**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
ДОНЕЦКОЙ НАРОДНОЙ РЕСПУБЛИКИ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

На правах рукописи

Бурлаева Екатерина Игоревна

СОВЕРШЕНСТВОВАНИЕ МЕТОДОВ СИСТЕМНОГО АНАЛИЗА В ЗАДАЧАХ УПРАВЛЕНИЯ И СИСТЕМАТИЗАЦИИ СПЕЦИАЛИЗИРОВАННОЙ ИНФОРМАЦИИ

05.13.01– Системный анализ, управление и обработка информации
(по отраслям) (технические науки)

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата технических наук

Донецк – 2019

Работа выполнена в ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ», г. Донецк.

Научный руководитель: доктор технических наук, доцент
Зори Сергей Анатольевич
профессор кафедры «Программная инженерия»
ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ», (г. Донецк)

Официальные оппоненты: **Фамилия, Имя, Отчество**
ученая степень, ученое звание, организация/место работы, должность

Фамилия, Имя, Отчество
ученая степень, ученое звание, организация/место работы, должность

Ведущая организация:

Полное название организации (в соответствии с уставом)

Защита состоится «__» _____ 2019 г. в 00.00 часов на заседании диссертационного совета Д 01.024.04 при ГОУВПО «ДОННТУ» и ГОУВПО «ДОННУ» по адресу: 283001, г. Донецк, ул. Артема, 58, корп. 1, ауд. 203 Тел./факс: 380(62) 304-30-55, e-mail: uchensovet@donntu.org.

С диссертацией можно ознакомиться в библиотеке ГОУВПО «ДОННТУ» по адресу: 283001, г. Донецк, ул. Артема, 58, корп. 2. Адрес сайта университета: http://donntu.org

Автореферат разослан «___» _____20__ г.

Учёный секретарь
диссертационного совета Д.01.024.04
кандидат технических наук Т.В. Завадская

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Обмен информацией является важнейшей частью современной деятельности человека. По оценкам экспертов около 70% накопленной и используемой информации находятся в несистематизированной текстовой форме, вследствие чего затрудняется получение требуемых сведений по конкретной тематике. Таким образом, возникает острая необходимость в создании систем, позволяющих автоматически систематизировать специализированную информацию.

В этой ситуации особую актуальность приобретают работы по созданию систем анализа, систематизации и управления специализированной текстовой информацией, так как даже высококвалифицированные эксперты испытывают затруднения по организации поиска документов и распределении полученных текстовых данных по тематикам.

Одним из способов систематизации, управления данными и их анализа является классификация информации, состоящая из сортировки текстовых документов по заранее определенным категориям.

Связь работы с научными программами, планами, темами. В основу диссертационного исследования положены работы, выполненные в Донецком Национальном техническом университете в рамках научно-исследовательской работы кафедры искусственного интеллекта и системного анализа Г/Т № Н 17-18 «Информационные технологии в системах моделирования и управления организационными и техническими объектами»; Г/Т № Н 17-13 «Разработка теоретических способов и методов создания современных информационных систем», в которых соискатель принимал участие как исполнитель.

Степень разработанности темы исследования. В исследованиях, посвященных применению методов машинного обучения для классификации текстов, применяются в основном универсальные алгоритмы, которые применимы для широкого круга задач анализа, управления и обработки специализированной информации. Качество рубрикации для систем, основанных на машинном обучении, является довольно высоким для небольших рубрикаторов, и значительно уменьшается с увеличением количества рубрик и усложнением структуры классификатора.

Цель методов машинного обучения для задачи классификации текстовых документов заключается в построении модели классификации на основе обучающего набора и применении ее для предсказания класса или набора классов, релевантных для нового документа. Разработке и тестированию моделей данного вида, а также связанными с ней алгоритмами обработки текстовой информации в настоящее время посвящены труды таких авторов как Агеев М.С., Кураленок И.Е., Joachims T., Schapire R., Schutze H., Scbastiani F и др.

Исследования показали, что эффективная обработка и анализ специализированной информации практически невозможны без разработки

комплексной модели процесса систематизации на основе машинного обучения и знаний эксперта, т.е. парадигматических и синтагматических подходов, т.к. состав и содержимое анализируемых документов постоянно изменяется.

В условиях роста информационного пространства и необходимости автоматизации информационных процессов повышение эффективности существующих моделей и методов для построения систем классификации и управления текстовой информацией является важной и актуальной научно–технической задачей, имеющей отраслевое значение.

Целью исследования является модернизация моделей, методик и алгоритмов системного анализа, искусственного интеллекта и компьютерной лингвистики для повышения эффективности автоматизации, систематизации специализированной текстовой информации больших объёмов с учетом особенностей русскоязычных текстовых документов.

Для достижения цели диссертационного исследования поставлены и решены следующие **задачи**:

1. Проанализировать существующие методы, модели и алгоритмы, используемые для решения задач систематизации и управления специализированной текстовой информации.
2. Разработать модифицированную модель автоматической обработки специализированной информации, основываясь на объединении достоинств парадигматических и синтагматических подходов.
3. Выявить закономерности изменения качества работы модели обработки текстовой информации при внесении изменений в ее конфигурацию.
4. Осуществить исследование выполненной разработки усовершенствованной модели обработки текстовой информации.
5. Обосновать подход к повышению эффективности применения предложенных моделей и методов для систем автоматической обработки специализированной информации.
6. Проанализировать качество работы предложенных средств при решении практических задач классификации текстов.

Объект исследования – процессы анализа, систематизации и управления текстовой информацией.

Предмет исследования – модели, методы и алгоритмы автоматизации процессов анализа, управления и систематизации специализированной текстовой информации.

Методы исследований. В процессе исследований использованы: методы машинного обучения и предварительной обработки текста; методы и алгоритмы анализа лингвистических особенностей языка; экспертный подход для решения поставленных задач управления и обработки информации; элементы аналитической алгебры и теории множеств.

Научная новизна полученных результатов:

1. Впервые предложена усовершенствованная общая модель автоматической систематизации и управления информацией, основанная на объединении достоинств синтагматических и парадигматических подходов. Использование новой модели позволяет повысить полноту и точность работы модели в среднем на 32,5% и 31,5% соответственно.

2. Получила дальнейшее развитие и модернизирована модель классификации текстовой информации на основе внесения изменений в структуру алгоритма её построения, что позволяет повысить полноту и точность работы модели еще на 5,5% и 8,5 % соответственно.

3. Экспериментально обоснована модель вычислительной композиции распределения веса слов в текстовом документе. При разном распределении веса терма, повышение качества работы предложенной общей модели систематизации и управления информацией варьируется в пределах 10% в зависимости от используемой композиции.

4. Обосновано решение задачи повышения эффективности предложенной общей модели систематизации и управления текстовой информацией на основе разработанных модернизированных моделей классификации информации за счет внесения изменений в структуру алгоритма её построения и вычислительных композиций распределения веса слов в документах, правил отбора неинформативных признаков и способов взвешивания термов. Применение предложенных усовершенствований обеспечивает дополнительное повышение качества распределения информации на 27,5%.

