**Федеральное государственное бюджетное образовательное учреждение  
высшего образования  
«Саратовский государственный технический университет имени  
Гагарина Ю.А.»**

*На правах рукописи*

**Петров Андрей Сергеевич**

МАТЕМАТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ ИЗВЛЕЧЕНИЯ ТЕРМИНОВ ИЗ РУССКОЯЗЫЧНЫХ ТЕКСТОВ

|  |
| --- |
| *(наименование темы научно-квалификационной работы)* |

09.06.01 Информатика и вычислительная техника

|  |
| --- |
| *(код и наименование направления)* |

Математическое моделирование, численные методы и  
комплексы программ

|  |
| --- |
| *(наименование направленности)* |

Научный доклад

Научный руководитель д.ф.-м.н., профессор, проф. каф. ИКСП

*(ученая степень, ученое звание, должность)*

Шульга Татьяна Эриковна

*(ФИО полностью)*

Рецензент д.ф.-м.н., профессор

*(ученая степень, ученое звание, должность)*

Бровко Александр Валерьевич

*(ФИО полностью)*

Рецензент

*(ученая степень, ученое звание, должность)*

*(ФИО полностью)*

Саратов 2019

## ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

**Актуальность темы.** Специалисты по терминоведению и компьютерной лингвистике длительное время ведут споры по поводу природы понятия «термин». Под термином обычно понимается слово или словосочетание на естественном языке, описывающее понятие конкретной предметной области. Извлечение терминов из текста — это процесс выделения из текста и распознавания ключевых слов и терминов, описывающих понятия определенной предметной области. Извлечение терминов из текста является важным этапом в решении ряда задач, связанных с обработкой текстов предметной области, среди которых — информационный поиск, машинный перевод, классификация и кластеризация документов, построение онтологий, глоссариев и тезаурусов.

Существует большое количество методов автоматического извлечения терминов из текста на естественном языке. Часть из них создавалась с учетом особенностей ряда таких языков, относящихся к индоевропейским языкам романо-германской группы, как английский и французский, и не учитывает особенности грамматики и морфологии русского языка. Кроме того, эффективность работы данных методов остается низкой и варьируется в зависимости от предметной области и стиля исходного текста. Также большое количество методов извлечения терминов из текстов для обработки используют входные данные, размеченные вручную. Данные особенности и недостатки ограничивают их область применимости.

Большая часть существующих методов относятся к категории статистических методов извлечения терминов из текста. Исследованиям в области применения статистических критериев для выбора ключевых слов и понятий посвящены работы Roberto Navigli, Paola Velardi, Katerina Frantzi, Sophia Ananiadou, Hideki Mima, David A. Evans, Robert G. Lefferts. В исследованиях указанных авторов основным критерием, использующимся для процесса извлечения терминов из текста, является критерий частоты вхождения кандидатов в рассматриваемую коллекцию текстовых документов. Большая часть понятий в предметных областях являются многословными терминами. Однако в текстах, написанных на естественных языках, вне зависимости от предметной области такие понятия встречаются значительно реже по сравнению с однословными понятиями. Существующие методы по-разному компенсируют данный эффект. Так, методы, основывающиеся на подсчете меры ассоциации, учитывают вероятность совместного употребления слов в составе определенного термина. К таким метрикам относятся, в частности, взаимная информация (Mutual Information, MI), логарифмическое правдоподобие (Loglikelihood Ratio). В ряде методов учитывается вложенность понятий предметной области. К таким, например, относится C-Value. Другим подходом к улучшению результатов работы методов извлечения ключевых слов из текста является учёт контекста вхождений. Одним из методов, использующим данный подход, является, например, NC-Value.

Еще одним способом улучшения результатов извлечения многословных терминов является частота вхождений слов и словосочетаний во внешнюю по отношению к рассматриваемому тексту коллекцию документов с точки зрения предметной области. Причем документы подбираются таким образом, чтобы их тематика не ограничивалась определенной предметной областью. То есть в случае обработки исходного текста военно-исторической предметной области подбирается корпус новостей или художественной литературы. Текстовый корпус – это сформированная по определенным правилам, структурированная коллекция текстов. К методам, учитывающим данную метрику, относятся, в частности, Domain Relevance, Weirdness, Relevance, Domain Pertinence.

