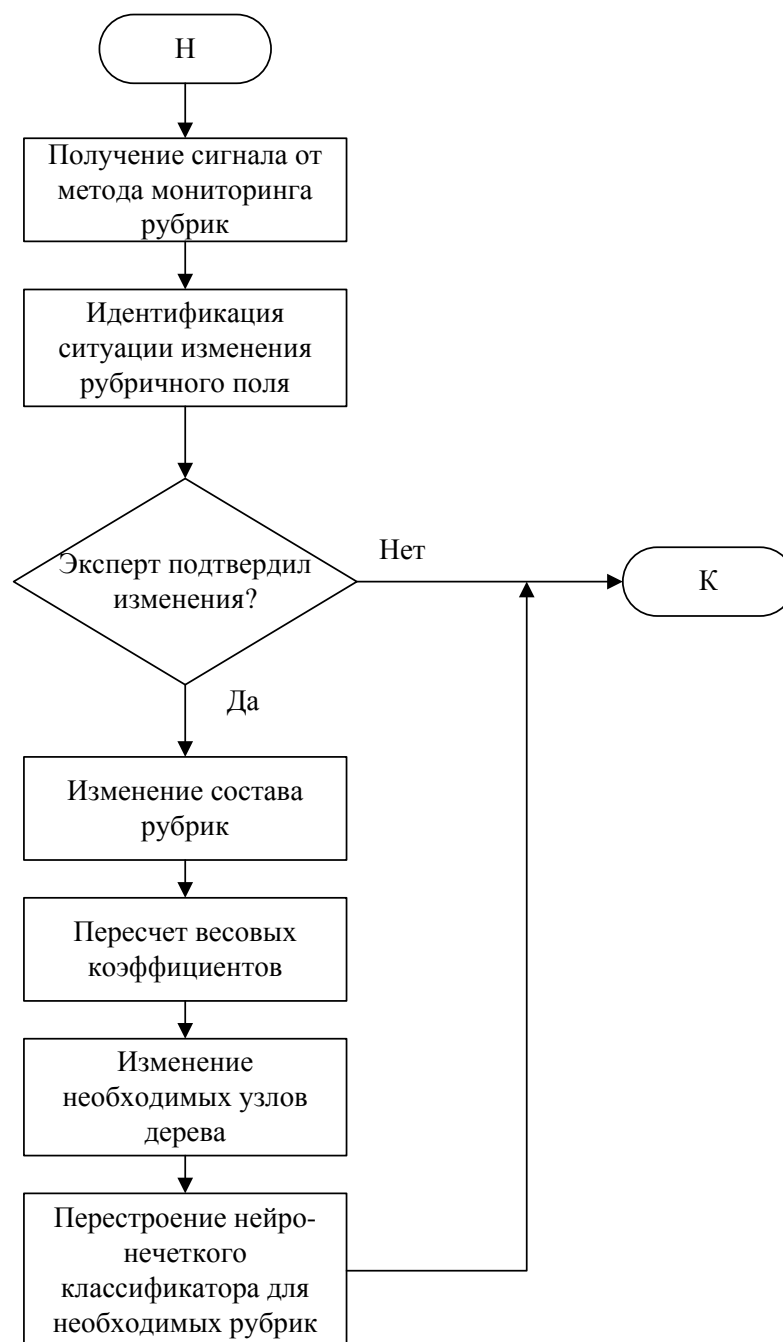


Рисунок 3.19 – Алгоритм динамического изменения моделей рубрицирования

В результате работы данного алгоритма выполняется адаптация моделей рубрицирования к новым условиям, что позволяет поддерживать актуальность состава рубричного поля.

3.4 Выводы по главе

Разработан алгоритм предварительного анализа электронных неструктурированных текстовых документов, который включает в себя такие этапы анализа, как сегментация, морфологический анализ, синтаксический анализ. Синтаксический анализ осуществляется при помощи синтаксического мультязычного парсера MalpParser. Морфологический словарь, а так же модель русского языка для парсера построены на основе Национального корпуса русского языка СинТагРус. Данный алгоритм предварительного анализа позволяет использовать предложенные для мультимодельного метода анализа классификаторы.

Разработан алгоритм классификации коротких электронных неструктурированных текстовых документов на основе нейро-нечеткой сети, входами для которого является формализованное представление текстового документа в виде весовых коэффициентов первых двадцати ключевых слов, а выходом - степени принадлежности данного текстового документа существующим рубрикам.

Разработан алгоритм классификации коротких электронных неструктурированных текстовых документов на основе нечеткого дерева решений, входом для которого являются формализованное представление текстового документа в виде весовых коэффициентов всех ключевых слов, а также их синтаксические роли в предложениях и синтаксические связи, выходом являются степени принадлежности данного текстового документа существующим рубрикам.

4 РЕЗУЛЬТАТЫ ПРАКТИЧЕСКОГО ИСПОЛЬЗОВАНИЯ АЛГОРИТМОВ АНАЛИЗА (РУБРИЦИРОВАНИЯ) ЭЛЕКТРОННЫХ НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ

4.1 Структура средств информационной системы автоматизированного анализа электронных неструктурированных текстовых документов

Для практической реализации предложенных мультимодельного метода и нечетких моделей разработана архитектура информационной системы автоматизированного рубрицирования электронных текстовых документов в условиях изменения рубрик Artex 1.0, основные элементы которой приведены на рисунке 4.1.

Для реализации разработанной ИС автоматизированного рубрицирования ЭНТД, представляемых на естественном языке, создана dll библиотека, которая содержит основные классы, процедуры и функции, позволяющие системе корректно функционировать.

На рисунках 4.2 и 4.3 представлена диаграмма основных классов dll библиотеки разработанного программного средства XMLibrary.dll.

На рисунке 4.2 представлены классы для реализации алгоритмов управления рубриками и словами в документах и словарях. Из рисунка видно, что разработанная библиотека обладает универсальностью и позволяет централизованно модифицировать разработанные алгоритмы, классы и т.п.

Класс TInfBlocks содержит информацию о рубриках, которые включаются в позволяющий ими управлять класс TListInfBlocks.

К каждой рубрике относятся ЭНТД, описанные классом TListDoc, состоящие из отдельных текстовых документов TDoc, который в свою очередь состоит из абзацев TAbzac, предложений TPredloj и слов TWordsFromDocument.

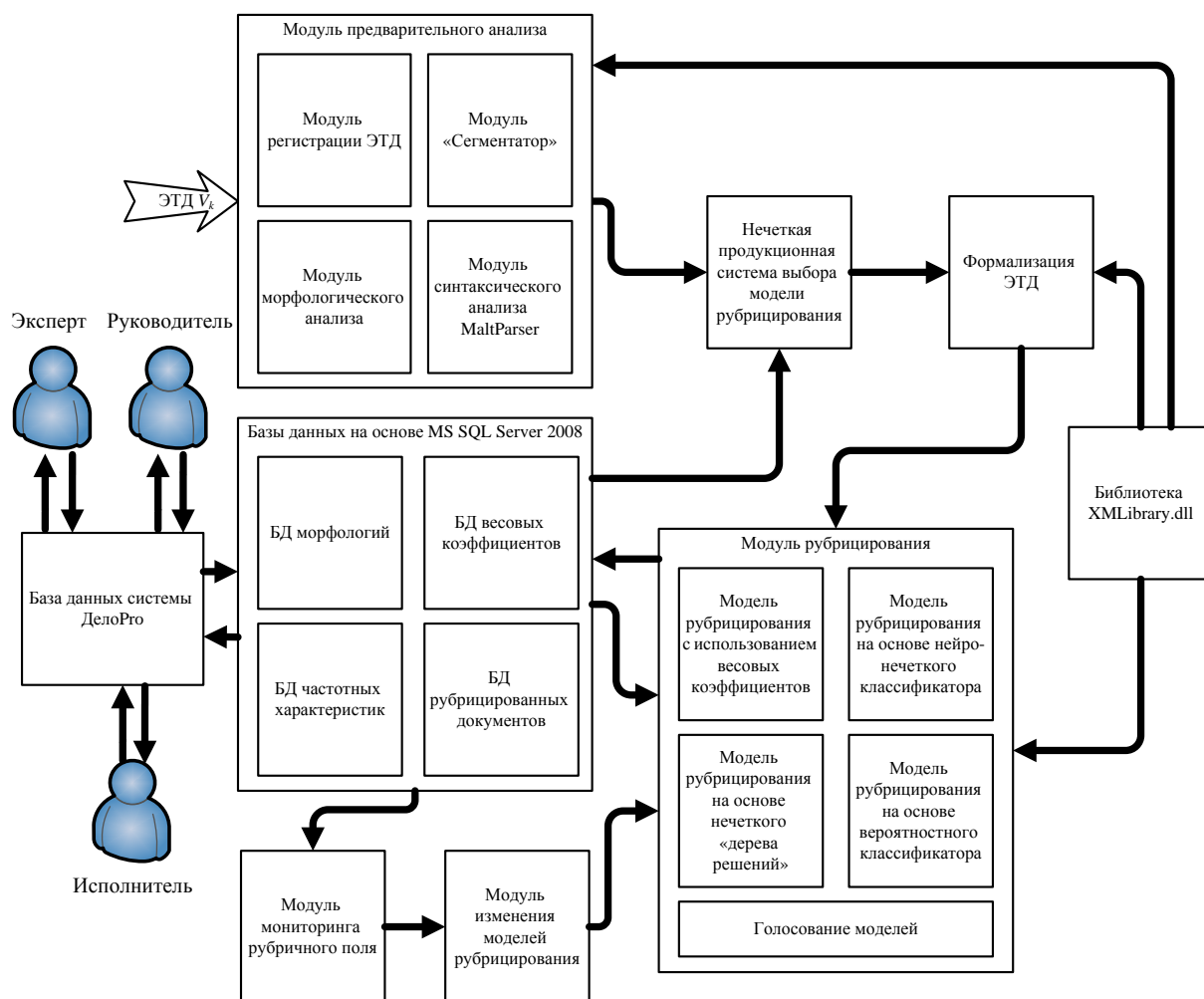


Рисунок 4.1 – Основные компоненты ИС автоматизированного рубрицирования ЭНТД Artex 1.0

Класс `TWordsFromDB` описывает информацию о весовых коэффициентах и статистических характеристиках значимых слов, хранящихся в БД. Класс `TListWordsFromDB` позволяет ими управлять.

Класс `mXMLLibrary` упрощает работу с файлами в формате XML, позволяет считывать и записывать ЭНТД в требуемой разметке, описанной в главе 3.

Класс `mWordsLibrary` осуществляет вспомогательные процедуры и функции для предварительного этапа анализа ЭНТД "Сегментатор", данный класс может быть легко модифицирован для улучшения процедуры "Сегментатор".

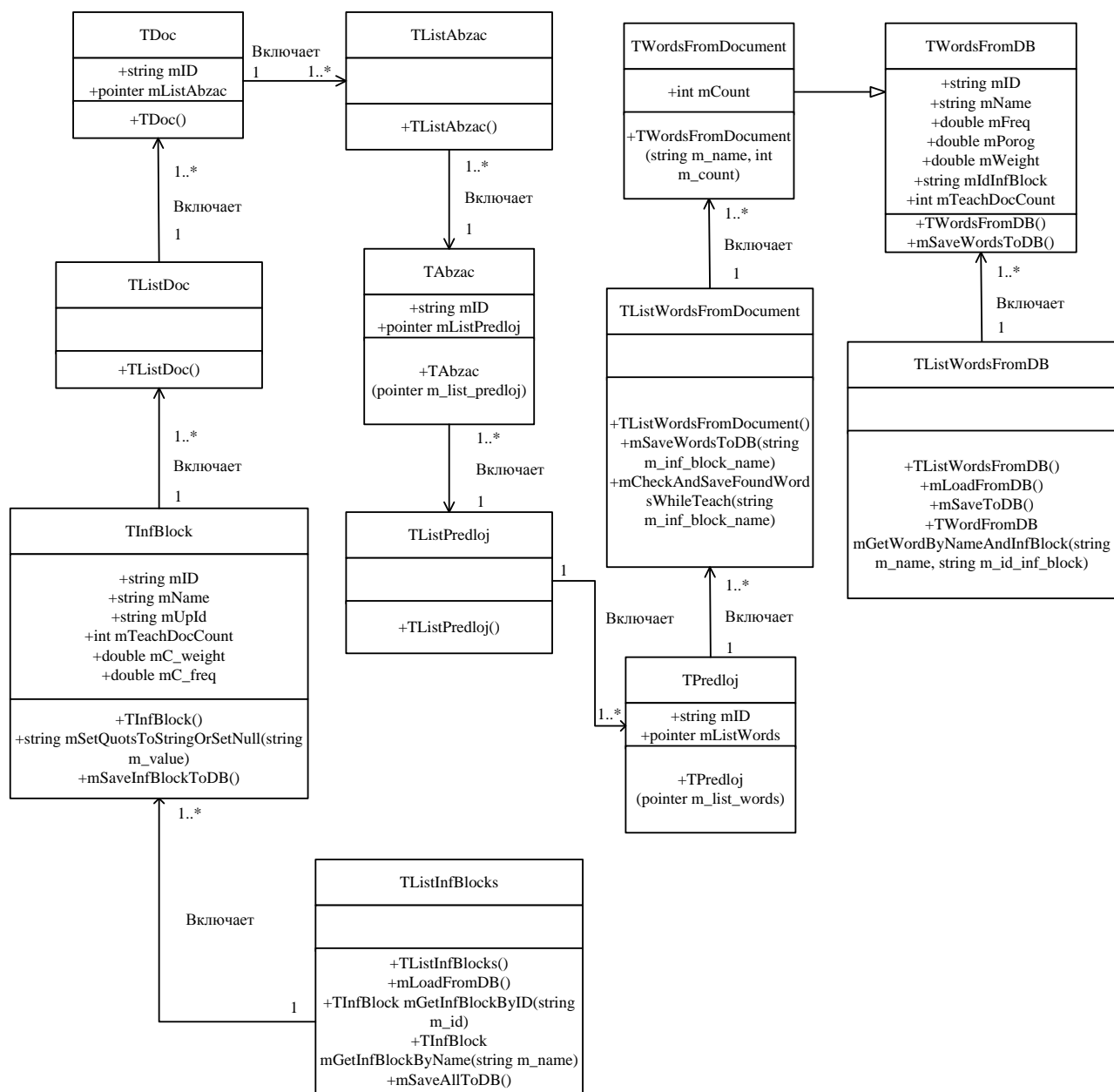


Рисунок 4.2 – Диаграмма классов библиотеки XMLibrary.dll

На рисунке 4.3 представлены классы, реализующие дополнительные вспомогательные процедуры и функции отдельных модулей ИС.

Класс **mFilesLibrary** реализует процедуры считывания и записи текстовых файлов.