Теоретическая значимость работы. Предложенная комплексная методика построения модели автоматической классификации и статистического анализа является совершенствованием существующих подходов к обработке информации и в дальнейшем может быть расширена и дополнена функциями автоматического и автоматизированного тематического анализа потоков текстовой информации для расширения количества тематик, по которым распределяются текстовые документы, а также повышением качества модели автоматической обработки информации. Структура статистических баз данных, формируемых с помощью предложенной технологии, позволяет ставить и решать большой спектр статистических и математических расчетных задач, и задач, связанных с принятием решений, имеющих место в информационных системах. Развитие данной разработки может осуществляться путем дополнения ее новыми решениями в области морфологического, синтаксического и семантического анализа языков, для усовершенствования методов управления и систематизации специализированной текстовой информации.

Практическая значимость полученных результатов. Результаты исследований имеют широкий спектр применения для различных предметных областей. Предложенная практическая реализация усовершенствованной модели систематизации и управления информацией позволяет формировать

текстовые базы данных классифицированной информации в автоматическом режиме. На основании результатов классификации имеется возможность формировать аналитические задачи и статистические базы данных по результатам обработки текстов, автоматизировать работу специалистов–аналитиков, осуществляющих тематический анализ текстовой информации, и ведение аналитических задач в различных предметных областях, что может послужить функциональным дополнением и развитием информационных систем различных организаций.

Практическое значение полученных результатов подтверждается:

– внедрением в практику организации информационных массивов и баз данных с целью совершенствования компьютерной технологии прогноза в отделе сдвижения земной поверхности и охраны подрабатываемых объектов (СЗПО) Республиканского академического научно–исследовательского и проектно–конструкторского института горной геологии, геомеханики, геофизики и маркшейдерского дела (РАНИМИ) (справка о внедрении № 01/140 от 20.05.2019 г.);

– внедрением в учебный процесс ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ» (справка № 52.1–05/19 от 14.05.2019 об использовании в учебном процессе при чтении лекций и проведении практических занятий на кафедрах «Искусственный интеллект и системный анализ» и «Прикладная математика» по дисциплинам: «Организация баз данных и знаний», «Информационные системы и технологии», «Стандартизация и сертификация в сфере информационных технологий», «Распределённые информационно–аналитические системы», «Корпоративные информационные системы»).

Методология исследования. В процессе исследования выполнялся анализ закономерностей морфологии естественного языка (русского), анализ структуры существующих словарей, поисковых запросов и логических моделей возможных запросов, математический анализ методики оценки релевантности качества классификации, применялись методы концептуального анализа и управления в системах автоматической систематизации специализированной информации и оценки эффективности их работы, современные методы автоматического анализа и управления текстовыми документами.

Положения, выносимые на защиту:

– обоснована новая модель автоматической систематизации и управления информацией, основанная на объединении достоинств синтагматических и парадигматических подходов, использование которой позволяет повысить полноту и точность работы модели в среднем на 32,5% и 31,5% соответственно;

– применение вычислительной композиции методов в модели распределения веса слов в тексте позволяет повысить качество работы модели в среднем на 10% в зависимости от используемой композиции;

– усовершенствованная конфигурация модели систематизации информации на основе правил отбора неинформативных признаков и способов взвешивания термов за счет внесения изменений в структуру алгоритма её построения и вычислительных композиций распределения веса слов в документах обеспечивает дополнительное повышение качества распределения информации на 27,5%.

По направлению исследований, содержанию научных положений и выводов, существу полученных результатов диссертационная работа соответствует паспорту специальности 05.13.01 – Системный анализ, управление и обработка информации (по отраслям) согласно разделам:

2. Формализация и постановка задач системного анализа, оптимизации, управления, принятия решений и обработки информации.

4. Разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации.

6. Методы идентификации систем управления на основе ретроспективной, текущей и экспертной информации.

Степень достоверности и апробация результатов обеспечивается полнотой анализа теоретических и практических исследований, положительной оценкой на научных конференциях и семинарах, выполненными публикациями.

Основные результаты настоящего диссертационного исследования представлены в 9 публикациях, в том числе 5 статей в изданиях, рекомендованных ВАК ДНР, основные положения и научные результаты диссертационной работы докладывались, обсуждались и получили положительную оценку на следующих 4 конференциях:

– VIII Международная научно–техническая конференция «Информационные перспективы Донбасса» (г. Донецк, 2017);

– V Международная научно–техническая конференции «Современные информационные технологии в образовании и научных исследованиях» (г. Донецк, 2017);

– XXV международная научно–техническая конференция «Машиностроение и техносфера XXI века» (г. Севастополь, 2018);

– Научно–техническая конференция «Донецк будущего глазами молодых ученых» (г. Донецк, 2018).

Личный вклад. Основные научные результаты диссертации включают в себя разработку новой модели систематизации и управления информацией на основе объединения парадигматических и синтагматических подходов, комплекс вычислительных композиций в модели распределения веса слов в тексте для предсказания релевантной для документа тематики или набора тематик, а так же усовершенствованную конфигурацию модели систематизации информации на основе модернизации алгоритма её построения и вычислительных композиций распределения веса слов в документах, применение которых повышает качество и скорость автоматической систематизации и обработки информации.

Все выносимые на защиту положения получены автором лично.

Публикации. Основные научные результаты диссертации опубликованы автором самостоятельно и в соавторстве в 9 научных изданиях, 5 из них в рецензируемых научных изданиях: в том числе 2 – в рецензируемых научных журналах и изданиях, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата и доктора наук в Российской Федерации, и 3 - в специализированных научных изданиях, рекомендованных ВАК ДНР, 4 – по материалам научных конференций.

Объем и структура диссертации. Диссертация изложена на 161 странице машинописного текста и состоит из введения, четырех глав, заключения, списка литературы, приложений. Работа иллюстрирована 28 рисунками, содержит 19 таблиц. Указатель литературы включает 140 источников.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Введение содержит общую характеристику работы. Обоснована актуальность темы диссертации, сформулированы цели и задачи исследований. Показана научная новизна и практическое значение полученных результатов.

В **первой главе** выполнен анализ основных подходов, тенденций и закономерностей управления информацией, обеспечивающих более точное и быстрое извлечение и поиск полезных сведений в документальных базах.

Анализ и систематизация документов, оформляемых в виде текстового описания, предполагает работу с информацией, ее глубоким осмыслением, принятием адекватных решений по анализу и управлению той или иной ситуацией. Такие действия могут быть выполнены только экспертом, но с учетом огромного числа электронных документов их эффективная обработка человеком становится весьма затратной как по времени, так и по используемым ресурсам. С другой стороны, отсутствие возможности вовремя и быстро получить необходимую информацию по нужной теме делает бесполезной большую часть накопленных знаний, вследствие чего появляется необходимость в управлении, анализе и систематизации текстовой информации, выполняемой автоматическими системами обработки информации.

