Цель работы – разработка программного комплекса, позволяющего автоматизировать сбор и структурирование информации на естественном языке с тематических интернет-ресурсов – тематического агрегатора данных.

В результате проделанной работы была анализирована предметная область и современное состояние исследований в области агрегации данных и извлечения именованных сущностей, разработан программный комплекс – тематический агрегатор данных, библиотека контекстно-свободных грамматик, спроектирован пользовательский интерфейс, а также проведена апробация системы.

Эффективность разработанной системы заключается в автоматизации сбора структурированной информации с тематических интернет-ресурсов, обеспечения оптимальной организации процесса сбора данных, повышения удобства работы с информацией для пользователя.

Разработанный программный комплекс может быть применен для решения задачи автоматического извлечения и структурирования (агрегации) данных из документов на естественном языке.

# **ВВЕДЕНИЕ**

В настоящее время массивы информации, доступные человеку, многократно выросли благодаря развитию сети Интернет. Классификация/рубрикация информации (отнесение порции информации к одной или нескольким категориям из ограниченного множества) является традиционной задачей организации знаний и обмена информацией. В больших информационных коллекциях имеет смысл говорить только об автоматической рубрикации.

Поиск нужной информации среди большого объёма данных на множестве интернет-ресурсов становится все более сложной и затратной по времени задачей для человека. Информацию необходимо собирать, извлекать ее из множества неструктурированных источников близких между собой тематик, обрабатывать, структурировать, стремиться к повышению удобства и эффективности ее восприятия.

Решение этих задач возможно с применением обработки естественных языков и реализацией ее алгоритмов. Одна из наиболее важных задач обработки естественных языков, связанная с агрегацией данных и необходимая для ее реализации – это извлечение именованных сущностей.

***Цель работы -*** разработка программного комплекса, позволяющего автоматизировать сбор и структурирование информации на естественном языке с тематических интернет-ресурсов – тематического агрегатора данных.

Для достижения поставленной цели необходимо решить ***следующие задачи***:

1. Проанализировать современное состояние исследований в области агрегации данных и анализа естественных языков.
2. Провести анализ существующих научных и практических решений в выбранной области, изучить методы, принципы и технологии извлечения именованных сущностей, возможности их применения для агрегации данных.
3. Спроектировать компонентную реализацию программной системы для агрегации тематических данных.
4. Разработать программный комплекс – тематический агрегатор данных.
5. Апробировать разработанный программный комплекс на практике, сделать вывод об эффективности созданной системы агрегации данных.

***Решение поставленных задач определило структуру работы.*** В первой части проведено исследование ключевых аспектов применения автоматизированного анализа текстов, приводится обзор методов и алгоритмов извлечения информации из текстов на естественном языке. Во второй – предложена методика выделения фактов из текстов. В третьей главе представлена архитектура системы агрегации данных из текстов на естественном языке с применением контекстно-свободных грамматик.

В результате решением поставленных задач стали контекстно-свободные грамматики, реализующие задачи извлечения именованных сущностей (фактов и связанных с ними атрибутов) из текстов на естественном языке для различных тематик.

***Объект исследования*** – агрегация неструктурированных массивов информации

***Предмет исследования*** – методы и алгоритмы обработки естественных языков и извлечения именованных сущностей, применимые для агрегации данных.

***Методология и методы исследования.*** При выполнении работы использованы методы классификации текстов на естественном языке с использованием технологии машинного обучения, объектно-ориентированное программирование, системный анализ.

***Научная новизна и теоретическая значимость работы.*** Новизна данной работы заключается в следующем:

* Спроектирована новая компонентная реализация программной системы для агрегации тематических данных, *отличием* которой является наличие взаимодействующих модулей парсинга текста на естественном языке и работы с базой данных.
* Разработан программный комплекс, реализующий агрегацию данных из текстов на естественном языке на основе извлечения именованных сущностей с помощью контекстно-свободных грамматик.
* Проведено тестирование программного комплекса «Тематический агрегатор данных» на практических примерах.

***Практическая значимость.*** В ходе выполнения работы был создан программный комплекс – тематический агрегатор данных с применением алгоритмов анализа естественного языка, применимый для извлечения структурированной информации из текстов, была создана библиотека контекстно-свободных грамматик для извлечения разных типов фактов и работы с текстами разной тематики. Данный программный комплекс позволяет автоматически обрабатывать поступающие материалы по выбранным тематикам.

# **1 Анализ предметной области**

* 1. **Концепция машинного обучения**

Классификация/рубрикация информации (отнесение порции информации к одной или нескольким категориям из ограниченного множества) является традиционной задачей организации знаний и обмена информацией. В больших информационных коллекциях имеет смысл говорить только об автоматической рубрикации.

В настоящее время предложено много методов для решения данной задачи посредством автоматических процедур. Существующие методы можно разделить на два принципиально различных класса: методы машинного обучения и методы, основанные на знаниях (также иногда именуемые «инженерный подход»).