Класс **mSQLLibrary** позволяет подключаться к базам данных морфологических словарей, а также к базе данных весовых коэффициентов и статистических характеристик значимых слов рубрик.

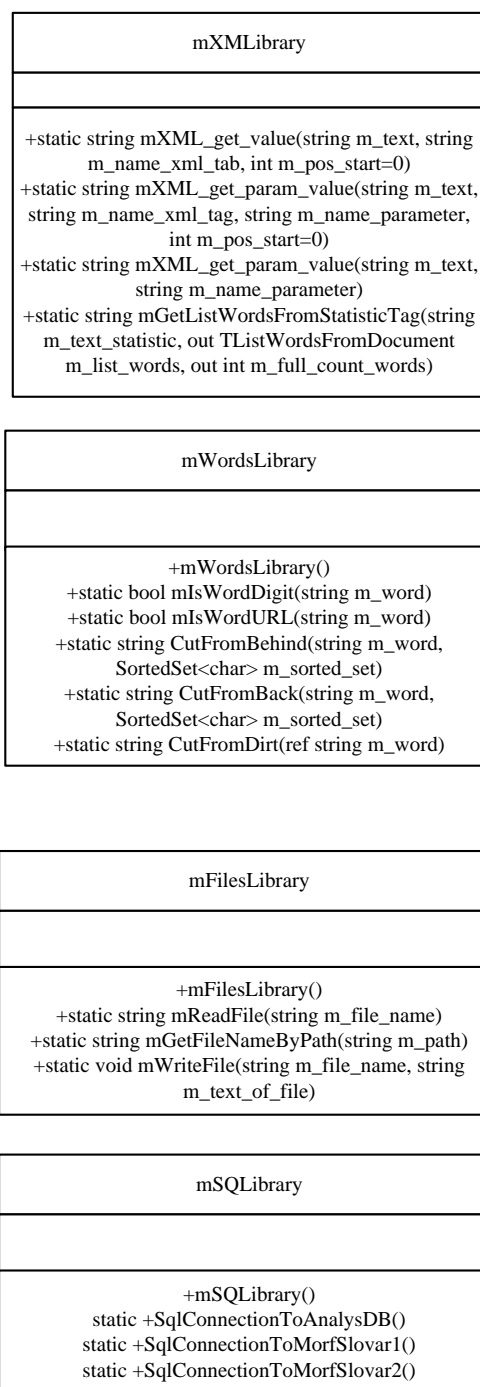


Рисунок 4.3 – Диаграмма классов библиотеки XMLibrary.dll

Из рисунка 4.3 также следует, что разработанная библиотека позволяет централизованно модифицировать правила обработки текстовых файлов, способы подключения и запросы к базам данных.

На рисунке 4.4 представлена UML-диаграмма активности процедуры рубрицирования разработанной системы Artex 1.0.

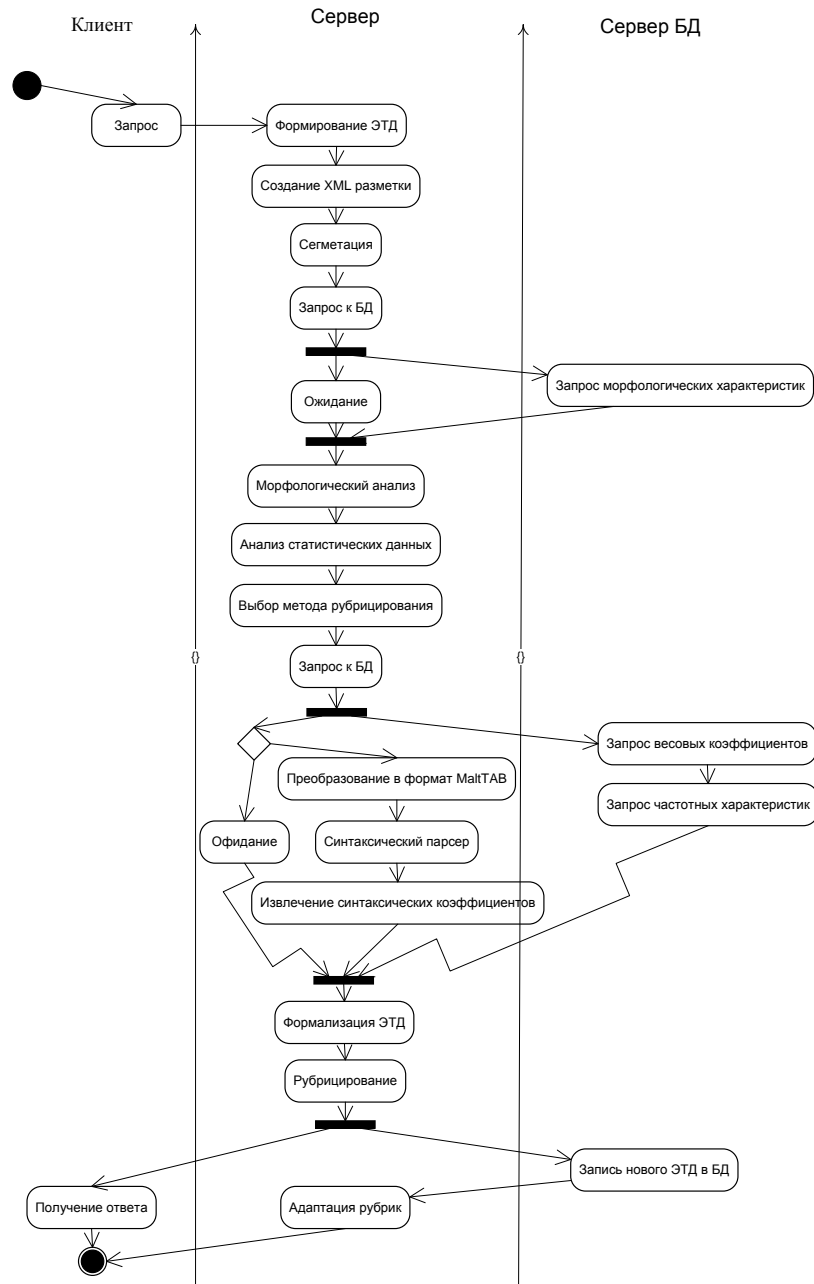


Рисунок 4.4 – Диаграмма активностей процедуры рубрицирования ИС Artex 1.0

Данная процедура активностей показывает последовательность выполнения этапов рубрицирования ЭНТД в разработанной системе. На диаграмме также отображены запросы, выполняемые сервером к базам данных морфологий, весовых коэффициентов, статистических характеристик ЗС и рубрицированных ЭНТД.

4.2 Оценка точности рубрицирования электронных неструктурированных текстовых документов с использованием разработанных алгоритмов и средств

Для проверки точности автоматизированного рубрицирования ЭНТД с использованием разработанных алгоритмов и средств проведена серия вычислительных экспериментов с использованием тестовых выборок из наборов данных Newsgroup-20 (пакета "19997", который содержит 18846 документов, отсортированных в пропорции 60% для обучающей выборки и 40% для тестирования). Среднее количество слов в тестовых документах – 161, среднее количество символов без пробелов – 871.

Используемый для тестирования разработанных алгоритмов пакет ЭНТД содержит сообщения по 20 разным рубрикам: mac-оборудование, pc-оборудование, windows-ос, windows-разное, автотехника, атеизм, бейсбол, ближний-восток, компьютерная-графика, космос, криптография, медицина, мотоциклы, политика-разное, политическое-оружие, продается, религия-разное, хоккей, христианство, электроника.

При исследовании ситуаций наличия взаимосвязанных рубрик использовалась матрица, представленная в таблице 4.1. Каждая ячейка таблицы содержит наиболее близкие по составу рубрики.

Таблица 4.1. Матрица взаимосвязей рубрик

Компьютерная-графика Windows-разное Pc-оборудование Mac-оборудование Windows-ос	Автотехника Мотоциклы Бейсбол Хоккей	Криптография Электроника Медицина Космос
Продается	Политика-разное Политическое-оружие Ближний-восток	Религия-разное

На рисунке 4.5 представлено нечеткое дерево решений для указанных рубрик в таблице 4.1.

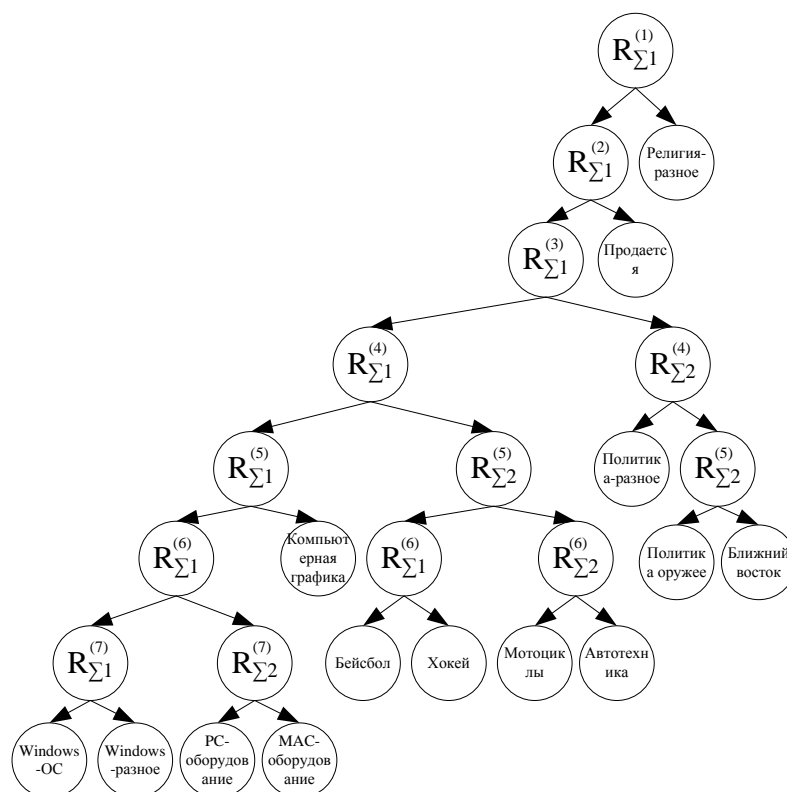


Рисунок 4.5 – Нечеткое дерево решений для выбранных рубрик

Для настройки алгоритма, реализующего модель рубрицирования с использованием весовых коэффициентов, были выбраны следующие начальные настройки:

- уникальным словам соответствует весовой коэффициент равный $r_{un}=1$; неуникальным – $r_{rr}=0.45$; общим – $r_{cm}=0,15$;
- порог отбора общих слов соответствует $\beta=80\%$;

В Приложении 1 на рисунке П.1 представлена экранная форма функционирования процедуры работы морфологического анализа ЭНТД, которая позволяет определить морфологические характеристики ЗС ЭНТД.

На рисунке П.2 представлена экранная форма работы сборщика статистических характеристик ЗСЮ, которая необходима для работы модели рубрици-

Экранная форма программной реализации ручной настройки весовых коэффициентов представлена на рисунке П.4.

Был проведен ряд вычислительных экспериментов, результаты которых представлены в таблицах 4.2 и 4.3. Таблица 4.2 содержит результаты рубрицирования при использовании взаимосвязанных рубрик, таблица 4.3 – невзаимосвязанных.

Таблица 4.2 – Результаты рубрицирования для взаимосвязанных рубрик с использованием разработанной ИС Artex 1.0, % правильно рубрицированных ЭНТД

<div>Размер обучающей выборки</div> <div>Модель</div>	На основе вероятностного классификатора	На основе весовых коэффициентов	На основе нейронечеткого классификатора	На основе НДР
2000	62	65	66	69
5000	73	71	74	76
8000	84	73	81	82
12000	87	76	84	85

Как видно из таблицы 4.2, в условиях взаимосвязанных рубрик в ситуациях с малым размером обучающей выборки (до 5000) модель рубрицирования на основе нечеткого дерева решений показывает более высокую точность рубрицирования ЭНТД по сравнению с остальными.

Таблица 4.3 – Результаты рубрицирования для невзаимосвязанных рубрик с использованием разработанной ИС Artex 1.0, % правильно рубрицированных ЭНТД

<div> <div>Размер обучающей выборки</div> <div> <div>Модель</div> </div> </div>	На основе вероятностного классификатора	На основе весовых коэффициентов	На основе нейронечеткого классификатора	На основе нечеткого дерева решений
2000	65	67	73	71
5000	82	71	80	78
8000	88	73	86	84
12000	91	76	89	88

На рисунках 4.5 и 4.6 представлены графики зависимостей точности алгоритмов рубрицирования от объема обучающей выборки при использовании моделей рубрицирования, перечисленных в таблицах 4.2 и 4.3, в условиях взаимосвязанных и невязанных рубрик соответственно.