Для создания эффективных систем управления и систематизации информацией требуется применение методов классификации на основе машинного обучения и методов, основанных на знаниях. Поскольку состав и содержимое анализируемых документов постоянно изменяется, одним из направлений адаптации к этой динамике является использование методов машинного обучения. Цель методов машинного обучения для задачи классификации текстовых документов заключается в построении модели классификации на основе обучающего набора и применении построенной модели для предсказания класса или набора классов, релевантных для

документа. Такой подход может обеспечивать качество классификации, сравнимое с производимым человеком.

Описание сложных систем осуществляется с помощью методологий семейства *IDEF*, использование которой позволит наглядно продемонстрировать внутреннюю архитектуру исследуемой системы, систематизирующей информацию, что предоставит возможность детализации каждого блока модели, осуществляющей управление и систематизацию информации.

Формализация единого универсального системного способа представления знаний позволит создать соответствующие алгоритмы и инструментальные средства для обработки знаний различного типа единообразным способом и с помощью единого формального аппарата, построение которого осуществляется на основе парадигматических и синтагматических подходов, являющимся бурно развивающимся научным направлением в рамках анализа специализированной текстовой информации. На основе выполненного анализа сформулирована актуальная цель и вытекающие из нее задачи исследования, а также необходимость в построении модернизированной модели для предсказания тематики или набора тематик для нового документа.

Во **второй главе** проведен анализ основных подходов обработки текстовой информации и применения методов оценки результатов классификации, проанализированы базовые технологии машинного обучения и лингвистические процессы естественного языка, а также обоснован выбор метода для построения модели модернизированной системы систематизации и управления текстовой информацией.

Лингвистические процессы естественного языка, являются неотъемлемой частью модели, используемой для систематизации текстовой информации. Этапы, составляющие структуру систем анализа текстов, последовательно обрабатывают получаемую информацию (рисунок 1).

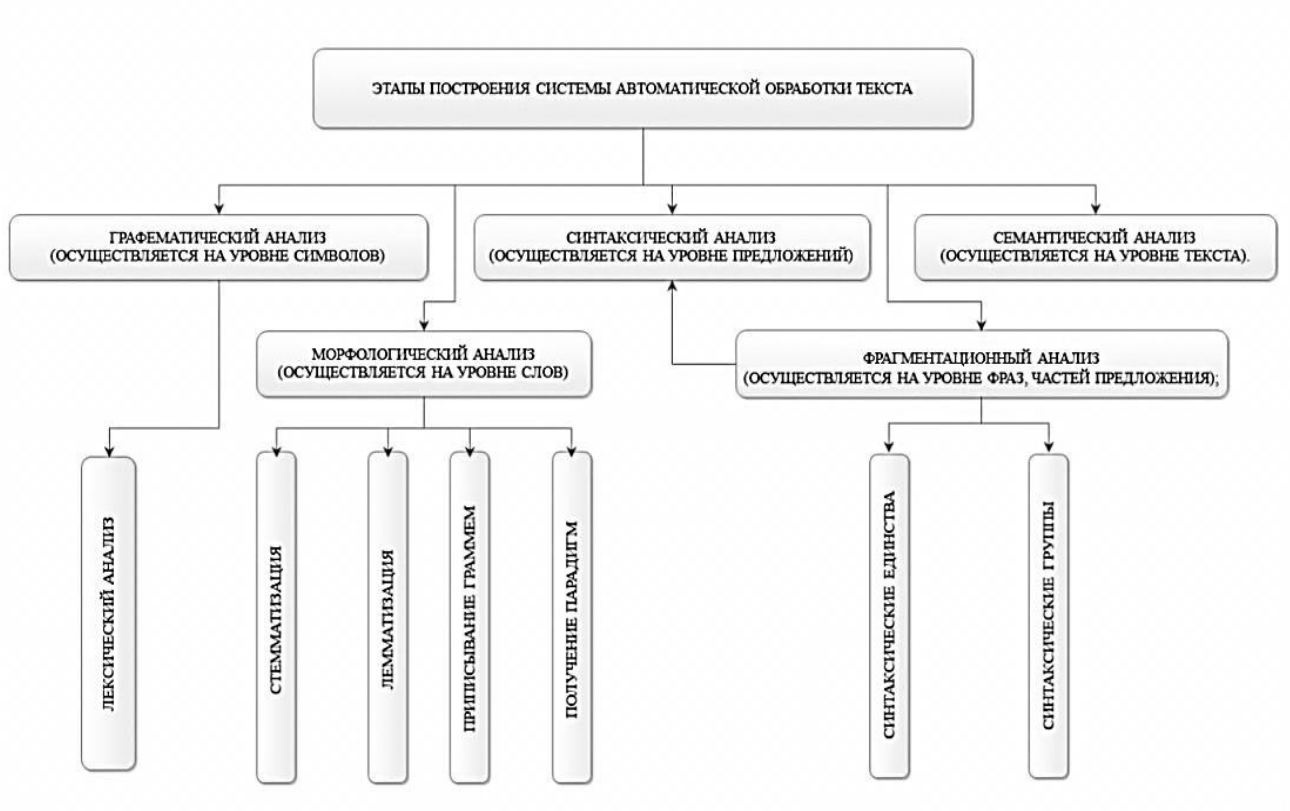


Рисунок 1 – Этапы построения структуры систем анализа текстов

Текстовые документы в непосредственном виде не подходят для интерпретации классификатором или алгоритмом построения классификатора, поэтому необходимо применение процедуры индексации. Трансформация теста в векторы, является стандартным представлением документа, используемым в машинном обучении для систематизации информации.

Выбор основного алгоритма осуществляется на основании выполненного сравнительного анализа алгоритмов машинного обучения (рисунок 2). Некоторые алгоритмы обучения делают определенные предположения о желаемых результатах и структуре данных. Если найти алгоритм, который соответствует потребностям, то можно уменьшить время обучения и получить более точные результаты и прогнозы.