Часть методов извлечения терминов из текста применяют внешние ресурсы: например, поисковые машины или энциклопедию Википедия. Большая часть подобных методов не использует заранее подготовленную коллекцию текстовых документов, применяя исключительно внешние ресурсы для поиска и извлечения терминов. Однако, при таком подходе без предварительного составления текстового корпуса достаточно сложно добиться высокой точности извлечения перечня терминов.

Таким образом, значительная часть существующих методов извлечения терминов из текста, основывающихся на использовании статистических метрик для выбора понятий, работает только с текстами выбранной предметной области. Как правило, тексты, используемые в процессе обработки, обычно не являются размеченными текстовыми корпусами, в связи с чем не содержат в себе необходимого объема, в том числе, лингвистической и морфологической информации для автоматического извлечения терминов. Ряд методов для решения этой проблемы используют внешние ресурсы, такие как поисковые машины или контрастные корпуса текстов других предметных областей. Текстовые документы в большинстве своем не имеют структуры и позволяют использовать только статистическую информацию о частоте встречаемости слов и словосочетаний в тексте без учета особенностей рассматриваемой предметной области.

В качестве модели представления исходных текстовых данных большая часть статистических методов извлечения терминов из текста использует модель «мешок слов» (англ. Bag of words). Данная модель представляет текстовый документ в виде множества составляющих его слов, а также их частот встречаемости. Разновидностью модели Bag of words является частотная модель текста, в которой каждому слову соответствует весовой коэффициент, который определяется в зависимости от выбранной метрики: например, частоты вхождения слова в документе TF (англ. Term frequency), логарифма частоты вхождения слова Log TF или обратной частоты документов IDF (англ. Inverse document frequency). Однако вышеописанные модели не учитывают порядок слов и могут быть использованы для извлечения лишь однословных понятий из текста.

Для решения этой проблемы вводится модель коллокации N-грамма (англ. N-gram). Коллокация — словосочетание, являющееся синтаксически и семантически целостной единицей. Данная модель позволяет представить словосочетания из N слов, при N=2 словосочетание имеет название биграммы, при N=3 — триграммы. Таким образом, для извлечения многословных терминов возможно использование частотной модели текста вместе с моделью коллокаций N-грамма. Однако, данные модели не принимают во внимание грамматические особенности русского языка, с учётом которых возможно повышение качества результатов методов извлечения терминов из текста.

Вышесказанное определило актуальность настоящей работы, а также, следующие из неё цели и задачи.

**Цель работы** – cодержательная постановка задачи автоматического извлечения терминов из текста, разработка математической модели русскоязычного текстового документа, которая может быть использована для решения рассматриваемой задачи, проведение исследования статистических методов выделения понятий на текстах исторической предметной области

**Задачи работы**. Для достижения поставленной цели были поставлены и решены следующие задачи:

1. Исследовать существующие математические модели и методы, применяемые при автоматическом извлечении терминов из текста.
2. Разработать математическую модель русскоязычного текстового документа.
3. Создать комплекс проблемно-ориентированных программ, реализующий разработанную математическую модель и методы автоматического извлечения терминов из русскоязычного текста.

**Объект исследования** – процесс извлечения терминов из русскоязычных текстовых документов.

**Предмет исследования** – математические модели и методы, применяемые при выделении ключевых слов и понятий.

**Методы исследования**: аппарат теории множеств, методы обработки многомерных данных, методы параллельного и объектно-ориентированного программирования, методы статического анализа.

**Научная новизна работы** (соответствует пунктам 1, 4 паспорта специальности 05.13.18) заключается в следующем:

Разработана математическая модель текстового документа на русском языке, учитывающая грамматические особенности русского языка, описанная в форме терминологии с использованием аппарата теории множеств. С помощью разработанного комплекса проблемно-ориентированных программ исследована применимость методов автоматического извлечения терминов из текста на текстах военно-исторической предметной области.

**Практическая значимость работы** обусловлена созданием комплекса проблемно-ориентированных программ для извлечения терминов из коллекции текстовых документов.

**Основные положения и результаты**, выносимые на защиту:

1. Математическая модель предметной области текстового документа на русском языке, учитывающая грамматические особенности русского языка, описанная в форме терминологии с использованием аппарата теории множеств.
2. Разработанный комплекс проблемно-ориентированных программ «Программный комплекс для извлечения терминов из текстов», на основе предложенной модели.

**Достоверность результатов**.

Достоверность и обоснованность результатов работы определяется корректной постановкой задач, применением методов статистического анализа, обработки многомерных данных. Результаты исследования подтверждены вычислительными экспериментами и практической апробацией.