При применении методов машинного обучения для построения классификатора используется коллекция документов, предварительно отобранная человеком. Алгоритм машинного обучения строит процедуру классификации документов на основе автоматического анализа заданного множества текстов.

При использовании методов, основанных на знаниях, правила отнесения документа к той или иной рубрике задаются экспертами на основе анализа рубрикатора и, возможно, части текстов, подлежащих рубрицированию.

Любые методы автоматической классификации текстов в той или иной форме используют знания о свойствах текста на естественном языке и знания об особенностях текстов, принадлежащих той или иной рубрике.

Принципиальная разница между двумя группами методов состоит в том, что методы машинного обучения используют математические методы для извлечения знаний из обучающей коллекции текстов, в то время как "инженерный подход" использует знания эксперта о свойствах текстов, принадлежащих рубрикам. Знания эксперта основываются, в первую очередь, на предыдущем опыте, в частности, на большой коллекции прочитанных ранее текстов, и во вторую очередь, на части текстов, подлежащих рубрицированию.

В исследованиях, посвященных применению методов машинного обучения для классификации текстов, применяются универсальные алгоритмы, которые применимы для широкого круга задач анализа и обработки информации. Для задачи классификации текстов эти методы работают с абстрактной векторной моделью документа и не учитывают особенностей задачи тематической классификации текстов и структуры рубрикатора. Тем не менее, во многих случаях методы машинного обучения дают весьма высокие результаты. Качество рубрикации для систем, основанных на машинном обучении, является довольно высоким для небольших рубрикаторов, и сильно падает с увеличением количества рубрик и усложнением структуры рубрикатора.

Во многих случаях, даже при наличии заранее отобранной коллекции документов, методы машинного обучения неприменимы и используется значительно более трудоемкий инженерный подход. Инженерный подход обычно обеспечивает высокое качество рубрицирования и "прозрачность" алгоритма — результаты обработки легко.

**1.2 Концепция машинного обучения**

Машинное обучение ‒ это научное исследование алгоритмов и статистических моделей, которые компьютерные системы используют для эффективного выполнения конкретной задачи без использования явных инструкций, опираясь на шаблоны и выводы.

Абстрактный процесс машинного обучения начинается с подачи входных данных для алгоритма. Эти данные могут быть разными, начиная с более простых, такие как числовые или текстовые данные, заканчивая более сложными, например, видеофайл или изображение. Эти данные считываются электронно-вычислительной машиной и впоследствии обрабатываются различными математическими и статистическими алгоритмами. После обработки компьютер выводит эти же данные в виде информации, которые могут быть полезны для анализа, принятия решений, иллюстрации и т.д.

Машинное обучение лежит на стыке информатики, инженерии и статистики и часто появляется в других дисциплинах. Это инструмент, который может быть применен ко многим проблемам. Любая область, которая должна интерпретировать информацию и действовать на основе полученных данных, может извлечь выгоду из методов машинного обучения.

Алгоритмы машинного обучения строят математическую модель на основе выборочных данных, известных как «обучающие данные», чтобы делать прогнозы или решения без явного программирования для выполнения задачи.

## **1.3 Машинное обучение для анализа текстовых данных**

Обработка естественного языка (Natural Language Processing) ‒ интеллектуальный анализ данных и методы машинного обучения используются вместе для автоматической классификации и обнаружения шаблонов из электронных документов. Анализ текста состоит из нескольких задач, таких как правильная аннотация к документам, соответствующее представление документа, уменьшение размерности для обработки алгоритмических вопросов и соответствующая функция классификатора для получения обобщения.

Процесс предварительной обработки состоит в том, чтобы очистить границу каждой языковой структуры и максимально устранить языковые факторы, зависящие от языка.

Представление документов является одним из методов предварительной обработки, который используется для уменьшения сложности документов и облегчения их обработки, документ должен быть преобразован из полнотекстовой версии в вектор документа. Текстовое представление является важным аспектом в классификации или кластеризации документов.

Текстовый документ обычно представляется в виде вектора весов терминов (словарных признаков) из набора терминов (словаря), где каждый термин встречается хотя бы один раз в определенном минимальном количестве документов. Основной характеристикой проблемы классификации/кластеризации текста является чрезвычайно высокая размерность текстовых данных. Количество потенциальных возможностей часто превышает количество документов. Определение документа состоит в том, что он состоит из терминов, которые имеют различные шаблоны возникновения. Предварительная обработка включают в себя такие этапы как: извлечение признаков (характеристик) и выбор признаков.

Извлечение признаков является первым этапом предварительной обработки, который используется для представления текстовых документов в формате слов.