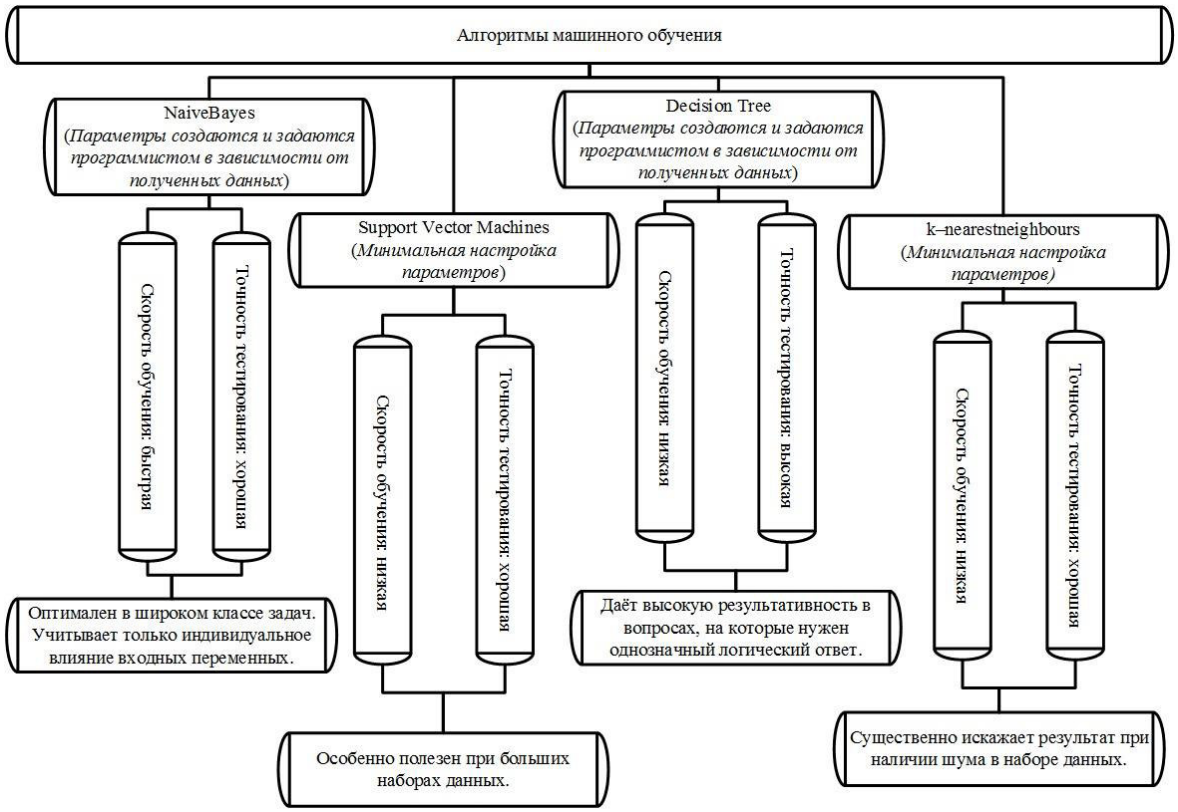


Рисунок 2 – Сравнительный анализ алгоритмов машинного обучения

Проведенный анализ позволил сделать следующие выводы:

– для улучшения характеристик классификатора для систематизации текста необходимо последовательное объединение нескольких алгоритмов классификации;

– при формировании последовательности необходимо оптимизировать критерии качественного обучения каждого метода классификации;

– для увеличения точности классификации при объединении методов и взаимодействии между собой необходимо их оптимальное сочетание.

Для дальнейших исследований на основании проведённого анализа сделан выбор одного из наиболее эффективных методов систематизации информации – *Support Vector Machines* (*SVM*). Использование метода *SVM* стало отправной точкой для его применения в разрабатываемой модели систематизации и управления специализированной текстовой информацией.

Третья глава посвящена созданию обобщенной модели систематизации и управления специализированной информацией для оптимизации управления и автоматической ее обработки с целью повышения эффективности систематизации, управления и своевременного доступа к актуальной информации больших объёмов.

Создание модели позволяет отобразить работу реальной системы управления и систематизации специализированной информацией и возможность исследовать ее функциональные характеристики с помощью

конструирования множества имитационных моделей, описывающих влияние того или иного воздействия на поведение системы.

При построении модернизированной модели управления и систематизации информации предложено использование оптимального способа анализа знаний, представляемого в виде стандарта *IDEF0*. Формализация данных единым универсальным системным способом представления знаний позволяет создать соответствующие алгоритмы и подобрать инструментальные средства для обработки знаний с помощью единого формального аппарата (рисунок 3).

Диаграмма *A–0*, отображает связи объекта моделирования с окружающей средой. В этот блок входит компоновка методов, позволяющая сократить временные затраты на обработку текстовых документов, и повысить точность отнесения документа к той или иной тематике. В основной блок входят функции импорта и классификации.

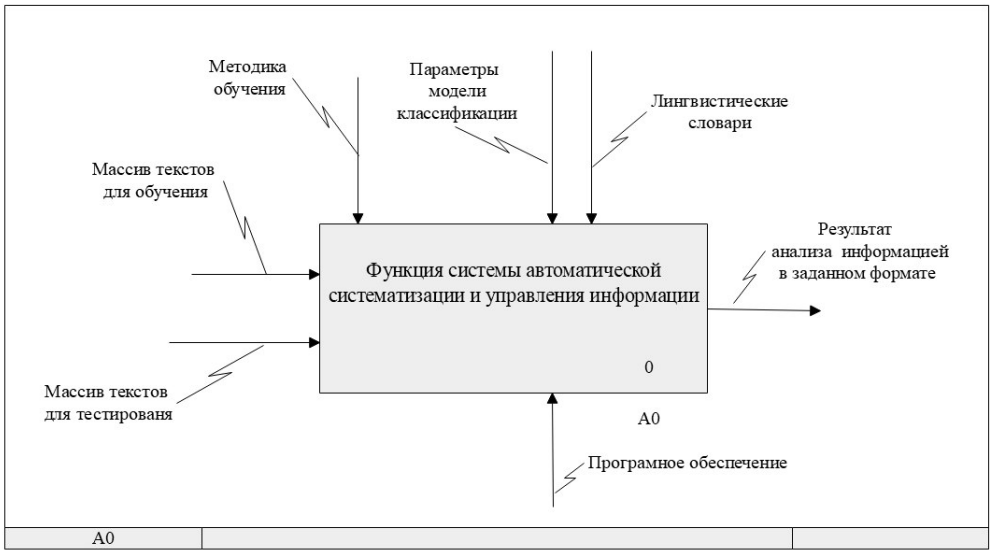


Рисунок 3 – Модернизированная модель управления и систематизации информации в *IDEF0 A–0*

Функция импорта представляет собой предварительную обработку текстовых документов, на основе правил отбора неинформативных признаков и способов взвешивания термов (таблица 2), использование которых повысит скорость и качества работы модели, систематизирующей информацию. Здесь:

Коэффициент *L* – показывает вес слова в классе и вычисляется так:

$$L = \frac{n_i}{|(d_i \supset w_i)|}, \quad (1)$$

где

n_i — число вхождений слова i в документ,

$|(d_i \supset w_i)|$ — количество документов, в которых встречается слово w_i

Под коэффициентом *Q* понимается словарь терминов конкретной рубрики, коэффициент *Z*–словарь со словами, входящими в название класса.

TF–IDF – статистическая мера, используемая для оценки важности слова в контексте документа.

Ш – метод «Шумовых слов» (в процессе предварительной обработки шумовые–слова удаляются из текста).

С – метод «Стемминга» (приведение слова к общей форме и отбрасывание окончаний).