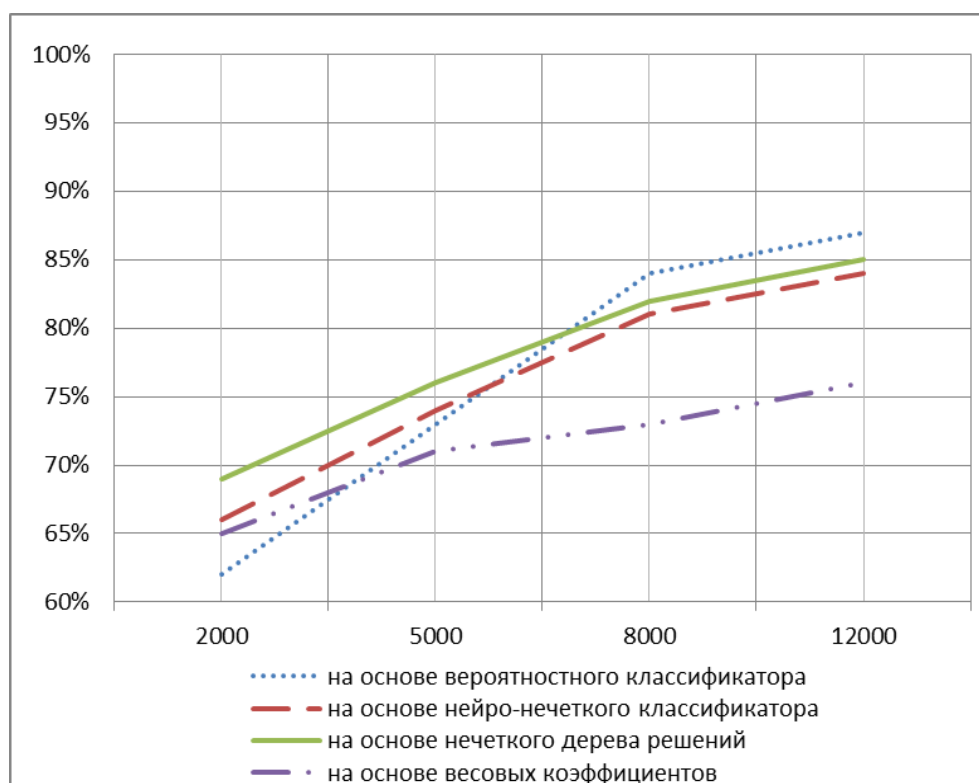


Рисунок 4.6 – График зависимости точности рубрицирования от объема обучающей выборки при использовании моделей рубрицирования, перечисленных в таблицах 4.2 и 4.3, в условиях взаимосвязанных рубрик

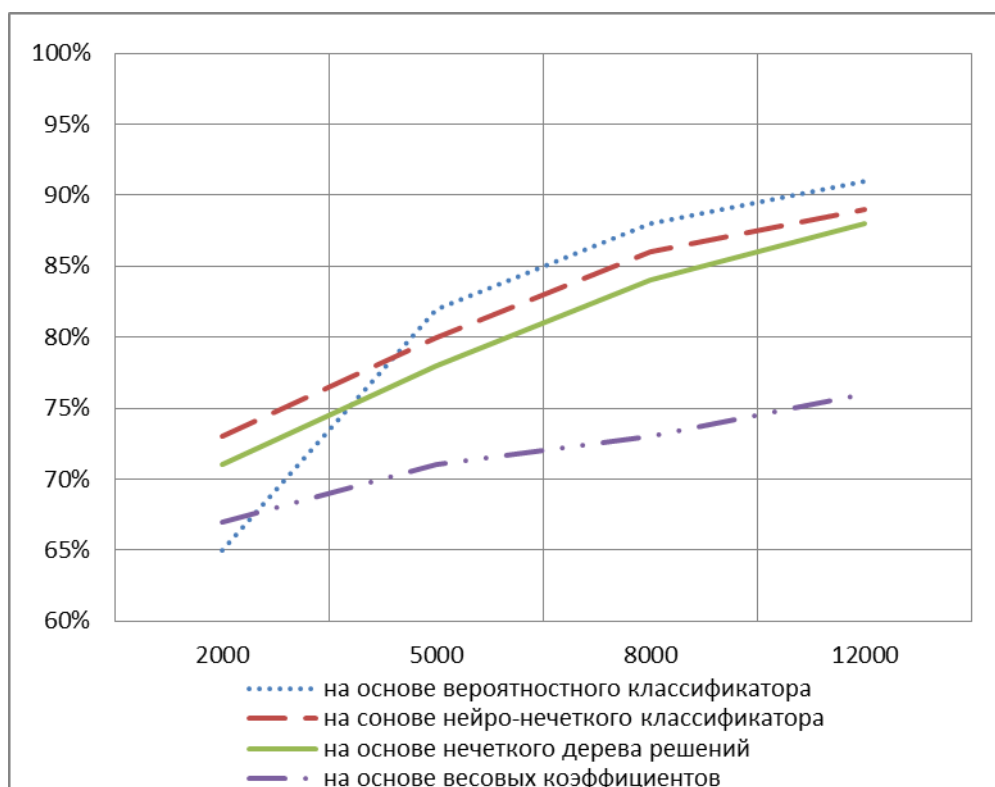


Рисунок 4.7 – График зависимости точности рубрицирования от объема обучающей выборки при использовании моделей рубрицирования, перечисленных в таблицах 4.2 и 4.3, в условиях невязанной рубрик

Как видно из таблицы 4.3, в условиях невязанной рубрик в ситуациях с малым размером обучающей выборки (до 5000) модель рубрицирования на основе нейро-нечеткого классификатора показывает более высокую точность рубрицирования ЭНТД по сравнению с остальными.

4.3 Результаты практического использования разработанных алгоритмов рубрицирования неструктурированных электронных текстовых документов в Администрации Смоленской области

Разработанные методы и модели для рубрицирования ЭНТД были практически использованы в ИС Администрации Смоленской области при автоматизации процедур обработки обращений (жалоб, заявлений и предложений) граждан и организаций.

В настоящее время в Администрации Смоленской области (далее- Администрация) для подачи обращений в электронном виде используются интернет-портал (<http://www.smoladmin.ru/gostyam-i-zhitelyam/obrascheniya-grazhdan/internet-priemnaya/vopros/>) и электронная почта smol@smoladmin.ru).

Общая процедура обработки обращений и подготовки ответа, которая регламентируется Федеральным законом от 02.05.2006 № 59-ФЗ «О порядке рассмотрения обращений граждан Российской Федерации» [120] и административными регламентами [121], состоит из следующих этапов:

Этап 1. Регистрация документа (заведение на документ регистрационной карточки).

Этап 2. Передача (доклад) документов руководителю.

Этап 3. Рассмотрение документов руководителем (резолюция).

Этап 4. Внесение сведений из резолюции (фамилии исполнителей, сроки исполнения) в регистрационную карточку.

Этап 5. Передача документов исполнителю.

Этап 6. Контроль исполнения документов.

Этап 7. Исполнение документа.

Этап 8. Списание документа в дело.

Указанная процедура предполагает не более чем в 3-дневный срок регистрацию ЭНТД (управлением по работе с обращениями граждан Аппарата Администрации Смоленской области; рубрицирование документа с целью определения департамента или организации для подготовки ответа (не более чем за 30 дней с момента регистрации ЭНТД); отправка ответа автору ЭНТД;

Для контроля процедур обработки указанных ЭНТД используется система документооборота ДелоPro, которая решает задачи: документирования; управления документооборотом; организации архивирования ЭНТД с возможностью быстрого поиска и извлечения.

Функциональные возможности указанной системы позволяют при участии сотрудников практически на каждом этапе движения обращения, посту-

пившего в Администрацию, осуществлять ведение регистрационной карточки по данному ЭНТД [122, 123].

На рисунке 4.8 показана схема обработки обращений в Администрации с использованием системы ДелоPro.

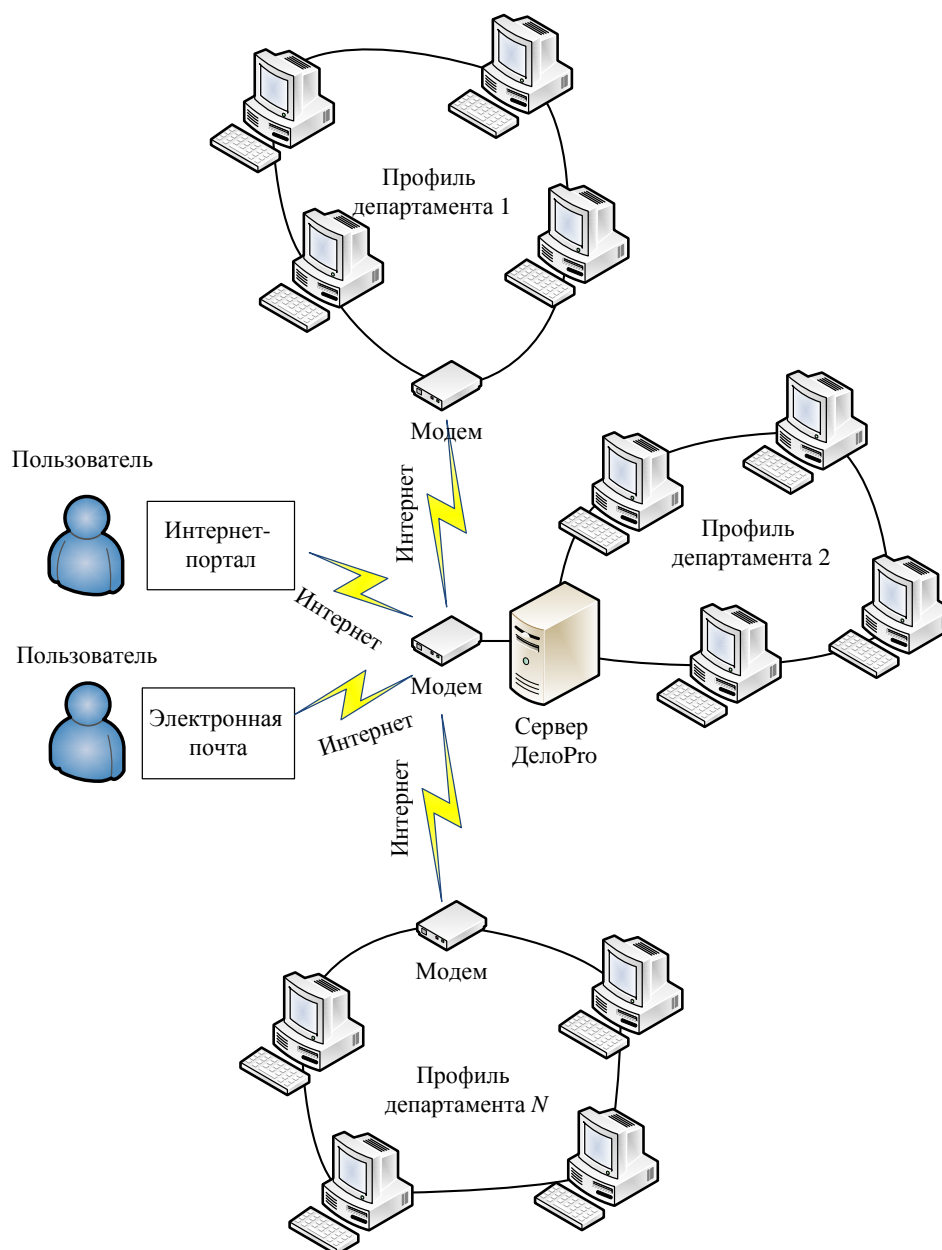


Рисунок 4.8 – Схема обработки обращений в Администрации с использованием системы ДелоPro

Учитывая существенный рост числа обращений в электронной форме, для выполнения временных условий их ручной обработки возникает необходимость увеличения штата сотрудников и, следовательно, фонда оплаты труда.

При этом значительные ресурсы расходуются именно на реализацию этапа рубрицирования ЭНТД, так как он предполагает детальное изучение ЭНТД. В связи с этим представляется целесообразным использование разработанных алгоритмов для автоматизации рубрицирования ЭНТД рассматриваемого типа, в том числе на основе интеграции предлагаемой информационной системы Artex 1.0 и системы ДелоPro (рисунок 4.9).

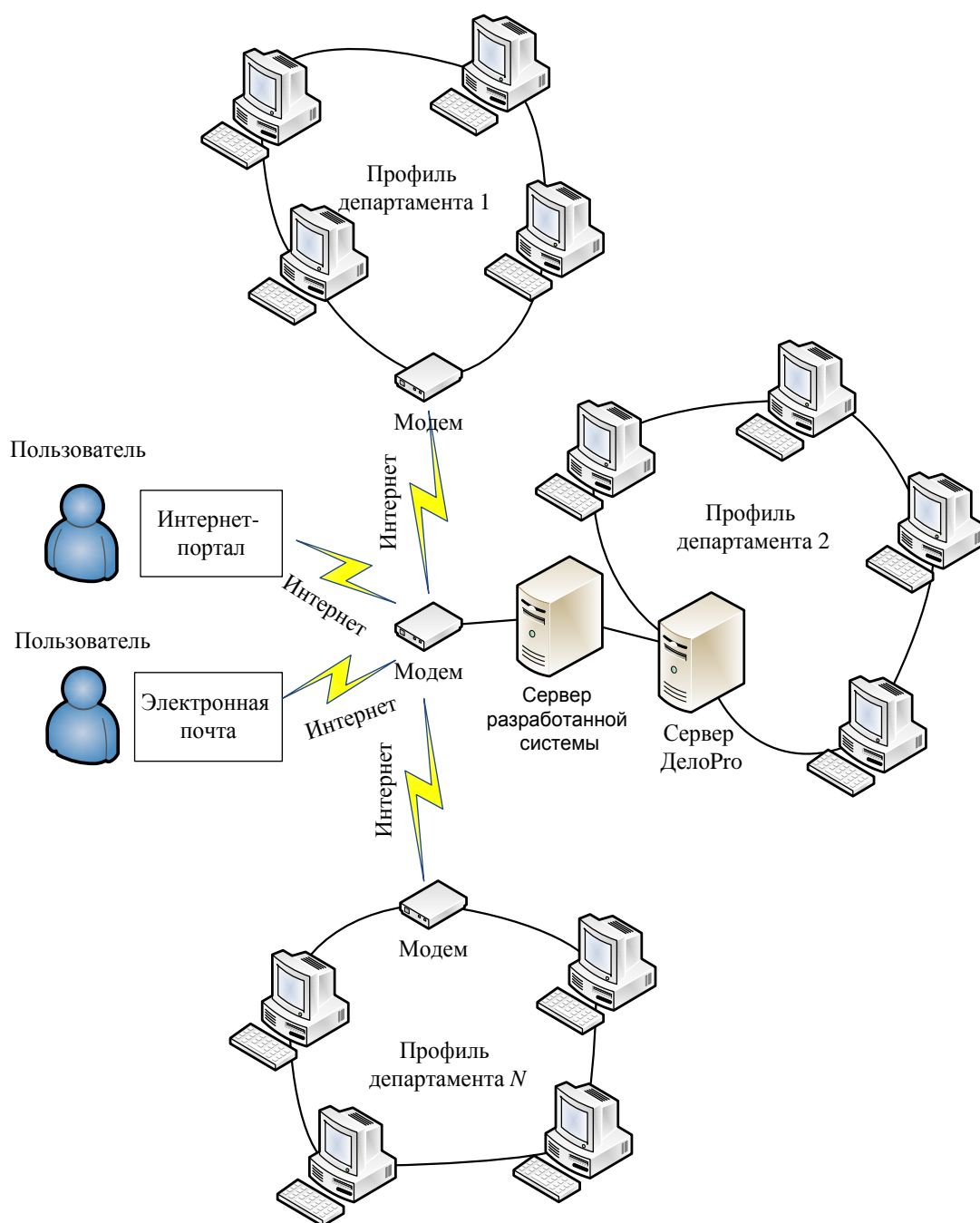


Рисунок 4.9 – Схема интеграции системы Artex 1.0 и системы ДелоPro

Процесс обработки обращений в Администрации состоит из 11 этапов и представлен на рисунке 4.10.