**Апробация работы**. По основным результатам работы сделано 4 доклада на 4 международных, всероссийских и региональных конференциях:

1. Международная научно-практическая конференция «Проблемы управления в социально-экономических и технических системах» (ПУ-2016).
2. Международная научная конференция «Информационно-коммуникационные технологии в науке, образовании и производстве» (ICIT-2016).
3. Международная научная конференция «Информационно-коммуникационные технологии в науке, образовании и производстве» (ICIT-2017).
4. Международная научная конференция «Проблемы управления, обработки и передачи информации» (УОПИ-2017).

**Публикации**

Результаты работы опубликованы в 5 изданиях, 1 из которых является изданием, рекомендованным ВАК, 4 индексируются в базе РИНЦ. Имеется свидетельство о государственной регистрации программы для ЭВМ №2017663283 от 28 ноября 2017 г.

**Соответствие темы диссертации требованиям паспорта специальностей научных работников.** Диссертационная работа выполнена в соответствии с паспортом специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ, п. 1. Разработка новых математических методов моделирования объектов и явлений; п. 4. Реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента.

**Структура и объем работы.** Научно-квалификационная работа состоит из введения, трех глав, заключения, списка использованной литературы и приложения. Работа содержит 53 страницы, включая 2 рисунка, библиографический список из 58 наименований.

## ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

**Во введении** показано современное состояние решаемой проблемы, обоснована актуальность темы диссертационной работы, сформулирована цель и основные задачи работы, обоснованы новизна полученных результатов, научная и практическая ценность работы, названы методы исследований, приведены сведения об апробации работы и структуре диссертации, сформулированы выносимые на защиту результаты и положения.

### **В первой главе** приведен обзор существующих определений базовых понятий для задачи извлечения терминов, определена cодержательная постановка задачи автоматического извлечения терминов из текста, проанализированы существующие методы ее решения.

Определен основной стандарт, фиксирующие понятия терминологии (ISO 1087-1:2000). Данный стандарт даёт следующие определения основным терминам данной области.

Термин - это словесное обозначение определенного понятия в определенной предметной области.

Обозначение - это представление понятия с помощью символьного обозначения.

Понятие - это единица знаний, образованная путем уникального объединения характеристик.

Характеристика - это абстракция свойств объекта или множества объектов.

В ходе анализа работа был выявлен ряд признаков понятия «термин», отличительных по сравнению с общеупотребительной лексикой. В работах по данной предметной области исследователями выделяются следующие группы признаков:

1. Синтаксические признаки
   1. Номинативность.
   2. Нормативность.
   3. Терминологическая инвариантность.
   4. Мотивированность или самообъяснимость термина.
2. Семантические признаки
   1. Системность.
   2. Соответствие обозначаемому понятию.
   3. Однозначность или моносемантичность термина.
   4. Содержательная точность.
3. Прагматические признаки
   1. Внедрённость или общепринятость.
   2. Дефиницированность.
   3. Независимость от контекста.
   4. Вариационная устойчивость.
   5. Благозвучность.

Проведенный анализ работ других исследователей показал, что большая часть существующих методов извлечения терминов из текста используют статистические метрики подсчета весов кандидатов в термины. Выделен основной алгоритм работы статистических методов извлечения терминов из текста, согласно которому существуют следующие этапы:

1. Сбор кандидатов в термины

Извлечение слов и словосочетаний из коллекции текстовых документов по определенным условиям, задаваемым лингвистическим фильтров.

Фильтрация кандидатов по частоте, содержанию и длине слов кандидата или содержанию в них особых символов.

1. Подсчет признаков.

перевод каждого терминологического кандидата в вектор пространства признаков.

1. Вывод на основе признаков

Сортировка кандидатов по рассчитанной оценке и отбор заранее определенного числа кандидатов.

Существующие математические модели, использующиеся в существующих методах извлечения терминов из текста, позволяют описать текстовый документ. Однако, они не принимают во внимание грамматические особенности русского языка, с учётом которых возможно повышение качества результатов методов извлечения терминов из текста.

### **Во второй главе** приведено описание математической модели русскоязычного текстового документа, учитывающей особенности грамматики русского языка.

### Разработана математическая модель русскоязычного текстового документа.

### Разработанная модель может использоваться для описания текстового корпуса, который будет являться входными данными для методов извлечения ключевых слов и терминов из текста. Также модель учитывает лингвистическую информацию, необходимую для работы лингвистического фильтра на этапе сбора терминологических кандидатов.