Таблица 2 – Композиции моделей предварительный обработки текста для функции импорта

Сокращение	Краткое описание
1	2
T–I	Использование метода $TF-IDF$ для преобразования текста в вектор. По формуле: $\frac{n_i}{\sum_k n_k} \times \frac{\lvert D\rvert}{\lvert(d_i \supset w_i)\rvert}$.
T–I+Ш	Удаление слов, не несущих смысловой нагрузки.
T–I+C	Приведение слов к единой основе.
T–I+C+Ш	Приведение слов к единой основе. Удаление шумовых слов, не несущих в себе смысловую нагрузку.
((*T–I*)+*C*+*Ш*)+*L*–1	Использование коэффициента *L* для изменения веса слов. По формуле: $\frac{\mathrm{n_i}+\frac{\mathrm{n_i}}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert}}{\sum_k \mathrm{n_k}} \times \frac{\lvert \mathrm{D}\rvert}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert} + \mathrm{C} + \mathrm{Ш}$.
((*T–I*)+*C*+*Ш*)+*L*–2	Использование коэффициента *L* для изменения веса слов. По формуле: $\frac{\mathrm{n_i}*\frac{\mathrm{n_i}}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert}}{\sum_k \mathrm{n_k}} \times \frac{\lvert \mathrm{D}\rvert}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert} + \mathrm{C} + \mathrm{Ш}$.
((*T–I*)+*C*+*Ш*)+*L*–3	Использование коэффициента *L* для изменения веса слов. По формуле: $\frac{\mathrm{n_i}}{\sum_k \mathrm{n_k}} \times \frac{\lvert \mathrm{D}\rvert}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert} + \frac{\mathrm{n_i}}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert} + \mathrm{C} + \mathrm{Ш}$.
((*T–I*)+*C*+*Ш*)+*L*–4	Использование коэффициента *L* для изменения веса слов. По формуле: $\frac{\mathrm{n_i}}{\sum_k \mathrm{n_k}} \times \frac{\lvert \mathrm{D}\rvert}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert} \times \frac{\mathrm{n_i}}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert} + \mathrm{C} + \mathrm{Ш}$.
(((*T–I*)+*C*+*Ш*)+*L*–3)+*Q*+*Z*–1	Использование коэффициентов *L*, *Q* и *Z* для изменения веса слов. По формуле: $\frac{\mathrm{n_i}+\mathrm{Q}+\mathrm{Z}}{\sum_k \mathrm{n_k}} \times \frac{\lvert \mathrm{D}\rvert}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert} + \frac{\mathrm{n_i}}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert} + \mathrm{C} + \mathrm{Ш}$.
(((*T–I*)+*C*+*Ш*)+*L*–3)+*Q*+*Z*–2	Использование коэффициентов *L*, *Q* и *Z* для изменения веса слов. По формуле: $\frac{\mathrm{n_i}}{\sum_k \mathrm{n_k}} \times \frac{\lvert \mathrm{D}\rvert}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert} + \frac{\mathrm{n_i}}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert} + \mathrm{C} + \mathrm{Ш} + \mathrm{Q} + \mathrm{Z}$.
(((*T–I*)+*C*+*Ш*)+*L*–3)+*Q*+*Z*–3	Использование коэффициентов *L*, *Q* и *Z* для изменения веса слов. По формуле: $\frac{\mathrm{n_i}+\mathrm{Q}+\mathrm{Z}}{\sum_k \mathrm{n_k}} \times \frac{\lvert \mathrm{D}\rvert}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert} + \frac{\mathrm{n_i}+\mathrm{Z}+\mathrm{Q}}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert} + \mathrm{C} + \mathrm{Ш}$.
(((*T–I*)+*C*+*Ш*)+*L*–3)+*Q*+*Z*–4	Использование коэффициентов *L*, *Q* и *Z* для изменения веса слов. По формуле: $\frac{\mathrm{n_i}+\mathrm{Q}+\mathrm{Z}}{\sum_k \mathrm{n_k}} \times \frac{\lvert \mathrm{D}\rvert}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert} + \frac{\mathrm{n_i}+\mathrm{Z}\times\mathrm{Q}}{\lvert(\mathrm{d_i}\supset \mathrm{w_i})\rvert} + \mathrm{C} + \mathrm{Ш}$.

Продолжение таблицы 2

Сокращение	Краткое описание
1	2
(((*T*–*I*)+*C*+*Ш*)+*L*–3)+*Q*–1	Использование коэффициентов *L* и *Q* для изменения веса слов. По формуле: $\frac{n_i}{\sum_k n_k} \times \frac{\lvert D\rvert}{\lvert(d_i \supset w_i)\rvert} + \frac{n_i}{\lvert(d_i \supset w_i)\rvert} + C + Ш + Q + Z.$
(((*T*–*I*)+*C*+*Ш*)+*L*–3)+*Q*–1	Использование коэффициентов *L* и *Q* для изменения веса слов. По формуле: $\frac{n_i+Q}{\sum_k n_k} \times \frac{\lvert D\rvert}{\lvert(d_i \supset w_i)\rvert} + \frac{n_i}{\lvert(d_i \supset w_i)\rvert} + C + Ш.$
(((*T*–*I*)+*C*+*Ш*)+*L*–3)+*Z*–1	Использование коэффициентов *L* и *Z* для изменения веса слов. По формуле: $\frac{n_i+Z}{\sum_k n_k} \times \frac{\lvert D\rvert}{\lvert(d_i \supset w_i)\rvert} + \frac{n_i}{\lvert(d_i \supset w_i)\rvert} + C + Ш.$
(((*T*–*I*)+*C*+*Ш*)+*L*–3)+*Z*–2	Использование коэффициентов *L* и *Z* для изменения веса слов. По формуле: $\frac{n_i}{\sum_k n_k} \times \frac{\lvert D\rvert}{\lvert(d_i \supset w_i)\rvert} + \frac{n_i}{\lvert(d_i \supset w_i)\rvert} + C + Ш + Z.$

Предложенная модернизация алгоритма (функции) классификации объединяет в себе методы *SVM* (поиск оптимальной гиперплоскости, отделяющей положительные примеры от отрицательных, гарантируя, что граница между ближайшими позитивами и негативами является максимальной) и дерево принятия решений (последовательное разбиение множества обучающих документов на классы до тех пор, пока в классе не останется документов только одной из категорий). Использование предложенной модификации–композиции повышает полноту и точность работы модели систематизации информации (таблица 3).

Таблица 3 – Композиции моделей классификации

Сокращение	Краткое описание
1	2
SVM	Использование метода *SVM*.
SVMK	Использование метода *SVM* с внесенными в него коррективами на коэффициент *k*.
SVM+NTREE	Использование метода *SVM* и *Decision tree* состоящего из *U* уровней.
SVMK+NTREE	Использование метода *SVM* с внесенными в него коррективами на коэффициент *k* и *Decision tree* состоящего из *U* уровней.

Разработанные композиции модели позволяют отобразить работу реальной системы управления и систематизации специализированной информацией.

Для улучшения характеристик модели, осуществляющей систематизацию информации, предложены и обоснованы композиции методов, включающие в себя:

- удаление шумовых слов (позволит уменьшить размер документов и повысить скорость работы);
- удаление окончаний, а также внесение изменений путем корректировки приоритетности слова в методе – *TF–IDF*, определяющий вес слова в тексте (позволяет более корректно распределить вес термов в документе, для повышения качества работы классификатора);
- модернизацию метода *SVM* при помощи дополнения его алгоритмом дерево принятия решений с *U* – уровнями (позволяет повысить полноту и точность работы классификатора).