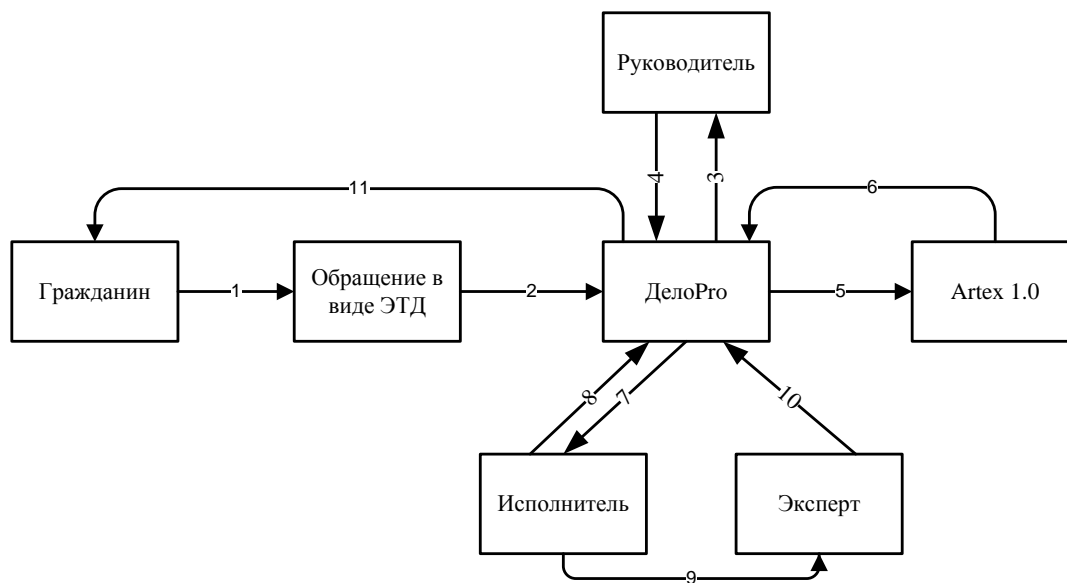


Рисунок 4.10 – Процесс обработки обращений в Администрации

Этап 1. Поступление обращения в электронном виде на сайт Администрации или отправка по электронной почте.

Этап 2. Обращение в виде ЭНТД поступает в систему ДелоPro, где для него заводится карточка.

Этап 3. Обращение направляется руководителю для контроля.

Этап 4. Руководитель может сделать некоторые корректировки в карточке.

Этап 5. Обращение поступает в ИС Artex 1.0 для рубрицирования.

Этап 6. Artex 1.0 осуществляет корректировки в карточке с пометкой рубрики.

Этап 7. Карточка поступает исполнителю.

Этап 8. Карточка с пометкой о выполнении поступает обратно в ДелоPro.

Этап 9. В случае неправильного рубрицирования обращение передается эксперту для ручного анализа.

Этап 10. Эксперт делает изменения в карточке и отправляет в ДелоPro для повторного назначения исполнителя.

Этап 11. Ответ на обращение направляется заявителю.

Для проверки точности рубрицирования ЭНТД при помощи информации

онной системы Artex 1.0 были проанализированы поступившие в 2016–2017 гг. 5062 жалобы и предложения, присланные в Администрацию Смоленской области через интернет-портал и по электронной почте. Анализ показал наличие 17 различных взаимосвязанных рубрик: общие вопросы общества и политики (R_1), разграничение полномочий и функций в Администрации (R_2), социальная сфера (R_3), образование (R_4), предложения по улучшению города Смоленска к 1150-летию (R_5), семья (R_6), культура (R_7), физическая культура и спорт (R_8), жилищно-коммунальная сфера (R_9), содержание и обеспечение коммунальными услугами (R_{10}), жилищный фонд (R_{11}), нежилой фонд (R_{12}), обеспечение права на жилище (R_{13}), экономика (R_{14}), хозяйственная деятельность (R_{15}), природные ресурсы (R_{16}) и охрана окружающей среды (R_{17}).

На рисунке 4.11 представлено нечеткое дерево решений для анализируемых рубрик.

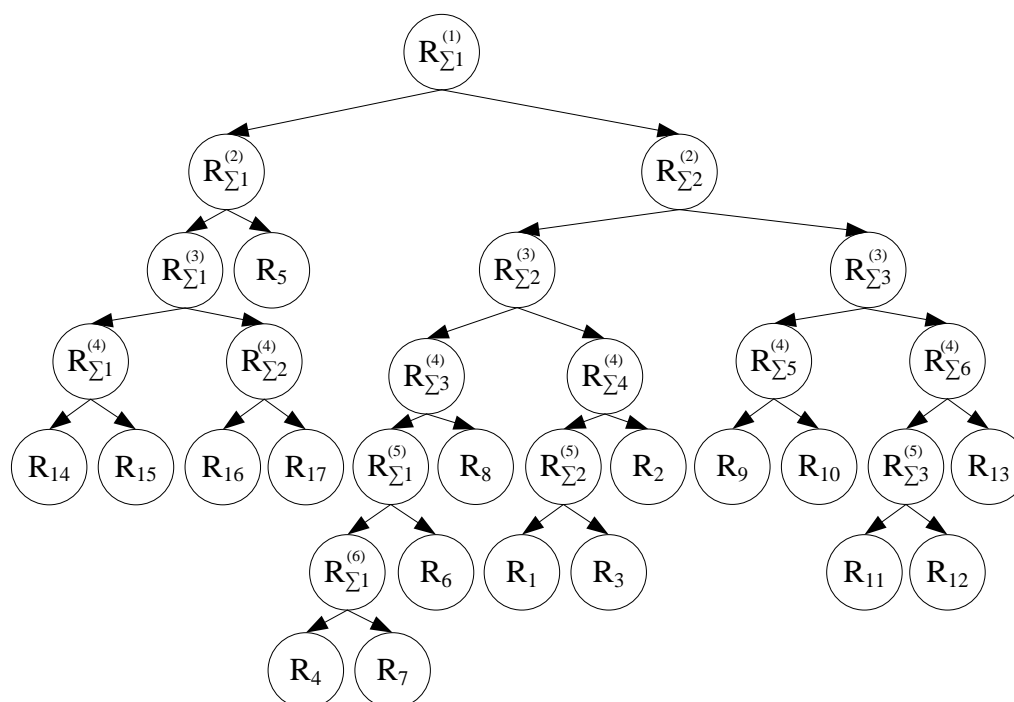


Рисунок 4.11 – Нечеткое дерево решений для анализируемых рубрик

Из рисунка 4.11 видно, что рубрики взаимосвязаны, и можно выделить несколько групп.

Анализ данных ЭНТД с помощью описанного способа анализа динамики изменения рубрик (слияния, разделения, появления новых и ликвидации) для электронных неструктурированных текстовых документов, отличающиеся использованием процедур динамической кластеризации этих документов с учетом синтаксических ролей слов, позволит выявить следующие изменения рубричного поля.

Изменение 1. Появление новых рубрик, условно названных «Автомобильные сигнализации» (R_{18}) и «Парковочные места» (R_{19}). Появление данных рубрик иллюстрируется наличием ЭНТД, позиционирующихся в начале координат на графиках для всех выделенных ранее рубрик (рисунок 4.12).

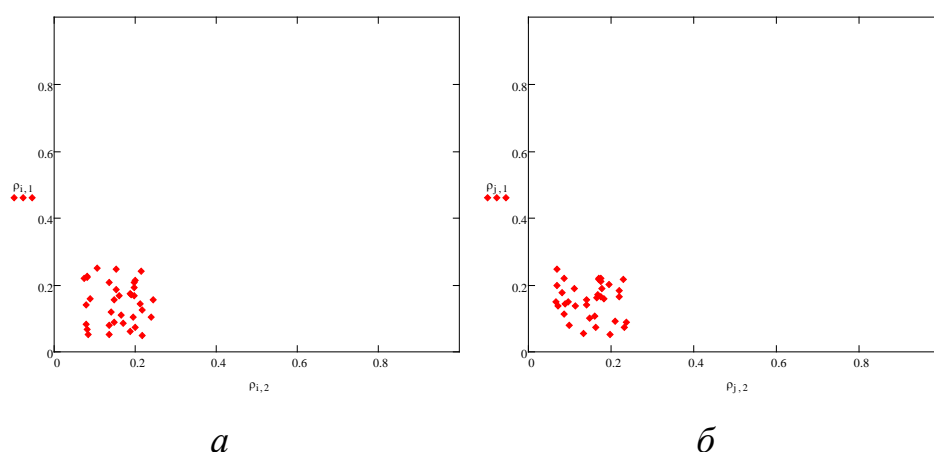


Рисунок 4.12 – Поле для рубрики «Общие вопросы общества и политики»

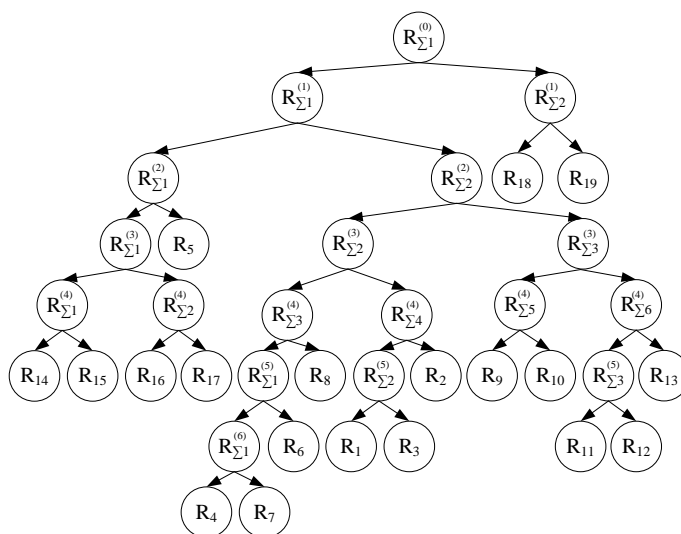


Рисунок 4.13 – Нечеткое дерево решений после появления новых рубрик

Из рисунка 4.13 видно, что наиболее близкими рубриками оказался корневой узел, т.к. данные две рубрики слабо связаны с остальными.

Изменение 2. Расщепление рубрики «Экономика», из которой выделилась еще одна рубрика «Информация и информатизация» (R_{20}). Данный вывод был сделан потому, что ЭНТД, предположительно относящиеся к рубрике «Экономика», попадали в середину рубричного поля, а на поля для всех остальных существующих на указанный момент времени рубрик – в окрестности начала координат (рисунок 4.14).

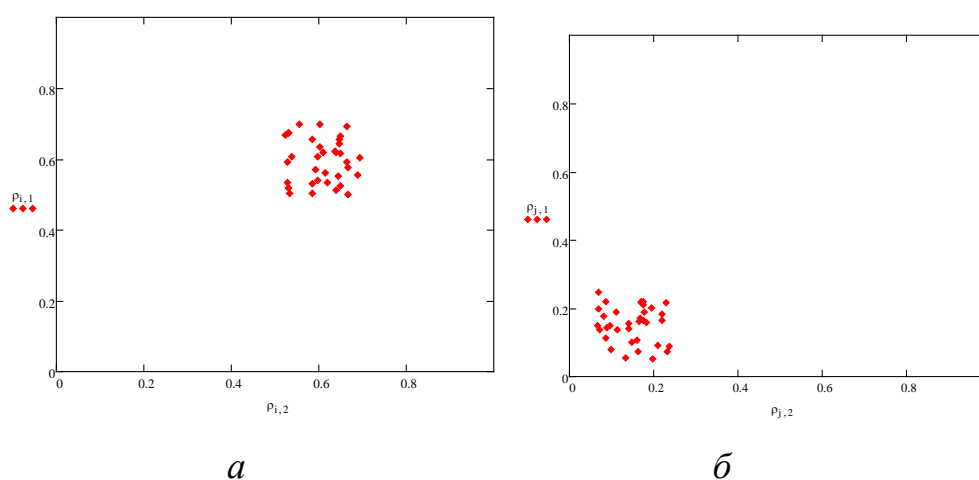


Рисунок 4.14 – Иллюстрация ситуации целесообразности разделения для рубрики «Экономика» и «Разграничение полномочий и функций Администрации»

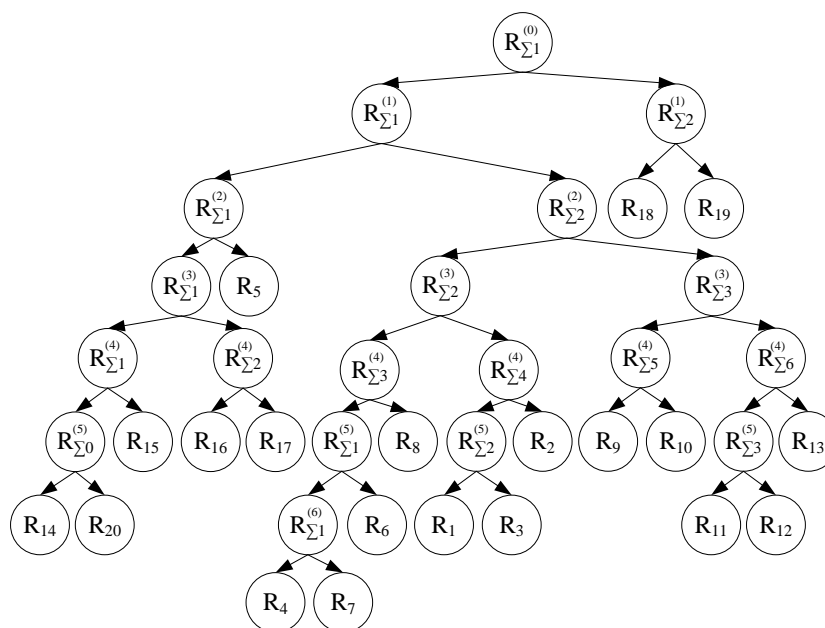


Рисунок 4.15 – Нечеткое дерево решений после расщепления рубрики

Из рисунка 4.15 видно, что лист, относящийся к рубрике «Экономика» стал узлом, из которого теперь выходят две рубрики: «Экономика» и «Информация и информатизация».

Изменение 3. Ликвидация рубрики «Предложения по улучшению города Смоленска к 1150-летию», так как поступающие ЭНТД перестали позиционироваться в правом верхнем углу графика для указанной рубрики.