Пусть дан текстовый документ определенной предметной области *D,* состоящий из множества предложений *si*.

*D* = {*s1,…,* *si,…,sn* },

где *si* - предложение

Предложение *si* состоит из множества слов *wij*, разделенных между собой символами-разделителями слов *sep'*ib′, а также из символов конца предложения sepik .

*,*

где *i* - номерпредложения,

n - количество предложений в текстовом документе *D*

k′- номер символа конца предложения,

k - количество символов конца предложения si

b′ - номер символа-разделителя слова,

b - количество символов-разделителей слов wij в предложении si

*wij* - слово, *wij* ∈ W

Опишем слово *wij* предложения *si* как множество, состоящее из символов русского алфавита и символа дефиса

,

где j - номер слова в предложении si,

m - количество слов в предложении si

lija′ - символ в слове wij,

*L* – множество символов кириллического алфавита, а также символ дефиса (‘-’)

a′ - номер буквы в слове wij,

a - количество символов в слове wij

sepik′ - символ окончания предложения si,

sep′ib′ - разделитель слов wij и wij+1 в предложении si,

Sep - множество символов окончания предложений.

Sep = {’.’, ’!’, ’?’}

Sep′ - множество символов-разделителей слов в предложении.

Sep′ = {’ ’, ’,’, ’;’, ’-’, ’:’}

Каждое слово обладает морфологической парадигмой *Wf* — системой словоформ, образующих одну лексему.

Словоформа – это слово в некоторой грамматической форме

Лексема — это единица словаря языка, объединяющая разные формы одного слова (например, метод, метода, методу, методы и т.д.), а также разные смысловые варианты слова, зависящие от контекста.

Таким образом, *Wf* - это множество словоформ одного слова.

,

где *wfe'* – словоформа,

e - количество словоформ одной парадигмы

e′ - номер словоформы, .

Опишем множество частей речи русского языка Pos. Элементами данного множества являются:

Pos = (’Существительное’, ’Местоимение’, ’Прилагательное’, ’Предлог’, ’Частица’, ’Союз’, ’Наречие’, ’Числительное’, ’Причастие’, ’Деепричастие’, ’Глагол’).

Каждая часть речи характеризуется особой системой грамматических категорий. Таким образом, в зависимости от части речи словоформа *wfe'* может быть описана по-разному. Пусть дано множество *Pos'*, Pos′ ∈ *Pos*, элементами которого являются именные части речи, а также причастие.

Pos′ = (’Существительное’, ’Местоимение’, ’Прилагательное’, ’Причастие’, ’Числительное’).

Для именных частей речи, а также для такой отглагольной части речи, как причастие, словоформа может быть описана следующим образом.

для

,

где w - слово, w ∈ W,

case - падеж, case ∈ Case,

n - число, n ∈ N

*g* – род, g ∈ G

Case = (’именительный’, ’родительный’, ’дательный’, ’винительный’, ’творительный’, ’предложный’)

N = (’единственное’, ’множественное’)

G = (‘мужской’, ’средний’, ’женский’)

Глагольная словоформа может быть описана следующим образом

для ,

,

где w – слово,

*t –* время, t ∈ T

n - число, n ∈ N

*p* – лицо, p ∈ *P*

T = {‘прошедшее’, ‘настоящее’, ‘будущее’}

P = {‘первое’, ‘второе’, ‘третье’}

Глаголы в форме прошедшего времени характеризуются, помимо грамматических категорий времени, числа и лица, также и такой категорией, как род *g*, g ∈ *G* (например, читал, читала).

для

В русском языке деепричастие, наречие, предлог, частица и союз являются неизменяемыми частями речи. Поэтому морфологической парадигмой *Wf* слов, принадлежащих к данным частям речи, является само слово *w*. Таким образом, *Wf =* {*w*}.

У каждого слова w имеется лемма wfnf.

Лемма - это исходная, базовая или нормальная форма слова, зафиксированная в словаре. Для именных частей речи, а также для причастия, нормальной является форма именительного падежа единственного числа. Лемма является одной из словоформ слова w. Таким образом,

,

где case′ = ′именительный′,

n′ = ′единственное′

Нормальной формой глагола является инфинитив. В инфинитиве морфологически не выражены такие категории глагола, как время, лицо и число.