Использование предложенных композиций приводит к дополнительному влиянию на свойства предложенной модели управления и систематизации информации, и, как следствие, к повышению качества работы систем систематизации и управления текстовой информацией.

Четвертая глава посвящена разработке модернизированной модели анализа и управления информацией, способствующей улучшению качества работы и оптимизации решения задач систематизации, управления и автоматической обработки специализированной информации.

Прежде чем начинать практическую разработку системы автоматического управления и систематизации текстовой информации, необходимо синтезировать ее общую архитектуру на основе предложенной модели. При построении архитектуры выбор был остановлен на методологии семейства *IDEF* (рисунок 4), которая эффективно отображает модули деятельности сложных систем. Архитектура системы имеет модульную структуру, которая позволяет системе быть открытой, гибкой, и дает значительные преимущества в ее поддержке.

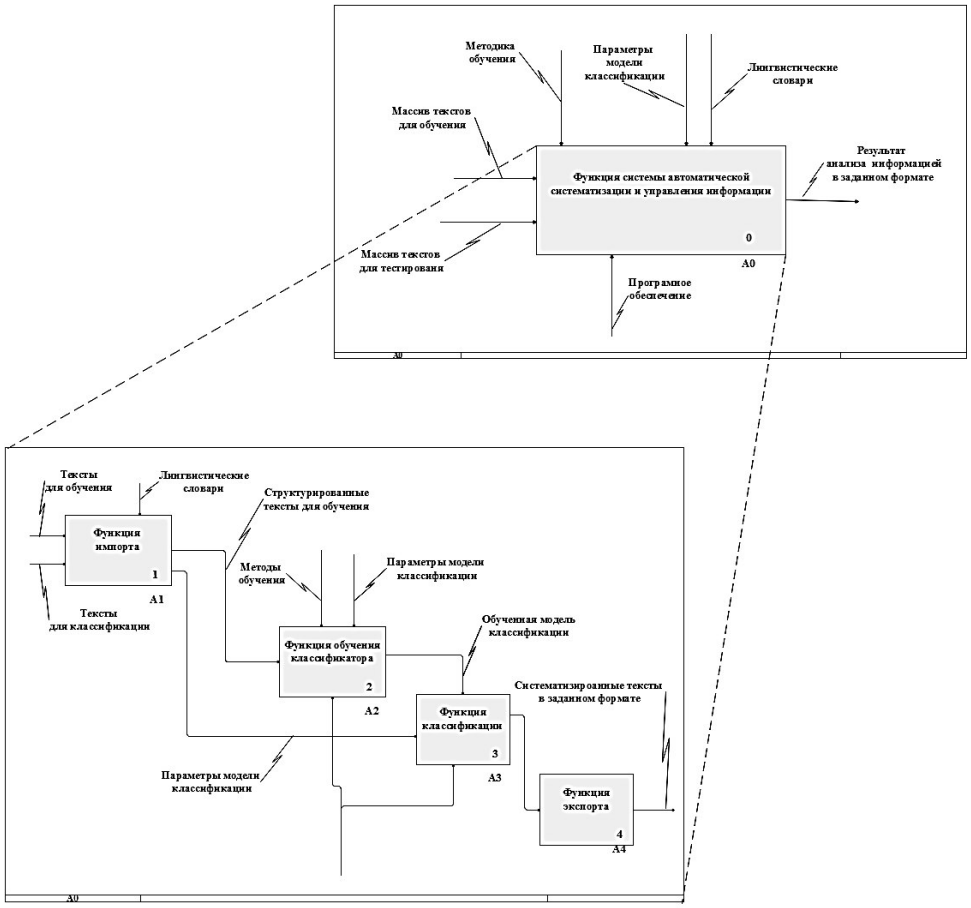


Рисунок 4 – Обобщенная модель автоматической систематизации и управления текстовой информацией

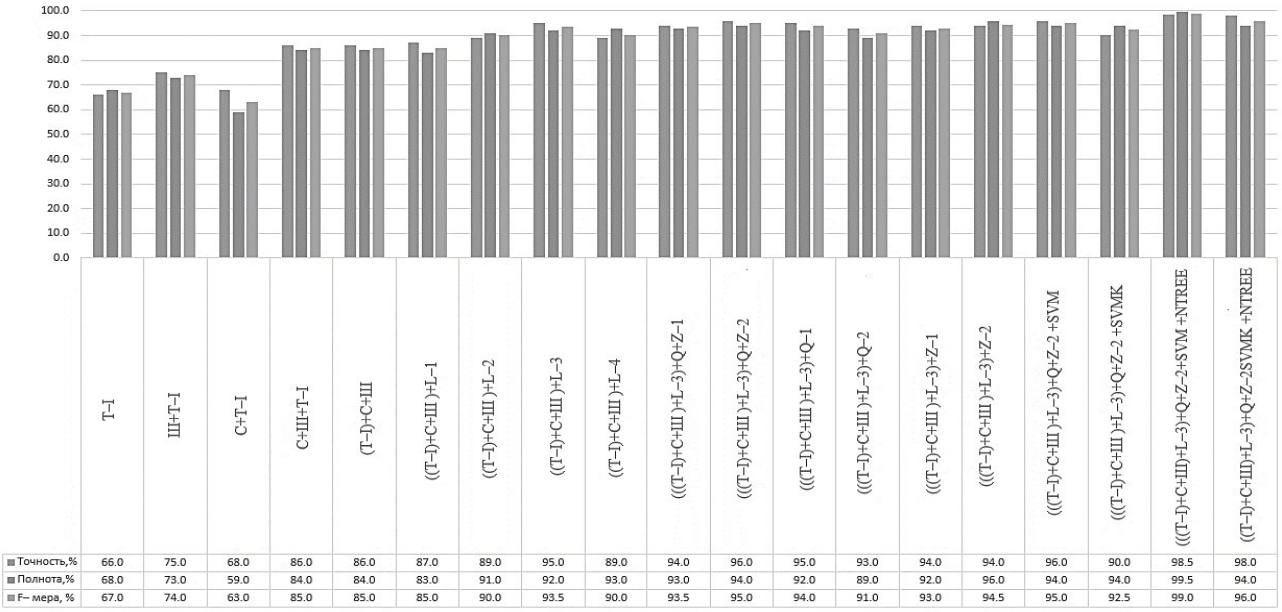
Для улучшения характеристик обобщенной модели, осуществляющей систематизацию информации, предложены и обоснованы модели композиций методов систематизации:

- удаление шумовых слов (позволяет уменьшить размер документов и повысит скорость работы);
- удаление окончаний, а также внесение изменений путем корректировки приоритетности слова в методе – *TF–IDF*, определяющий вес слова в тексте (позволяет более корректно распределить вес термов в документе, что в дальнейшем повысит качество работы классификатора);
- модернизацию метода *SVM* при помощи дополнения его алгоритмом дерево принятия решений с *U* – уровнями (позволяет повысить полноту и точность работы классификатора).