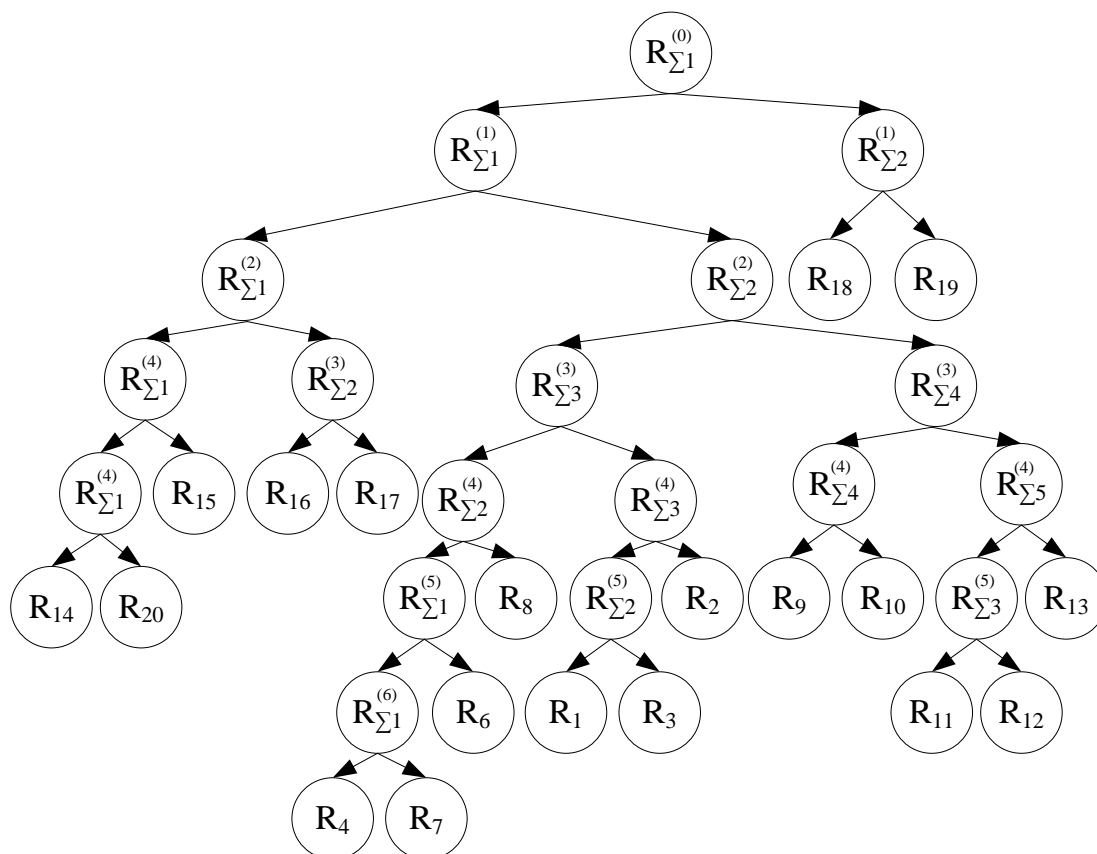


Рисунок 4.16 – Нечеткое дерево решений после ликвидации рубрики

Из рисунка 4.16 видно, что лист R_5 ликвидировали, а соседние узлы перенесли на уровень выше, чтобы не нарушать структуру дерева.

Изменение 4. Слияние рубрик «Природные ресурсы» и «Окружающая среда»: ЭНТД, поступающие в первую из указанных рубрику также попадали в окрестности точки (1; 1) для рубрики «Окружающая среда» (рисунок 4.17).

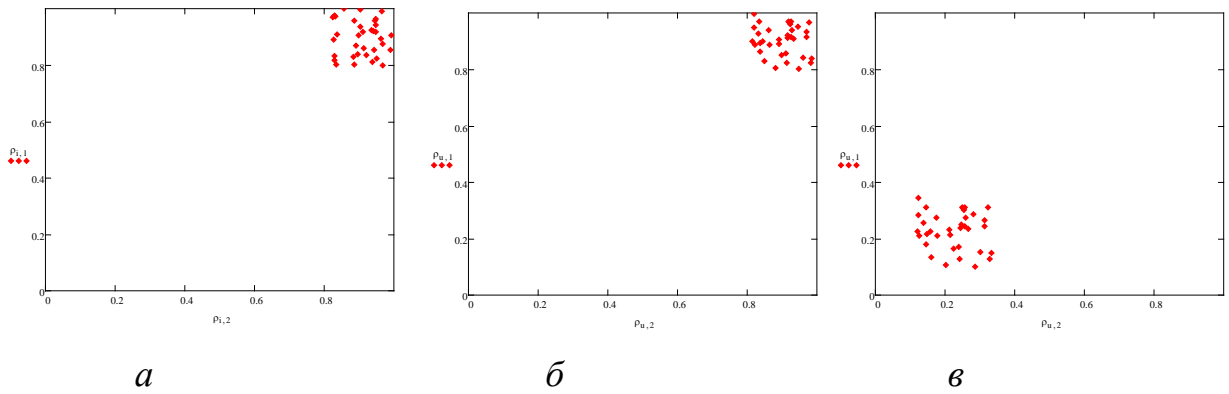


Рисунок 4.17 – Рубричное поле для рубрик «Природные ресурсы», «Окружающая среда» и «Семья»

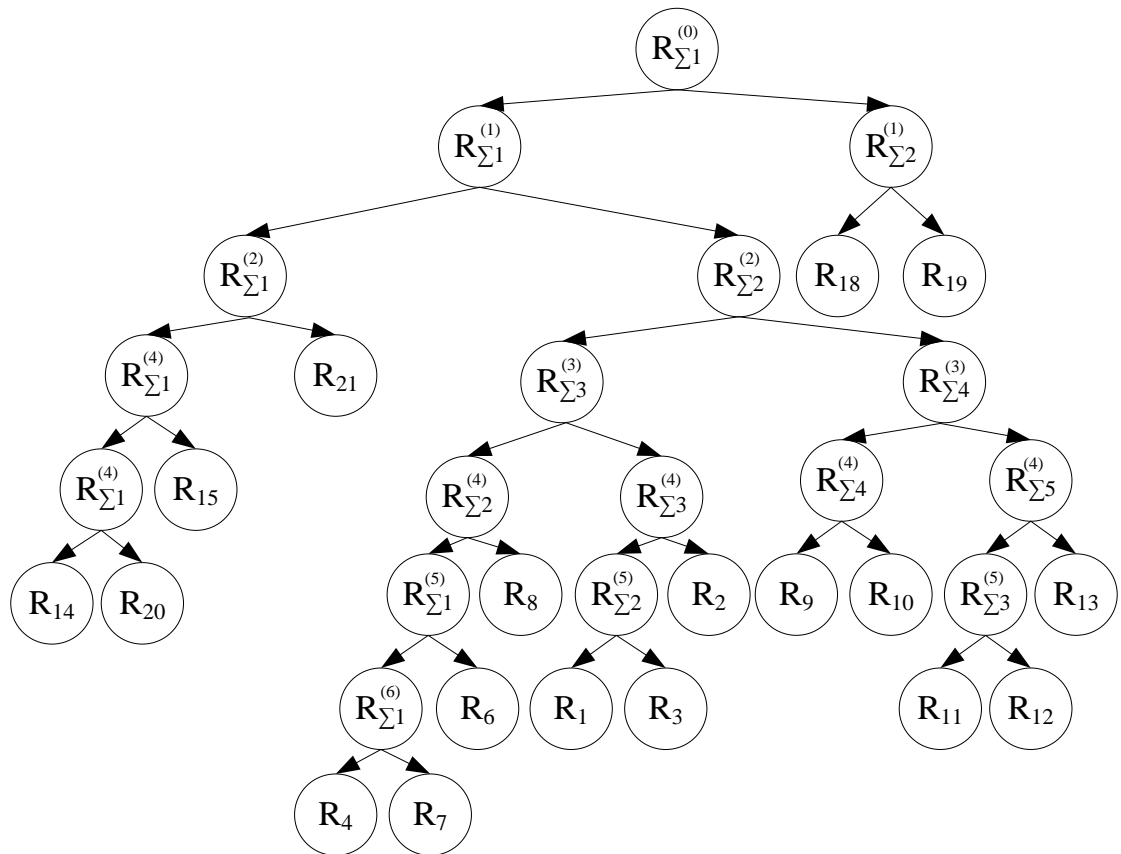


Рисунок 4.18 – Нечеткое дерево решений после слияния рубрики

R21 – это новая объединенная рубрика «Природные ресурсы и окружающая среда».

С использованием описанного выше набора поступивших в Администрацию ЭНТД было проведено тестирование разработанных моделей рубрицирования и ИС Artex 1.0, результаты тестирования представлены в таблице 4.4. Весь набор ЭНТД разделили по четырем ситуациям, описанным в подразделе 2.1. Далее ЭНТД по каждой ситуации разделили на обучающую выборку и тестовую в соотношении 60% и 40% соответственно. При настройке модели рубрицирования с использованием весовых коэффициентов применялись те же значения показателей, как и в примере в подразделе 4.1.

Таблица 4.4 – Результаты рубрицирования сообщений, поступивших в Администрацию Смоленской области, % правильно рубрицированных ЭНТД

Мо- дель Ситуация	На основе вероятностного классификатора	На основе весовых коэффициентов	На основе нейро-нечеткого классификатора	На основе нечеткого дерева решений	Artex 1.0
2	65	75	62	61	75
3	62	65	66	79	79
4	69	73	87	75	87
5	89	81	86	84	89

Исходя из анализа данных таблицы 4.4 можно заключить, что разработанная информационная система Artex 1.0 позволяет снизить число ошибочно рубрицированных ЭНТД в среднем на 13,3% по сравнению с известными системами, основанными на использовании вероятностных моделей.

Разработанная ИС Artex 1.0 также практически используется в учебном процессе филиала НИУ «МЭИ» в г. Смоленске при проведении лабораторных занятий по дисциплинам «Интеллектуальные информационные системы» и «Автоматизированные системы документооборота».

4.4 Выводы по главе

Разработана структура программного обеспечения информационной системы автоматизированного анализа электронных неструктурированных текстовых документов, позволяющая реализовать алгоритм предварительного анализа и алгоритмы рубрицирования текстовых документов, описанные в главе 3.

Проведена оценка точности рубрицирования электронных неструктурированных текстовых документов с использованием разработанных алгоритмов и программных средств на наборе данных Newsgroup-20, которая показала, что разработанные алгоритмы и программные средства обеспечивают более точное рубрицирование этих документов, чем известные алгоритмы, в условиях нехватки статистических данных и взаимосвязанности рубрик, а в условиях достаточного количества статистических данных показывают результаты не хуже, чем известные алгоритмы.

В таблице 4.4 приведены результаты использования алгоритмов анализа электронных неструктурированных текстовых документов на наборе из 1062 жалоб и предложений, поступивших в Администрацию города Смоленска через интернет-портал. Результаты показали, что разработанные методы, алгоритмы и программные средства обеспечивают повышение точности рубрицирования жалоб и предложений граждан в среднем на 13,3% по сравнению с моделью рубрицирования на основе вероятностного классификатора.

ЗАКЛЮЧЕНИЕ

В результате исследований решена научная задача, заключающаяся в разработке нейро-нечетких методов и алгоритмов анализа электронных неструктурированных текстовых документов в условиях изменения рубрик. При выполнении диссертации получены следующие основные результаты.

1. Выполнен анализ задач и методов автоматизированного рубрицирования текстовых документов и оценены их перспективы для анализа электронных неструктурированных текстовых документов с учетом особенностей жалоб и предложений граждан, поступающих в органы государственного и муниципального управления.

2. Разработан мультимодельный метод анализа электронных неструктурированных текстовых документов, обеспечивающий возможности комбинирования нечетких, нейро-нечетких и вероятностных моделей с учетом различного объема документов, степени пересечения рубрик и достаточности статистической информации о рубрицируемых документах.

3. Разработан метод мониторинга и изменения рубрик электронных неструктурированных текстовых документов в зависимости от идентифицированных ситуаций изменения рубричного поля на основе нечеткой динамической кластеризации этих документов.

4. Разработана каскадная нейро-нечеткая модель и алгоритмы анализа коротких электронных неструктурированных текстовых документов в условиях нехватки статистических данных для использования вероятностных методов.

5. Разработана нечетко-логическая модель и алгоритмы анализа электронных неструктурированных текстовых документов на основе нечетких деревьев решений с учетом синтаксических связей и ролей слов в предложениях в условиях взаимосвязанных рубрик и нехватки статистических данных.

6. Разработан комплекс алгоритмов, реализующих предлагаемый мультимодельный метод анализа электронных неструктурированных текстовых документов, а также метод мониторинга и изменения рубрик.

7. Проведена серия вычислительных экспериментов по проверке точности рубрицирования электронных неструктурированных текстовых документов с использованием разработанных методов, моделей, алгоритмов и программных средств, результаты которых позволили выделить области их применимости.

8. Представлены результаты практического использования разработанных алгоритмов и программных средств для автоматизированного анализа электронных неструктурированных текстовых документов, поступивших в Администрацию Смоленской области. Результаты показали, что разработанные методы, алгоритмы и программные средства обеспечивают повышение точности рубрицирования электронных сообщений, а так же оперативности подготовки ответа.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Об утверждении программы "Цифровая экономика Российской Федерации" [Электронный ресурс]. – Режим доступа: <http://government.ru/docs/28653/>. – Заглавие экрана. – (Документы – Правительство России).
2. Уэно Х., Кояма Т., Окамото Т. и др. Представление и использование знаний: Пер. с япон. – М.: Мир, 1989.
3. Рыжиков Ю.И. Вычислительные методы. СПб.: БХВ-Петербург, 2004.
4. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. – М.: МИЭМ, 2011. – 272 с.
5. Сегалович И., Маслов М. Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре форм. Казань: ООО «Хэтер», 1998.
6. Рубашкин В.Ш. Представление и анализ смысла в интеллектуальных информационных системах. – М.: Наука, 1989.
7. Бойцов Л. М. Синтез системы автоматической коррекции, индексации и поиска текстовой информации: дис. ... канд. технических наук: 05.13.01. – М., 2003. – 144 с.
8. Головкин Н. В. Формально-семантический анализ многозначной лексики как средство оптимизации систем автоматизированной обработки текстов: дис. ... канд. филологических наук: 10.02.19. – Ставрополь, 2011. – 194 с.
9. Мокроусов М.Н. Разработка и исследование методов и системы семантического анализа естественно-языковых текстов: дис. ... канд. технических наук: 05.13.01. – Ижевск, 2010. – 185 с.
10. Заболеева-Зотова А. В., Петровский А., Орлова Ю. А., Шитова Т. А. Автоматизированный анализ тематики текстов новостей // International Journal "Information Content and Processing", 2016. – V. 3. – N. 3.

11. Александров М. Ю. Методы автоматической классификации и статистического анализа входного потока текстовой информации в информационных системах: дис. ... канд. технических наук: 05.25.05. – М., 2008. – 203 с.
12. Николаева И. В. Автоматизация анализа массивов текстовых документов в информационно-коммуникационных средах: дис. ... канд. филологических наук: 10.02.21. – М., 2007. – 245 с.
13. Заболеева-Зотова А. В, Орлова Ю. А., Розалиев В. Л. Комплексный семантический анализ потока новостных текстов // Искусственный интеллект и принятие решений, 2015. – № 4. – С. 81-88.
14. От автоматической обработки текста к машинному пониманию [Электронный ресурс]. – Режим доступа: http://polit.ru/article/2013/03/26/vladimir_selegey/. – Заглавие с экрана. – (От автоматической обработки текста к машинному пониманию – ПОЛИТ.РУ).
15. Советов Б.Я., Цехановский В.В., Чертовский В.Д. Базы данных. Теория и практика (2 изд). – М.: Высшая школа, 2005.
16. Харрингтон Джен Л. Проектирование реляционных баз данных. – М.: Лори, 2006.
17. Cao Jian-fang, Wang Hong-bin Text categorization algorithms representations based on inductive learning // Information Management and Engineering (ICIME). 2010.
18. Андреев А.М., Березкин Д.В., Сюзев В.В., Шабанов В.И. Модели и методы автоматической классификации текстовых документов [Текст]. // Вестн. МГТУ. Сер. Приборостроение. – М.: Изд-во МГТУ, 2003. – №3.
19. Раскин Д. Интерфейс. Новые направления в проектировании компьютерных систем. – СПб.: Символ-Плюс, 2005.
20. Леонтьева Н.Н. Автоматическое понимание текстов: системы, модели, ресурсы: учеб. пособие для студ. лингв. фак. вузов / Н. Н. Леонтьева. – М.: Издательский центр "Академия", 2006. – 304 с.
21. Хапаева Т. Автоматическая классификация документов [Текст]. // Софтерра, 2002. – №2.