* **Токенизация** ‒ лексический (семантический) анализ текста, который находит минимальную единицу в тексте. Минимальная единица – это токен. В лексическом анализе токеном может быть одно слово, предложение или абзац.
* **Частота термина в документе** ‒ в семантическом анализе часто используют алгоритм счетчика слов, чтобы найти частоту каждого слова в данном тексте. Распределение частот слов является фундаментальным фенотипом языка.
* **Стеммитизация** – это преобразование морфологических форм слов в его корень, при помощи удаление окончание морфологических преобразований. Корень не обязательно должен быть существующим словом в словаре, но все его варианты должны соответствовать этой форме после завершения алгоритма.

После извлечения признаков важным шагом предварительной обработки текста является выбор признаков для построения векторного пространства, что повышает масштабируемость, эффективность и точность текстового классификации или кластеризации.

Основная идея выбора признака заключается в выборе подмножества объектов из исходных документов. Выбор признаков выполняется путем сохранения слов с наивысшим весом в соответствии с заданной мерой важности слова. Выбранные функции сохраняют первоначальный смысл и обеспечивают лучшее понимание данных и процесса анализа. Для классификации или кластеризации текста основной проблемой является высокая размерность пространства объектов

## **1.4 Анализ существующих программных средств для извлечения именованных сущностей из текстов на естественном языке**

В ходе изучения состояния исследований в области извлечения именованных сущностей мной были рассмотрены следующие программные средства:

* General Architecture for Text Engineering (GATE) – система, разработанная в Университете Шеффилда и предназначенная для обработки естественного языка. Система применима для решения задач выявления смыслового содержания текстов на естественном языке и его упорядочивания в структурированной форме с помощью аннотирования отдельных сегментом текста. GATE может использоваться для анализа кореферентности документов, семантической аннотации, работы с онтологиями, машинного обучения. Инструмент GATE, как и программные модули для него, написан на языке Java. Есть возможность обработки различных форматов текстовых документов: простой текстовый, HTML, DOC и др.. Для описания правил в системе GATE применяется язык Jape. [9] К минусам системы можно отнести архитектуру обработки естественных языков. На практике архитектура системы не подходит для обработки гибких языков наподобие русского, немецкого или латинского, так как GATE не имеет возможности обрабатывать морфологию. К плюсам системы можно отнести свободу, которую предлагает язык правил Jape, работу с огромным количеством входных форматов, даже таких обычно «проблемных» для парсинга, как pdf и docx, кроссплатформенность за счет платформы Java, на которой система была разработана. Также важным преимуществом является открытость кода системы GATE.
* Pullenti SDK. Pullenti SDK – комплекc средств разработки от компании «ООО Семантик», предназначенный для работы с неструктурированными данными. Pullenti SDK применим для решения задач анализа текста и извлечения именованных сущностей из неструктурированных русскоязычных текстов. Выделение именованных сущностей основано на правилах. Однако для некоторых типов сущностей можно подгружать внешние словари (онтологии), содержащие описания существующих сущностей, тогда система при выделении именованных сущностей будет пытаться привязываться к внешним сущностям, описанным в онтологиях. Плюсом системы является то, что она бесплатна для некоммерческого использования. [10]
* FreeLing – мультиязычная библиотека для обработки текстов. В инструменте содержатся два модуля для извлечения именованных сущностей – Basic-модуль и Bio-модуль. Первый, Bio-модуль, основан на машинном обучении. Он обеспечивает довольно высокую точность извлечения (свыше 90%), однако работает медленнее, чем Basic-модуль. Инструмент используется как библиотека, написанная на C/C++. В функционал, предоставляемый инструментом, входит разметка текста (токенизация), выделение предложений, орфологический анализ и т.д. Для работы FreeLing требуется наличие нескольких установленных внешних библиотек. FreeLing поддерживает русский и английский языки. Поддерживаемые платформы: GNU/Linux. [11]
* Stanford NER. Инструмент для обработки естественных языков, основанный на машинном обучении. Для извлечения именованных сущностей в инструменте есть три встроенных класса, основанных на регулярных выражениях, и один класс, предназначенный для разметки токенов через модель. Инструмент разработан на языке Java. Существуют также API для работы с инструментом, написанные и на других языках программирования: Python, Ruby, Perl, F#/ C#/ .NET. Поддерживает только английский язык. [12]
* ReVerb. Инструмент извлечения именованных сущностей, основанный на машинном обучении. На вход используемая в ReVerb нейросеть принимает простой текст, а результатом работы являются тройки вида: аргумент 1, связывающая фраза, аргумент 2. Инструмент разработан на языке Java. Приложение использует библиотеку анализа естественных языков OpenNLP для разметки текста по частям речи и выделения в нем именных групп. Поддерживает только английский язык. [13]

Таким образом, сравнительный анализ программных средств показывает, что каждое из них обладает недостатками. Потому, наиболее подходящим вариантом для решения задачи агрегации тематических данных в рамках данной работы является разработка собственного модульного программного комплекса, который позволит легко масштабировать и дорабатывать необходимый функционал.