Перейдем к описанию использованных функций. Обозначим DPos функцией определения части речи слова, *DNF* - функцией нормализации слова. Нормализация слова или лемматизация - это процесс приведения слова к лемме или начальной форме

### В третьей главе рассматривается процесс разработки комплекса проблемно-ориентированных программ.

Сформулированы следующие базовые функциональные требования для разрабатываемого программного обеспечения:

* чтение текстовых документов в форматах txt, html, pdf;
* выделение предложений в каждом из документов;
* морфологический разбор токенов в предложениях, хранение грамматических характеристик;
* реализация лингвистических фильтров Noun+  и (Adj|Noun)+Noun;
* реализация механизма фильтрации с помощью стоп-списка;
* реализация статистических методов C-Value, GlossEx, k-factor;
* сохранение промежуточной информации после проведения лексического и морфологического анализа в виде файла(ов);
* сохранение промежуточной информации после обработки терминологических кандидатов с помощью лингвистических фильтров в виде файла(ов);
* сохранение результирующего списка терминов в виде файла(ов).

Требуемый функционал рациональнее разделить на модули, что позволит следовать принципу единственности ответственности и упростит последующее тестирование. Таким образом, была спроектирована следующая структура ПО:

* Модуль Structures. Содержит классы, реализующие математическую модель русскоязычного текстового документа:
  + Case — падеж,
  + PartOfSpeech — часть речь,
  + TaggedWord — слово с грамматическими свойствама,
  + Separator — символ-разделитель.
  + Collocation — коллокация.
* Модуль TextImporter. Содержит классы для распознавания текста из исходных файлов формата pdf, html и txt. В случае, если в качестве входного файла будет передан неоцифрованный документ pdf, данный модуль должен произвести распознавание символов (англ. optical character recognition, OCR).
* Модуль LinguisticFilter. Содержит классы для реализации общего механизма фильтрации с помощью лингвистических фильтров, структур шаблонов, применяемых лингвистическими фильтрами, а также непосредственные реализации фильтров Noun+ и (Adj|Noun)+Noun.
* Модуль Stoplist. Содержит классы для реализации механизма фильтрации с помощью стоп-списка.
* Модуль Morph. Содержит классы для морфологического анализа слов.
* Модуль SentenceFetcher. Содержит класс для разбиения текста на предложения.
* Модуль StatMethods. Содержит классы, реализующие статистические методы C-Value, GlossEx, k-factor.

**В заключении** работы сформулированы основные результаты научно-квалификационной работы и сделанные выводы.

## ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Разработана математическая модель предметной области текстового документа на русском языке, учитывающая грамматические особенности русского языка, описанная в форме терминологии с использованием аппарата теории множеств. Данная модель предназначена для использования в качестве модели, описывающей текстовый корпус, являющийся входными данными для методов извлечения ключевых слов и терминов из текста.

## ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

*Публикации в центральных изданиях,  
включенных в перечень периодических изданий ВАК РФ*

1. **Петров А.С.,** Шульга Т. Э. Математическая модель русскоязычного текстового документа для решения задачи автоматического извлечения терминов из текста // Вестник Воронежского Государственного Университета. Серия: Системный анализ и информационные технологии. - 2017. - № 3. - С. 195-203.

*Свидетельства о регистрации программ для ЭВМ*

2. Шульга Т. Э., **Петров А.С.** Программный комплекс для извлечения терминов из текста // Свидетельство о государственной регистрации программы для ЭВМ №2017663283, 28.11.2017.

*Публикации в других изданиях*

3.**Петров А.С.,** Шульга Т. Э. О методах решения задачи автоматического извлечения терминов из русскоязычных текстовых документов // Проблемы управления, обработки и передачи информации (УОПИ-2017): сб. тр. V Междунар. юбилейн. науч. конф. 2017. - C. 334-340.

4.**Петров А.С.,** Шульга Т. Э. Применение статистических методов для решения задачи автоматического извлечения терминов из русскоязычных текстовых документов / А.С. Петров, Т.Э. Шульга // Информационно-коммуникационные технологии в науке, производстве и образовании ICIT-2017: материалы Международной научной конференции, Саратов, 21-22 сентября 2017 г. - 2017. - С. 239-246.

5. Шульга Т.Э., **Петров А.С.,** О задаче автоматического извлечения терминов из текста // Информационно-коммуникационные технологии в науке, производстве и образовании ICIT-2016: материалы Международной научно-практической конференции, Саратов, 23-28 августа 2016 г. - 2016. - С. 112-117.

6. **Петров А.С.** Обзор статистических методов извлечения терминов из текста // Проблемы управления в социально-экономических и технических системах: сборник научных статей - 2016. С. 163-167