22. Шайкевич А.Я. Дистрибутивно-статистический анализ в семантике. Принципы и методы семантических исследований. – М.: Наука, 1976.
23. Chi Wang , Marina Danilevsky , Nihit Desai , Yinan Zhang , Phuong Nguyen, Thrivikrama Taula, Jiawei Han. A phrase mining framework for recursive construction of a topical hierarchy // In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013.
24. Stein B. Meyer zu Eissen S. Document Categorization with Major Clust [Text]. // Proceedings of the 12th, 2002.
25. Андреев А.М. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа [Электронный ресурс] / А.М. Андреев, Д.В. Березкин, В.В. Морозов, К.В. Симаков. – Электрон. текст. дан. – Режим доступа: http://www.inteltec.ru/publish/articles/textan/57_simakov.shtml, свободный.
26. Бауман Е.В., Дорофеюк А.А. Классификационный анализ данных [Текст]. // Избранные труды Международной конференции по проблемам управления. – Т.1. – М.: СИНТЕГ, 1999. – С. 62-67.
27. Manning C., Raghavan P., Schutze H. Introduction to Information Retrieval [Text]. // Cambridge University Press, 2008. – P. 544.
28. Mitchell T. M. Machine Learning [Text]. // McGraw Hill, New York, 1997. – P. 414.
29. Joachims T. Learning to Classify Text using Support Vector Machines, Kluwer/Springer, 2002.
30. Joachims T. Making Large-scale support vector machines learning practical // Advances in Kernel Methods: Support Vector Machines [Text]. / B.Scholkopf. C.Burges, A.Smola (eds.) MIT Press: Cambridge, MA – 1998.
31. Quinlan J. Induction of decision trees [Text]. // Machine Learning, 1998. – V. 1, – N. 1. – P. 81-106.
32. Мухаметзянов И.З., Мешалкин В.П. Имитационная многоагентная нечетко-логическая модель принятия маркетинговых решений промышленного

предприятия в условиях неопределенности // Прикладная информатика. 2014. № 3 (51). С. 100-109.

33. Lewis D.D. An evaluation of phrasal and clustered representations on a text categorization task [Text] // Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval. ACM Press, US. 1992. – P. 37-50.

34. Sebastiani F. Machine Learning in Automated Text Categorization [Text]. // ACM Computing Surveys, 2002. – V. 34, – N. 1. – P. 1-47.

35. Симаков К. В. Модели и методы извлечения знаний из текстов на естественном языке: дис. ... канд. технических наук: 05.13.17. – М., 2008. – 267 с.

36. Soloshenko A.N., Orlova Y.A., Rozaliev V.L., Zaboleeva-Zotova A.V. Establishing semantic similarity of the cluster documents and extracting key entities in the problem of the semantic analysis of news texts // Modern Applied Science, 2015. – V. 9. – N. 5. – P. 246-268.

37. Adam Berger. Statistical Machine Learning for Information Retrieval [Text]. // Carnegie Mellon University, 2001. – P. 143.

38. Witten I. H., Frank E. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition) [Text]. // Morgan Kaufmann, 2005. – P. 525.

39. Протасов К.В. Статистический анализ экспериментальных данных. – М.: Мир, 2005.

40. Толчеев В. О. Систематизация, разработка методов и коллективов решающих правил классификации библиографических текстовых документов: дис. ... док. технических наук: 05.13.01 / Владимир Олегович Толчеев. – М., 2009.

41. Фальк В.Н., Бочаров И.А., Шаграев А.Г. Трансдуктивное обучение логистической регрессии в задаче классификации текстов // Программные продукты и системы, 2014. – № 2. – 20 с.

42. Шаграев А. Г., Фальк В. Н. Линейные классификаторы в задаче классификации текстов // Вестн. МЭИ, 2013. – № 4. – 204-209 с.

43. Гулин В. В. Исследование и разработка методов и программных средств классификации текстовых документов: дис. ... канд. технических наук: 05.13.11. – М., 2013. – 172 с.
44. Шабанов В. И. Модели и методы автоматической классификации текстовых документов: дис. ... канд. технических наук: 05.13.11. – М., 2003. – 225 с.
45. Епрев А. С. Исследование влияния разрешения лексической многозначности с помощью контекстных векторов на эффективность категоризации текстовых документов: дис. ... канд. физико-математических наук: 05.13.11. – Омск, 2011. – 118 с.
46. Круглов В.В., Борисов В.В. Искусственные нейронные сети. Теория и практика [Текст]. – М.: Горячая линия – Телеком, 2001.
47. Quinlan J. C4.5: Programs for Machine Learning [Text]. // Morgan Kaufmann, 1993. – P. 302.
48. Rocchio J.J. Relevance feedback in information retrieval [Text]. // The SMART Retrieval System: Experiments in Automatic Document Processing, 1971. – P. 313-323.
49. Sebastiani F. Text Categorization [Text]. // Text Mining and Its Applications. WIT Press, Southampton, UK, 2005. – P. 109-129.
50. Yang Y. Pedersen J.O. A comparative study on feature selection in text categorization [Text]. // Proceedings of ICML-97, 14th International Conference on Machine Learning. Morgan Kaufmann Publishers, San Francisco, US: Nashville, US. 1997. – P. 412-420.
51. Yang Y., Chute C. G. An example-based mapping method for text categorization and retrieval [Text]. // ACM Trans. Inform. Syst, 1994. – V. 12, – N. 3. – P. 252-277.
52. Yang Y. Expert network: Effective and Efficient learning from human decisions in text categorisation and retrieval [Text]. // Proceedings of SGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval, Dublin, Ireland, 1994. – P. 13-22.

53. Сидорова Е. А. Методы и программные средства для анализа документов на основе модели предметной области: дис. ... канд. физико-математических наук: 05.13.11. – Новосибирск, 2006. – 125 с.

54. Шелманов А. О. Исследование методов автоматического анализа текстов и разработка интегрированной системы семантико-синтаксического анализа: дис. ... канд. технических наук: 05.13.17. – М., 2015. – 210 с.

55. Тревгода С. А. Методы и алгоритмы автоматического реферирования текста на основе анализа функциональных отношений: дис. ... канд. технических наук: 05.13.01. – СПб., 2009. – 257 с.

56. Чугреев В. Л. Модель структурного представления текстовой информации и методы ее тематического анализа на основе частотно-контекстной классификации: дис. ... канд. технических наук: 05.13.01. – СПб., 2003. – 185 с.

57. Stein B. Niggemann O. On the Nature of Structure and its Identification [Text]. // P. Widmayer, G. Neyer, S. Eidenbenz (eds.). Graph-Theoretic Concepts in Computer Science. LNCS 1665. Springer-Verlag, 1999.

58. Yang Y., Liu X. A re-examination of text categorization methods [Text]. // Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, Berkeley, CA, 1999. – P. 42-49.

59. Текстомайнинг. Извлечение информации из неструктурированных текстов | КомпьютерПресс [Электронный ресурс]. – Режим доступа: <http://compress.ru/article.aspx?id=19605>. – Заглавие экрана. – (Текстомайнинг).

60. Чернявский А.Л., Бауман Е.В., Дорофеюк А.А. Методы динамического классификационного анализа данных [Текст]. // Искусственный интеллект, 2002. – № 2. – С. 290-297.

61. Singh. S. Dynamic Pattern Recognition for Temporal Data [Text]. // Proc. 5th European Congress on Intelligent Techniques and Soft Computing (EUFIT'97), Aachen, Germany, 1997. – Vol 3.

62. Alberto Muñoz Compound Key Word Generation from Document Databases Using A Hierarchical Clustering ART Model, 1997.

63. Ротштейн А.П., Познер М., Ракитянская А.Б. Нейро-нечеткая модель прогнозирования результатов спортивных игр [Текст]. // В сб. трудов 8-й Всероссийской конф. "Нейрокомпьютеры и их применение" НКП-2002. – М., 2002. – С. 251-263.
64. Кофман А. Введение в теорию нечетких множеств [Текст]. – М.: Радио и связь, 1982.
65. Круглов В.В., Дли М.И., Голунов Р.Ю. Нечеткая логика и искусственные нейронные сети [Текст]. – М.: Физматлит, 2001.
66. Нечеткие множества в моделях управления и искусственного интеллекта [Текст]. / Под ред. Д.А. Поспелова. – М.: Наука, 1986.
67. Блишун А.Ф., Знатнов Ю.С. Обоснование операций теории нечетких множеств [Текст]. // Сб. науч. тр. "Нетрадиционные модели и системы с нечеткими знаниями". – М.: Энергоиздат, 1991. – С. 21-23.
68. Харламов А.А. Ассоциативный процессор на основе нейроподобных элементов для структурной обработки информации [Текст]. // Информационные технологии, 1997. – № 8.
69. Bauman E.V., Dorofeyuk A.A. Types of fuzzines in clustering [Text] // Intelligent Techniques and Soft Computing. Verlag Mainz, Aachen, 1997.
70. Гаврилова Т.А., Червинская К.Р. Извлечение и структурирование знаний для экспертных систем [Текст]. – М.: Радио и связь, 1992.
71. Дж Солтон. Динамические библиотечно-поисковые системы [Текст]. – М.: Мир, 1979.
72. Мелихов А.Н., Бернштейн Л. С., Коровин С.Я. Ситуационные советующие системы с нечеткой логикой [Текст]. – М.: Наука, 1990.
73. Нечеткие множества и теория возможностей. Последние десятилетия [Текст] / Под ред. Р.Р. Ягера. – М.: Радио и связь, 1986.
74. Прикладные нечеткие системы [Текст] / Под ред. Т. Терано, К. Асаи, М. Сугэно. – М.: Мир, 1993.
75. Девятков В.В. Системы искусственного интеллекта [Текст]. – М.: Изд-во МГТУ им. Н. Э. Баумана, 2001.

76. Леоненков А.В. Нечеткое моделирование в среде MATLAB и fuzzyTECH [Текст]. // СПб.: БХВ-Петербург, 2003.
77. Ялов В.П. Методы и алгоритмы адаптивной нечеткой классификации сложных объектов: Автореф. дис. канд. техн. наук. М., 2002.
78. Круглов В.В., Дли М.И. Интеллектуальные информационные системы: компьютерная поддержка систем нечеткой логики и нечеткого вывода [Текст]. – М.: Физматлит, 2002.
79. Круглов В.В. Функциональное сходство систем нечеткого вывода и искусственных нейронных сетей [Текст]. // В сб. трудов 14-й междун. научн. конф. "Математические методы в технике и технологиях. ММТТ-14". Т. 2. Смоленск, 2001. – С. 163-165.
80. Дюк В., Самойленко А. Data mining (учебный курс). // СПб.: Питер, 2001.
81. Роб П., Коронел К. Системы баз данных. Проектирование, реализация и управление. – СПб.: БХВ-Петербург, 2004.
82. Daniel Ramage , Christopher D. Manning , Susan Dumais Partially labeled topic models for interpretable text mining (2011) // In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011.
83. Попов Э. В., Фоминых И. Б., Харин Н. П., Виньков М. М. Управление знаниями. Аналитический обзор [Электронный ресурс] // Вестник РФФИ, 2004. – № 4. Режим доступа. – <http://www.rfbr.ru/pics/20742ref/uprznan.pdf>.
84. Стассман Поль А. Информация в век электроники: (Проблемы управления): Пер. с англ. с сокр. / науч. ред. и авт. предисл. Б.З. Мильнер. – М.: Экономика, 1987.
85. Сулицкий В.Н. Методы статистического анализа в управлении. – М.: Дело, 2002.
86. Норенков И.П. Задачи управления знаниями, извлекаемыми из текстовых документов. // Наука и образование. 2011. №9.
87. Шмулевич М. М. Методы автоматической кластеризации текстов, основанный на извлечении из текстов имен объектов и последующем построении

графов совместной встречаемости ключевых термов: дис. ... канд. физико-математических наук: 05.13.17. – М., 2009. – 120 с.

88. Тейз А., Грибомон П., Юлен Г. и др. Логический подход к искусственному интеллекту. От модальной логики к логике баз данных: Пер. с франц. – М.: Мир, 1998.

89. Искусственный интеллект. – В 3-х кн. Кн.2. Модели и методы: Справочник. / Под ред. Д.А. Поспелова. – М.: Наука, 1990.

90. Тейз А., Грибомон П., Юлен Г. и др. Логический подход к искусственному интеллекту. От модальной логики к логике баз данных: Пер. с франц. – М.: Мир, 1998.

91. Шеменков П. С. Разработка и исследование модели нейросетевого метода анализа текстовых документов: дис. ... канд. технических наук: 05.13.18. – СПб., 2009. – 153 с.

92. Мешкова Е. В. Разработка и исследование гибридных нейросетевых моделей для автоматической классификации текстовых документов: дис. ... канд. технических наук: 05.13.01. – Таганрог, 2009. – 164 с.

93. Корж В. В. Методы кодирования текстовой информации для построения нейросетевых классификаторов документов: дис. ... канд. технических наук: 05.13.06. – М., 2000. – 161 с.

94. Ф. Уссермен Нейрокомпьютерная техника. – М.: Мир, 1992.

95. Росин М.Ф., Булыгин В.С. Статистическая динамика и теория эффективности систем управления. – М.: Машиностроение, 1981.

96. Круглов В.В. Адаптивные системы нечеткого логического вывода [Текст]. // Нейрокомпьютеры: разработка и применение, 2003. – № 5. – С. 13-16.

97. Гимаров В.А. Методы динамической классификации сложных объектов [Текст]. – М.: Физматлит, 2003 – 224 с.

98. Гимаров В.А., Дли М.И. Динамическая кластеризация состояния химико-технологического оборудования [Текст]. // Известия Вузов. Химия и химическая технология, 2004 – т. 47. – №8. – С. 143-147.

99. Гимаров В.А., Дли М.И., Гимаров В.В. Задача распознавания динамически изменяющихся образов: формулировка задачи и перспективы решения [Текст]. // Программные продукты и системы. Приложение к журналу Проблемы теории и практики управления, 2001. – №3. – С. 28-30.