Использование предложенных композиций приводит к дополнительному влиянию на свойства предложенной модели управления и систематизации информации, и, как следствие, к повышению качества работы систем систематизации и управления текстовой информацией.

Рисунок 5 отображает качество работы различных разработанных моделей композиций систематизации информации, объединяющих в себе

предложенные методики и дополняющие работу друг друга с целью получения наилучшего качества при распределении данных по тематикам.



	Т–I	Ш+Т–I	С+Т–I	С+Ш+Т–I	(Т–Д)+С+Ш	((Т–Д)+С+Ш)+L–1	((Т–Д)+С+Ш)+L–2	((Т–Д)+С+Ш)+L–3	((Т–Д)+С+Ш)+L–4	(((Т–Д)+С+Ш)+L–3)+Q+Z–1	(((Т–Д)+С+Ш)+L–3)+Q+Z–2	(((Т–Д)+С+Ш)+L–3)+Q–1	(((Т–Д)+С+Ш)+L–3)+Q–2	(((Т–Д)+С+Ш)+L–3)+Z–1	(((Т–Д)+С+Ш)+L–3)+Z–2	(((Т–Д)+С+Ш)+L–3)+Q+Z–2 +SVM	(((Т–Д)+С+Ш)+L–3)+Q+Z–2 +SVMK	(((Т–Д)+С+Ш)+L–3)+Q–Z–2+SVM +NTREE	((Т–Д)+С+Ш)+L–3)+Q+Z–2SVMK +NTREE
Точность,%	66.0	75.0	68.0	86.0	86.0	87.0	89.0	95.0	89.0	94.0	96.0	95.0	93.0	94.0	94.0	96.0	90.0	98.5	98.0
Полнота,%	68.0	73.0	59.0	84.0	84.0	83.0	91.0	92.0	93.0	93.0	94.0	92.0	89.0	92.0	96.0	94.0	94.0	99.5	94.0
F– мера, %	67.0	74.0	63.0	85.0	85.0	85.0	90.0	93.5	90.0	93.5	95.0	94.0	91.0	93.0	94.5	95.0	92.5	99.0	96.0

Рисунок 5 - Итоговое качество предложенных композиций систематизации информации

На основе анализа итогового качества композиций была усовершенствована предложенная общая модель систематизации и управления информацией, вариантом с наилучшими показателями качества систематизации информации (2) (отображает 99% правильно отнесенных документов по тематикам в результате проведенных экспериментов):

$$(((T–I)+C+Ш)+L–3)+Q+Z–2 +SVM +NTREE, \quad (2)$$

где

T–I – использование метода TF – IDF,

C – удаление стемминга,

Ш – удаление шумовых слов,

L–3 – при распределении веса учитывает коэффициент *L*,

Q+*Z*–2 – при распределении веса учитывает коэффициенты *Q* и *Z*,

SVM+*NTREE* – объединяет метод *SVM* и *Decision tree* из *U* уровней.

Экспериментальное исследование предложенной модели автоматической систематизации и управления информацией, основанной на выполненном объединении достоинств синтагматических и парадигматических подходов, и всех предложенных модернизаций показало увеличение полноты и точности работы в среднем на 31,5% и 32,5% соответственно.

ЗАКЛЮЧЕНИЕ

Диссертация является законченной научно–исследовательской работой, в которой получено решение важной научно–технической задачи повышения эффективности управления и систематизации специализированной текстовой информации больших объёмов с учетом особенностей русскоязычных текстовых документов, реализованных с помощью модернизация моделей, методик, и алгоритмов системного анализа, искусственного интеллекта и компьютерной лингвистики.

Основные научные результаты и выводы, полученные при выполнении работы, состоят в следующем:

1. Управление и систематизация по содержанию большого количество текстов является сложной актуальной задачей. Ее выполнение ограниченным количеством специалистов и затрачиваемым временем в условиях постоянного поступления новой информации практически невозможно.
2. Анализ современных публикаций позволяет утверждать, что существует значительный разрыв между методами систематизации и управления информацией, основанными на машинном обучении, и методами, основанными на знаниях.
3. Выполнен анализ принципов построения систем управления, систематизации и обработки текстовой информации, изучены особенности их работы. Формализация единого универсального системного способа представления знаний позволяет создать соответствующие алгоритмы и инструментальные средства для обработки знаний различного типа единообразным способом и с помощью единого формального аппарата, построение которого осуществляется на основе методов и алгоритмов машинного обучения в рамках анализа специализированной текстовой информации.
4. Для решения задачи создания единых основ представления накопленных знаний и управления ими за счет интеграции и универсализации существующих способов систематизации таких знаний предлагается способ преобразования знаний, приведенных к единому виду, при помощи моделей в стандартах серии *IDEF*, выбор компонентов которой осуществляется среди возможных решений на основе специально разработанных приемов, методик и типовых моделей организации системы и принятия решений.
5. Показано, что разработка новой модели управления и систематизации специализированной текстовой информации, и модернизация применяемых в ней методик, методов и алгоритмов системного анализа, искусственного интеллекта и компьютерной лингвистики позволяет повысить эффективность управления, систематизации и обработки специализированной текстовой информации больших объёмов с учетом особенностей русскоязычных текстовых документов.
6. Обоснован выбор методов для построения модели модернизированной системы систематизации и управления текстовой

информацией. Для дальнейшего применения и модернизации в качестве базового выбран один из наиболее эффективных методов систематизации и управления информацией – *SVM*.

7. Разработана обобщенная модель управления и анализа информации, способствующая улучшению качества работы систем, решающих задачи этого класса, с целью оптимизации решения задачи управления и автоматической обработки специализированной информации. Благодаря применению разработанной модели появляется возможность оценить систему в состоянии равновесия и степень её чувствительности к различным факторам и внешним воздействиям, а также исследовать устойчивость поведения полученной модели в процессе принятия решений при обработке текстовой информации.

8. Для построения модернизированной модели управления и систематизации информации предложено использование оптимального способа анализа знаний, представляемого в виде стандарта *IDEF*. Формализация данных единым универсальным системным способом представления знаний позволяет создать соответствующие алгоритмы и подобрать инструментальные средства для обработки знаний с помощью единого формального аппарата.

9. Разработана обобщенная архитектура систематизации информации на основе предложенной модели, которая показывает общее представление ее построения и позволяет перейти к практической реализации ее прототипа и исследованию практической эффективности предложенных решений.

10. Для улучшения характеристик модели, осуществляющей систематизацию информации, предложены и обоснованы модели композиций методов систематизации, включающие в себя:

- удаление шумовых слов (позволяет уменьшить размер текстовых документов и повысит скорость работы);
- удаление окончаний, а также внесение изменений путем корректировки приоритетности слова в основном методе – *TF–IDF*, определяющий вес слова в тексте (позволяет более корректно распределить вес термов в документе, что в дальнейшем повысит качество работы классификатора);
- модернизацию метода *SVM* при помощи дополнения его алгоритмом дерево принятия решений с *U* – уровнями (позволяет повысить полноту и точность работы классификатора).

Использование предложенных композиций приводит к дополнительному влиянию на свойства предложенной модели управления и систематизации информации, и, как следствие, к повышению качества работы систем систематизации и управления текстовой информацией.

11. Экспериментальное исследование предложенных средств показало их позитивное влияние на усовершенствование предложенной общей модели систематизации и управления информацией на основе правил отбора неинформативных признаков, способов взвешивания термов и композиций методов систематизации.

12. Экспериментальное исследование моделей вычислительных композиций распределения веса слов в текстовом документе показало, что при разном распределении веса терма повышение качества систематизации информации варьируется в пределах 10% в зависимости от используемой композиции.

13. Экспериментальное исследование модели классификации текстовой информации на основе предложенных композиций методов систематизации и внесения изменений в структуру алгоритма её построения показало, что полнота и точность работы модели повышается на 5,5% и 8,5 % соответственно.

14. Экспериментальное исследование предложенной общей модели автоматической систематизации и управления информацией, основанной на выполненном объединении достоинств синтагматических и парадигматических подходов, и всех предложенных модернизаций показало увеличение полноты и точности работы в среднем на 32,5% и 31,5% соответственно.

15. Экспериментальное исследование предложенной общей модернизированной модели управления и систематизации текстовых документов на основе всех разработанных средств показало повышение скорости и качества обработки текстовой информацией в среднем более чем в 4 раза по сравнению с ручным способом (так на обработку одного тестового текстового документа вручную в среднем затрачивается 10 – 30 минут, в то время как автоматическая обработка на основе разработанных средств – в среднем затрачивает 3 – 7 минут).

16. Результаты исследований имеют широкий спектр применения для различных предметных областей. Предложенная практическая реализация усовершенствованной модели управления и систематизации информации позволяет формировать текстовые базы данных, содержащие классифицированную информацию, в автоматическом режиме. На основании результатов классификации появляется возможность автоматизировать работу специалистов–аналитиков, осуществляющих тематический анализ текстовой информации, и ведение аналитических задач в различных предметных областях, что может послужить функциональным дополнением и развитием информационных систем различных организаций.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ АВТОРОМ ПО ТЕМЕ ДИССЕРТАЦИИ.

– Публикации в рецензируемых изданиях ВАК Минобрнауки ДНР:

1. Бурлаева Е. И. Павлыш В. Н. Проект построения алгоритма классификации текстовых документов. / Проблемы искусственного интеллекта, №4 (7).– Донецк, 2017.– С 24–32.
2. Павлыш В. Н., Зори С. А., Бурлаева Е. И. Задача классификации информации при формировании баз данных в компьютерных

обучающих системах. / Проблемы искусственного интеллекта, №4 (11).– Донецк, 2018.– С 71–81.
3. Бурлаева Е. И., Зори С. А. Сравнение некоторых методов машинного обучения для анализа текстовых документов. /Проблемы искусственного интеллекта, №1 (12).– Донецк, 2019.– С 42–51.

– Публикации в рецензируемых изданиях ВАК Минобрнауки РФ:

4. Бурлаева Е. И. Обзор методов классификации текстовых документов на основе подхода машинного обучения. «Программная инженерия», том 8, №7.– Москва, 2017.– С 328–336.
5. Бурлаева Е. И., Павлыш В. Н. Анализ методов преобразования текстов в форму объектов векторного пространства. «Программная инженерия», том 10, №1.– Москва, 2019.– С 30–37.

– Публикации по материалам научных конференций:

6. Павлыш В. Н. Бурлаева К. И, Комбинированный подход к решению задач классификации текстовых массивов. Материалы V Международной научно–технической конференции «Современные информационные технологии в образовании и научных исследованиях» (СИТОНИ–2017).– Донецк: ДонНТУ, 2017. –442с.
7. Бурлаева Е.И., Ермоленко Т. В. Сопоставление методов автоматической обработки текста. Информатика, управляющие системы, математическое и компьютерное моделирование в рамках III форума «Информационные перспективы Донбасса» (ИУСМКМ – 2017): VIII Международная научно–техническая конференция, 25 мая 2017, г. Донецк: / Дон.нац. техн. ун–т; редкол. Ю.К. Орлов и др. – Донецк: ДонНТУ, 2017. – 802 с.
8. Павлыш В.Н., Бурлаева Е.И., Зори С.А Системный анализ и векторизация текстовой информации / Машиностроение и техно сфера XXI века // Сборник трудов XXV международной научно–технической конференции в г. Севастополе 10–16 сентября 2018 г. В 2–х томах. – Донецк: ДонНТУ, 2018. Т. 2. – с.32–37.
9. Бурлаева Е. И. Анализ работы классификаторов на русскоязычном массиве документов. Донбасс будущего глазами молодых ученых, г. Донецк, 20 ноября 2018 г. – Донецк: ДонНТУ, 2018. – 264 с.

АННОТАЦИЯ

Бурлаева Екатерина Игоревна. **Совершенствование методов системного анализа в задачах управления и систематизации специализированной информации.**– На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.01 – Системный анализ, управление и обработка информации (по отраслям) (технические науки) – ГОУВПО «ДОНЕЦКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ», Донецк, 2019 г.

Диссертация посвящена разработке модернизированной модели систематизации и управления информацией, представляющей собой эффективный инструмент осмысления информации и принятия адекватных решений по управлению ею, и имеющая широкий спектр применения для различных предметных областей.

В работе исследованы методы, модели и алгоритмы, используемые для решения задач систематизации и управления специализированной текстовой информацией. В результате для создания модернизированной модели систематизации информации, объединяющей в себе парадигматические и синтагматические подходы, появляется возможность эффективной автоматической систематизации информации с учетом особенностей лингвистических процессов естественного языка.

При использовании предложенных модернизаций произошло увеличение полноты и точности работы в среднем на 31,5% и 32,5% соответственно.

Ключевые слова: система, систематизация информации, управление информацией, модель, качество.

ANNOTATION

Burlaeva Ekaterina Igorevna. **Improving the methods of system analysis in management tasks and the systematization of specialized information**. – As a manuscript.

Thesis for the degree of Candidate of Technical Sciences in the specialty 05.13.01 – Systems analysis, management and information processing (by industry) (technical sciences) – GOUVPO «DONETSK NATIONAL TECHNICAL UNIVERSITY», Donetsk, 2019.

The thesis is devoted to the development of a modernized model of systematization and management of information, representing an effective tool for understanding information and making adequate decisions to manage it, which has a wide range of applications.

The paper studies the methods, models and algorithms used to solve problems of systematization and information management. When creating a modernized model of systematization of information that combines paradigmatic and syntagmatic approaches, it becomes possible to effectively automatically systematize information, taking into account the peculiarities of the linguistic processes of natural language.

Keywords: system, systematization of information, information management, model, quality.