100. Гимаров В.А. Применение нейросетей для распознавания динамических объектов [Текст]. // Актуальные вопросы управления техническими системами: Сб.тр. межвуз. сем., Смоленск: ВУ ОВПО МО РФ, 1999. – С. 40-42.

101. Гимаров В.А., Дли М.И., Битюцкий С.Я. Применение нейронных сетей для классификации динамических объектов [Текст]. // Математические методы в интеллектуальных информационных системах: Сб.тр. Межд. науч.кон. Смоленск, 2002. – С. 57-59.

102. Гимаров В.А., Дли М.И. Динамические задачи классификации объектов [Текст]. // Математические методы в технике и технологиях: Сб. трудов XVII Международ. науч. конф.: Т. 2. Кострома: изд-во Костр. гос. технол. ун-та, 2004. – С.33-36.

103. Федоров А.Г. Базы данных. – М.: КомпьютерПресс, 2001.

104. Поляков Д. В. Математические модели и алгоритмы эффективного поиска текстовой информации на основе кластеризации по нечетким коллокациям: дис. ... канд. технических наук: 05.13.17. – Тамбов, 2013. – 150 с.

105. Мешалкин В.П., Гимаров В.А., Дли М.И. Динамическая классификация сложных технологических систем: методы, алгоритмы и практические результаты. М.: Физматлит.2004. С. 344.

106. Ермаков А.Е., Плешко В.В. Синтаксический разбор в системах статистического анализа текста [Текст]. // Информационные технологии, 2002. – N. 7.

107. Bevainyte A., Butenas L. Document classification using weighted ontology // Materials Physics and Mechanics. 2010. №9.

108. Леонтьева Н.Н. К теории автоматического понимания естественных текстов. Ч.2: Семантические словари: состав, структура, методика создания – М.: Изд-во МГУ, 2001.

109. Леонтьева Н.Н. К теории автоматического понимания естественных текстов. Ч.3: Семантический компонент. Локальный семантический анализ. – М.: Изд-во МГУ, 2002.
110. Рубашкин В.Ш. Представление и анализ смысла в интеллектуальных информационных системах. – М.: Наука, 1989.
111. Lewis D.D. Representation and Learning in Information Retrieval [Text]. // PhD thesis, Computer Science Dept.; Univ. of Mass.; Amherst, M. A., 1992. – P. 91-93.
112. Орлов А.И. Экспертные оценки: учеб. пособие, М.: 2002. – 31 с.
113. М. В. Спика Создание Web-сайтов [Текст]. / А. В. Слепцов. – М.: Вильямс, 2007. – 288 с.
114. Национального корпуса русского языка [Электронный ресурс]. – Режим доступа: <http://www.ruscorpora.ru/index.html>. – Заглавие экрана. – (Национальный корпус русского языка).
115. Цыганов И.Г., Власов А.И., Ларютин А.А., Смирнов А.С., Иванов В.В., Петухов А.М., Кузнецов А.С. Исследование адаптивных систем категоризации электронных текстов. 2003. С. 62.
116. M. Marco Testing for concordance between several criteria [Text]. // Journal of Statistical Computation and Simulation, 2016. – P. 84.
117. Теслер Г.С. Интенсификация процесса вычислений // Математические машины и системы, 1999. – №2.
118. Теслер Г.С. Новая кибернетика. – Киев: Логос, 2004.
119. Торрес Р. Дж. Практическое руководство по проектированию и разработке пользовательского интерфейса. Пер. с англ. – СПб.: Вильямс, 2002.
120. Федеральный закон «О порядке рассмотрения обращений граждан Российской Федерации» [Электронный ресурс]. – Режим доступа: https://www.admin-smolensk.ru/obrascheniya_grazhdan/normativno-pravovie_akti/fz59/. – Заглавие экрана. – (Федеральный закон от 02.05.2006 N 59-ФЗ "О порядке рассмотрения обращений граждан Российской Федерации" – Администрация Смоленской области – Официальный портал органов власти).

121. Административный регламент обработки обращений граждан [Электронный ресурс]. – Режим доступа: <http://www.smoladmin.ru/gostyam-i-zhitelyam/municipalnye-uslugi/perechen-municipalnyh-uslug/>. – Заглавие экрана. – (Перечень муниципальных услуг. Муниципальные услуги. Гостям и жителям. Официальный сайт Администрации города-героя Смоленска).

122. Аналитическая справка о работе Аппарата Администрации Смоленской области с обращениями граждан [Электронный ресурс]. – Режим доступа: https://www.admin-smolensk.ru/obrascheniya_grazhdan/obzori_obrascheniy/news_16096.html. – Заглавие с экрана. – (Информационно-аналитический отчет о работе Администрации Смоленской области и Аппарата Администрации Смоленской области с обращениями граждан и оказанию гражданам бесплатной юридической помощи – Администрация Смоленской области – О).

123. Обзор обращений граждан Администрации города Санкт-Петербурга [Электронный ресурс]. – Режим доступа: <http://gov.spb.ru/gov/obrasheniya-grazhdan/otchet-obrasheniya/?page=1>. – Заглавие с экрана. – (Обзоры обращений – Администрация Санкт-Петербурга).

ГЛОСАРИЙ

1. **Теоретическая компьютерная лингвистика (КЛ-1)** – является дисциплиной, которая содержит весь перечень задач, область исследований со всеми вытекающими требованиями к степени формализации языковых описаний.

2. **Инженерная компьютерная лингвистика (КЛ-2)** – область, в которой исследуются методы обработки, изучения и решения некоторых полезных задач обработки ЕЯ. По объему привлекаемых данных и используемым методам выходит за пределы лингвистики, но существенно основывается на её моделях.

3. **Инструментальная компьютерная лингвистика (результат взаимодействия КЛ-1 и КЛ-2)** – лингвистика 21-ого века, которая радикально изменила методологию исследований. Благодаря компьютеру лингвистам стали доступны новые инструменты изучения (корпуса, парсеры, лингвистические ресурсы).

4. **Фонология** изучает звуки речи и правила их соединения при формировании речи.

5. **Морфология** занимается внутренней структурой и внешней формой слов речи, включая части речи и их категории.

6. **Синтаксис** изучает структуру предложений, правила сочетаемости и порядка следования слов в предложении, а также общие его свойства как единицы языка.

7. **Семантика и прагматика** тесно связанные области: семантика занимается смыслом слов, предложений и других единиц речи, а прагматика – особенностями выражения этого смысла в связи с конкретными целями общения.

8. **Лексикография** описывает лексикон конкретного естественного языка, а также его отдельные слова и их грамматические свойства и методы создания словарей.

9. **Лексический анализ текста** – выделение слов, знаков препинания, цифр и прочих текстовых единиц.

10. **Морфологический анализ** – определение грамматических характеристик лексем, а так же основных словоформ.

11. **Синтаксический анализ** – установление структуры предложения – системы связей между словами.

12. **Семантический анализ** – построение структуры, ассоциированной непосредственно с передаваемым значением – в границах языка.

13. **Прагматический анализ** – интерпретация семантической структуры в контексте модели текста и знаний о мире.

14. **Синтаксический анализ** описывает систему связей слов в предложении, называемую синтаксисом. В разных языках система синтаксических отношений, образующая синтаксическую структуру предложения, создается разными средствами – вспомогательными словами, грамматическими значениями, порядком слов, пунктуацией.

15. **Семантический анализ** – это переход от структуры поверхностных синтаксических связей к ее смысловой интерпретации, представленной глубинной семантической структурой. Это формализованное представление, соответствующее той глубине анализа, которая может быть примерно ассоциирована с информацией из толкового словаря языка.

16. **Прагматический анализ** – это интерпретация того, что получено в результате чисто языкового анализа уже в контексте ситуации или в рамках какой-то модели мира, которая стоит за текстом.

17. **Алгоритм классификации с учителем** – алгоритм категоризации, который использует обучающее множество, чтобы построить классификатор.

18. **Алгоритмы классификации без учителя** анализируют коллекцию полнотекстовых документов с целью разбиения их на группы так, чтобы внутри одной группы оказывались документы наиболее родственные в некотором смысле, а различные документы попадали в различные группы.

19. Вектор признаков VSM (VectorSpaceModel) текстового документа представляется в виде вектора, каждая координата которого соответствует частоте встречаемости одного из слов всей коллекции в этом тексте. Объединение всех таких векторов в единую таблицу приводит нас к прямоугольной матрице размером $n \times p$, где p – количество слов в коллекции (размерность пространства), а n – число документов.

20. Полиграммная модель со степенью n и основанием M предполагает представление текстового документа в виде вектора $\{f_i\}$, $i=1, \dots, M^n$, где f_i – частота встречаемости i -ой n -граммы в тексте, которая является последовательностью подряд идущих n – символов вида $a_1 \dots a_{n-1} a_n$, причем символы a_i принадлежат алфавиту, размер которого совпадает с M .

21. Модель терм-документ представляет модель, в рамках которой текст описывается лексическим вектором $\{\tau_i\}$ $i=1..N_w$, где τ_i – важность (информационный вес) термина w_i в документе, N_w – полное количество терминов в документной базе (словаре). Вес термина, отсутствующего в документе, принимается равным 0.

22. Дейтамайнинг позволяет извлекать новые знания (скрытые закономерности, факты, неизвестные взаимосвязи и т.п.) из больших объемов структурированной информации (хранимой в базах данных).

23. Текстомайнинг позволяет находить новые знания в неструктурированных текстовых массивах.

24. Неструктурированные текстовые документы – документы в форматах html, doc, rtf и подобных им. В этом случае изначально предполагается, что с содержимым таких документов будет знакомиться человек. Разумеется, речь не идет о непосредственном доступе человека к содержимому. Человек знакомится с экранным представлением таких документов в окне просмотра браузера Internet Explorer, например, либо в редакторе текстов Microsoft Word, либо в любой другой аналогичной среде.

25. Слабоструктурированные текстовые документы – для такого рода документов будет наиболее эффективным их представление в виде совокупно-

сти объектов с последующей возможной идентификацией текстовых фрагментов как атрибутов и значений данных.

26. Структурированные текстовые документы – документы в формате xml, rdf и подобных им. Эти форматы ориентированы на описание данных. Их содержимое аналогично содержимому таблиц базы данных. Поэтому при автоматизированной обработке таких документов не возникает проблем с идентификацией и доступом к данным, содержащимся в них. Причем сами эти документы, как правило, также получены как итог работы того или иного программного обеспечения.

27. Короткий текстовый документ – это текстовый документ, написанный на естественном языке и содержащий информацию в лингвистической или цифровой форме, объем которого не позволяет применять известные процедуры статистического анализа текстов, но допускает использование для его анализа экспертной информации, полученной в результате комплексирования знаний лингвистов и специалистов в рассматриваемых предметных областях.

28. Очень короткие для анализа – текстовые документы информации, в которых недостаточно для рубрицирования даже экспертными методами.

29. Длинные документы – текстовые документы, анализ которых статистическими методами даёт точный результат.

30. Уникальные – слова, встречающиеся только в текстовых документах одной рубрики.

31. Редкие – слова, встречающиеся в текстовых документах некоторой группы рубрик, занимающихся смежной отраслью.

32. Общие – слова, которые употребляются во всех или почти во всех предметных областях и которые не несут никакого признака рубрики.

33. Теги – именованная метка специальной языковой разметки.

34. Морфемы – наименьшая единица языка, имеющая некоторый смысл.

35. Лемма – начальная форма слова.

36. Рубричное поле – набор рубрик и их параметров, актуальных на момент рубрицирования ЭНТД.

ПРИЛОЖЕНИЕ 1 *Результаты тестирования разработанных алгоритмов автоматизированного рубрицирования ЭНТД*

```

Морфологический анализ документа ...

файл записан
слова указанного файла - .\2\output_for_teach\windows-разное-9979.txt6292.txt бы
ли преобразованы в морфемы
"пауза 2 пинга"
"ping" не является внутренней или внешней
командой, исполняемой программой или пакетным файлом.
"Найден новый документ : ".\2\output_for_teach\windows-разное-9980.txt6293.txt
"Третий этап - морфологический анализ ..."
создаю подключение к БД морфологического словаря ...
открываю подключение ...
начало программы 3
считали входной файл = .\2\output_for_teach\windows-разное-9980.txt6293.txt
нашли тэг текста документа
нашли тэг конца текста документа
выделен заголовок
выделен текст
выделен конец
Начинаем получать морфемы всех слов
затрачено времени 00:01:10.16
затрачено времени на перебор первого листа (морфемы слов) 00:01:06.71
затрачено времени на перебор второго листа (параметры слов) 00:00:00.00
среднее время поиска одной морфемы 0,0050791194281578 секунд
количество обработанных слов 13815шт
Закончили получать морфемы всех слов
новый файл = .\3Eng\output_for_teach\.\txt
файл записан
слова указанного файла - .\2\output_for_teach\windows-разное-9980.txt6293.txt бы
ли преобразованы в морфемы
"пауза 2 пинга"
"ping" не является внутренней или внешней
командой, исполняемой программой или пакетным файлом.
"Найден новый документ : ".\2\output_for_teach\windows-разное-9981.txt6294.txt
"Третий этап - морфологический анализ ..."
создаю подключение к БД морфологического словаря ...
открываю подключение ...
начало программы 3
считали входной файл = .\2\output_for_teach\windows-разное-9981.txt6294.txt
нашли тэг текста документа
нашли тэг конца текста документа
выделен заголовок
выделен текст
выделен конец
Начинаем получать морфемы всех слов

```

Рисунок П.1 – Экранная форма работы морфологического анализа

Результаты сегментации и морфологического анализа:

<XML>

<INFBLOCK>мас-оборудование<\INFBLOCK>

<HEAD>

<number_doc>3961<\number_doc>

<\HEAD>

<TEXT>

<ABZAC>

<PREDLOJ>

<WORD>i<\WORD>

<WORD>remember<\WORD>

<WORD>reading<\WORD>

<WORD>that<\WORD>

<WORD>apple<\WORD>

<WORD>also<\WORD>

<WORD>has<\WORD>

<WORD>a<\WORD>

<WORD>patent<\WORD>

<WORD>on<\WORD>

<WORD>their<\WORD>

<WORD>hardware<\WORD>

<WORD>and<\WORD>

<WORD>that<\WORD>

<WORD>the<\WORD>

<WORD>clones<\WORD>

<WORD>would<\WORD>

<WORD>therefore<\WORD>

<WORD>be<\WORD>

<WORD>lacking<\WORD>

<WORD>an<\WORD>

<WORD>port<\WORD>

<WORD>whatother<\WORD>

<WORD>patents<\WORD>

<WORD>does<\WORD>

<WORD>apple<\WORD>

<WORD>have<\WORD>

<WORD>on<\WORD>

<WORD>the<\WORD>

<WORD>mac<\WORD>

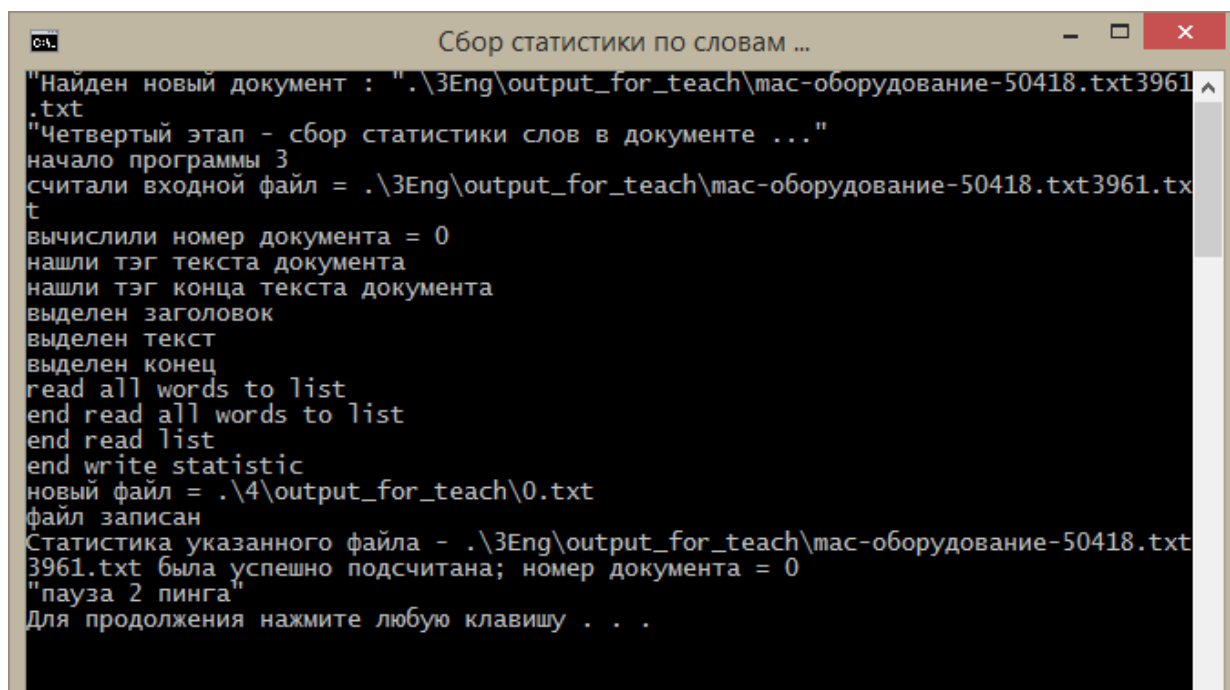
<\PREDLOJ>

<\ABZAC>

<\TEXT>

<\XML>

На рисунке 2 представлена экранная форма работы сборщика статистических характеристик.



```

Сбор статистики по словам ...
"Найден новый документ : ".\3Eng\output_for_teach\mac-оборудование-50418.txt3961
.txt
"Четвертый этап - сбор статистики слов в документе ..."
начало программы 3
считали входной файл = .\3Eng\output_for_teach\mac-оборудование-50418.txt3961.tx
t
вычислили номер документа = 0
нашли тэг текста документа
нашли тэг конца текста документа
выделен заголовок
выделен текст
выделен конец
read all words to list
end read all words to list
end read list
end write statistic
новый файл = .\4\output_for_teach\0.txt
файл записан
Статистика указанного файла - .\3Eng\output_for_teach\mac-оборудование-50418.txt
3961.txt была успешно подсчитана; номер документа = 0
"пауза 2 пинга"
Для продолжения нажмите любую клавишу . . .
  
```

Рисунок П.2 – Экранная форма сборщика статистических характеристик

Результаты работы сборщика статистических характеристик при морфологическом анализе слов?:

<STATISTICS>

from:2

subject:1

nutek:5

faces:1

apple's:3

wrath:1

article:1

organization:1

mapinfo:2

corporation:2

troy:1

lines:1

posting:1

host:1

writes:1

believe:2

apple:6

patent:3

region:3

features:1

implement:1

regions:5

this:3

possible:1

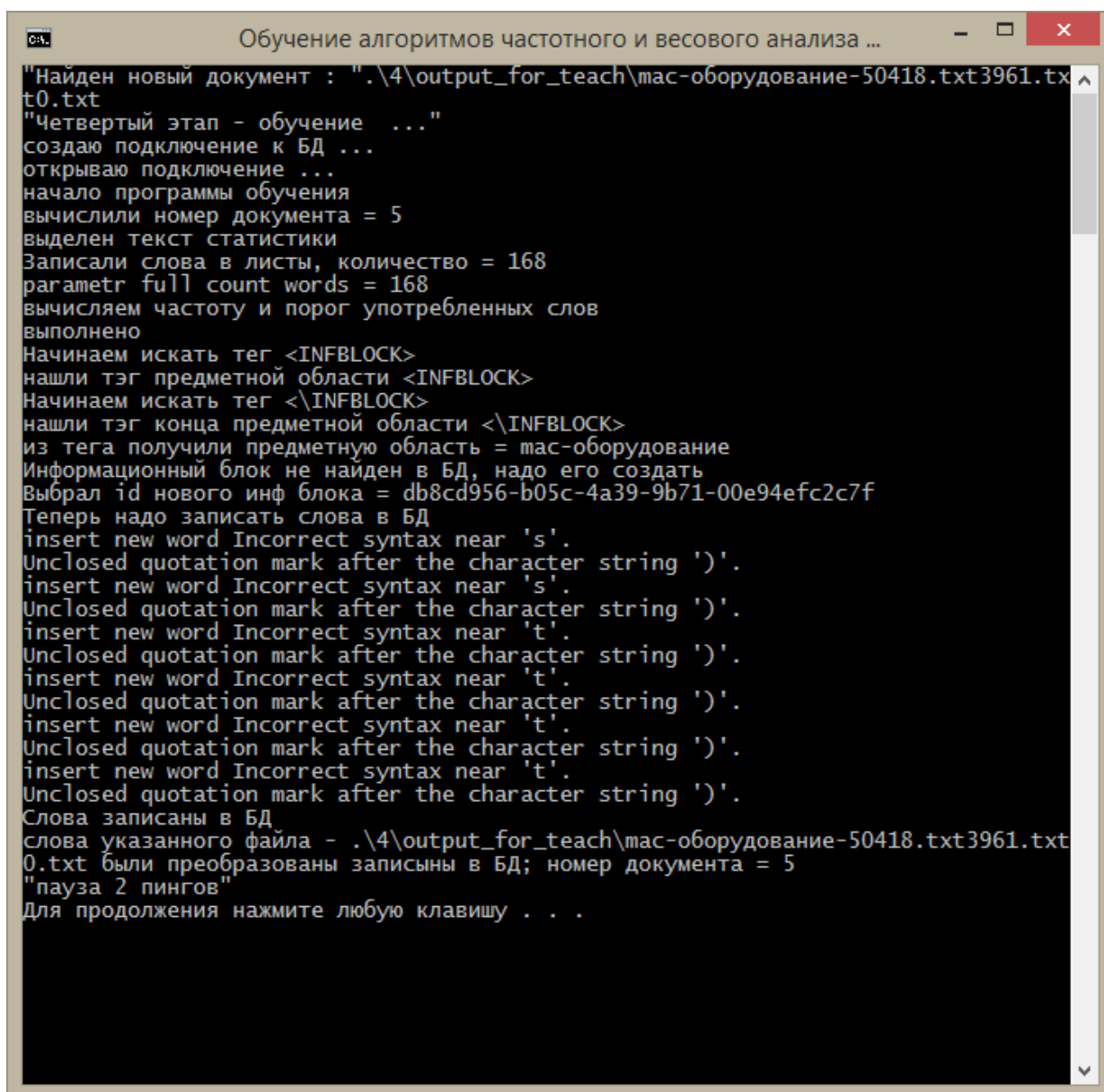
make:1

without:1

<FullCount>168<\FullCount>

<\STATISTICS>

Экранная форма обучения представлена на рисунке 3.



```

Обучение алгоритмов частотного и весового анализа ...
"Найден новый документ : ".\4\output_for_teach\mac-оборудование-50418.txt3961.tx
t0.txt
"Четвертый этап - обучение ..."
создаю подключение к БД ...
открываю подключение ...
начало программы обучения
вычислили номер документа = 5
выделен текст статистики
Записали слова в листы, количество = 168
parametr full count words = 168
вычисляем частоту и порог употребленных слов
выполнено
Начинаем искать тег <INFBLOCK>
нашли тэг предметной области <INFBLOCK>
Начинаем искать тег <\INFBLOCK>
нашли тэг конца предметной области <\INFBLOCK>
из тега получили предметную область = mac-оборудование
Информационный блок не найден в БД, надо его создать
Выбрал id нового инф блока = db8cd956-b05c-4a39-9b71-00e94efc2c7f
Теперь надо записать слова в БД
insert new word Incorrect syntax near 's'.
Unclosed quotation mark after the character string ')'.
insert new word Incorrect syntax near 's'.
Unclosed quotation mark after the character string ')'.
insert new word Incorrect syntax near 't'.
Unclosed quotation mark after the character string ')'.
insert new word Incorrect syntax near 't'.
Unclosed quotation mark after the character string ')'.
insert new word Incorrect syntax near 't'.
Unclosed quotation mark after the character string ')'.
insert new word Incorrect syntax near 't'.
Unclosed quotation mark after the character string ')'.
Слова записаны в БД
слова указанного файла - .\4\output_for_teach\mac-оборудование-50418.txt3961.txt
0.txt были преобразованы записаны в БД; номер документа = 5
"пауза 2 пингов"
Для продолжения нажмите любую клавишу . . .
  
```

Рисунок П.3 – Экранная форма пересчета весовых коэффициентов

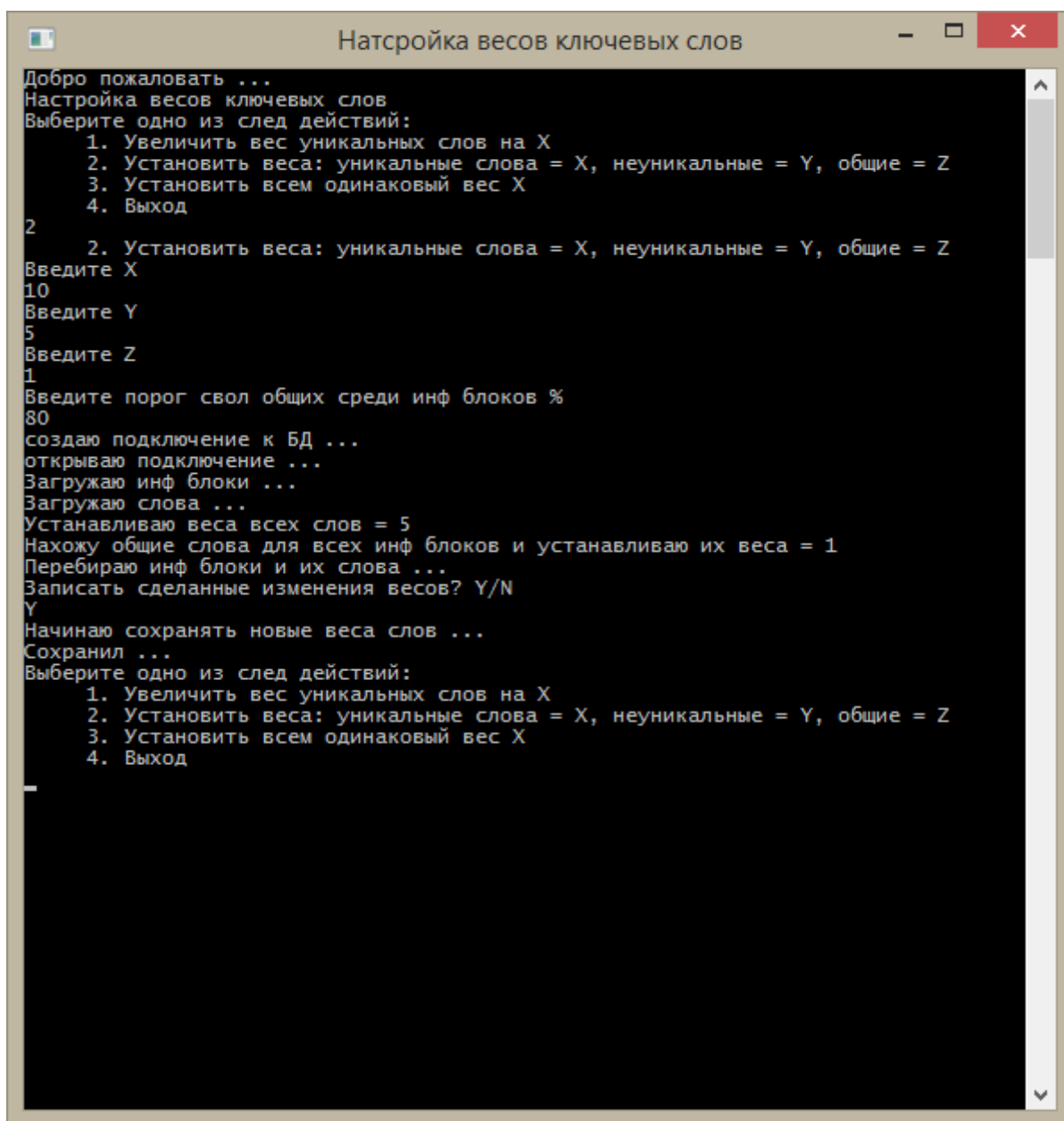


Рисунок П.4 – Экранная форма ручной настройки весовых коэффициентов

Результаты рубрицирования одного текстового документа:

<all inf blocks with freqs and weights>

<inf block id>db8cd956-b05c-4a39-9b71-00e94efc2c7f<\inf block id>

<inf block name>mac-оборудование<\inf block name>

<weight>854<\weight>

<freq>0,433084329089021<\freq>

<inf block id>a47124e0-457c-4a6b-be18-30e1deabcbf5<\inf block id>

<inf block name>windows-oc<\inf block name>

<weight>193<\weight>

<freq>0,129598131626516<\freq>

<inf block id>1291210c-afea-4fdb-b6ad-32eb7a57686d<\inf block id>

<inf block name>windows-разное<\inf block name>

<weight>71<\weight>

<freq>0,0728495615404262<\freq>

<inf block id>4cb29fac-d380-4d9c-9059-76e94c02c9fd<\inf block id>

<inf block name>pc-оборудование<\inf block name>

<weight>79<\weight>

<freq>0,0587049634668682<\freq>

<inf block id>954165b9-c903-414f-af2e-aed5ed54de5d<\inf block id>

<inf block name>автотехника<\inf block name>

<weight>104<\weight>

<freq>0,0376124757688858<\freq>

<\all inf blocks with freqs and weights>