**Министерство науки и высшего образования**

**Российской Федерации**

**ФЕДЕРАЛЬНОЕ Государственное Автономное**

**образовательное учреждение высшего образования**

**«новосибирский национальный исследовательский государственный университет»**

**Факультет информационных технологий**

Кафедра общей информатики

Направление подготовки 09.03.01 Информатика и вычислительная техника

Направленность (профиль) Программная инженерия и компьютерные науки

**ОТЧЕТ**

**о прохождении учебной практики (эксплуатационной практики)**

(указывается наименование практики)

**Обучающегося** Соломенникова Николая Александровича **группы №** 22204 **курса** 3

(Ф.И.О. полностью)

**Тема задания**: Исследование методов извлечения знаний о предметных областях для создания интеллектуальных помощников

**Место прохождения практики**: ФГБУН Институт математики им. С. Л. Соболева СО РАН, Лаборатория теории вычислимости и прикладной логики, 630090, Новосибирская обл., Новосибирск, пр. Академика Коптюга, 4

(полное наименование организации и структурного подразделения, индекс, адрес)

**Сроки прохождения практики:** с 06.02.2025 г. по 29.05.2025 г.

**Руководитель практики**   
**от профильной организации** Пальчунов Дмитрий Евгеньевич, в.н.с.

(Ф.И.О. полностью, должность)   (подпись)

**Руководитель практики от НГУ** Пальчунов Дмитрий Евгеньевич, зав. кафедрой

(Ф.И.О. полностью, должность)   (подпись)

**Руководитель ВКР** Пальчунов Дмитрий Евгеньевич зав. кафедрой

(Ф.И.О. полностью)   (должность)

**Оценка по итогам защиты отчета:**

(неудовлетворительно, удовлетворительно, хорошо, отлично)

**Отчет заслушан на заседании кафедры** общей информатики

(наименование кафедры)

**протокол** 12-25 **от** «3» декабря 2025 г.

Новосибирск 2024

**Содержание**

[Введение 3](#_Toc1055176473)

[1 Представление знаний 5](#_Toc1338136469)

[1.1 Интеллектуальные системы 5](#_Toc866885164)

[1.2 Понятие знаний в контексте ИИ 5](#_Toc1574401508)

[1.3 База знаний как центральный компонент 5](#_Toc382314715)

[1.4 Методы представления знаний 5](#_Toc543214346)

[1.5 Онтологии 7](#_Toc1070367916)

[2 Методы извлечения знаний 9](#_Toc1059530548)

[2.1 Методы извлечения знаний с участием человека 9](#_Toc851085731)

[2.1.1 Коммуникативные методы 9](#_Toc1761941480)

[2.1.2 Текстологические методы 10](#_Toc1129113134)

[2.2 Автоматические методы извлечения знаний 10](#_Toc1778106822)

[2.2.1 Методы для структурированных данных 10](#_Toc158507519)

[2.2.2 Методы для полуструктурированных данных 11](#_Toc391495644)

[2.2.3 Методы для неструктурированных данных 12](#_Toc1268133806)

[3 Точки роста в области извлечения знаний 13](#_Toc306063746)

[Заключение 14](#_Toc1211475799)

[Литература и интернет-ресурсы 15](#_Toc1541015496)

## **Введение**

**Введение в тему исследования**

Интеллектуальные помощники, такие как чат-боты, виртуальные ассистенты и экспертные системы, нуждаются в понимании контекста, терминологии и специфики предметной области, чтобы давать корректные ответы, предлагать решения и помогать пользователям. Без качественно организованных знаний помощник будет давать общие, неточные или бессмысленные ответы.

**Актуальность**

1. *Рост числа интеллектуальных помощников*:

Интеллектуальные помощники становятся неотъемлемой частью многих сфер жизни: от бытовых задач до сложных профессиональных областей, таких как медицина, финансы, и юриспруденция.

1. *Сложность и объем данных*:

Современный мир генерирует огромные объемы информации, большая часть которой не структурирована (статьи, форумы, веб-страницы). Преобразование этой информации в структурированную форму требует эффективных методов извлечения знаний, которые позволят интеллектуальным системам работать с ней.

1. *Узкоспециализированные приложения*:

Пользователи всё чаще ожидают, что помощники будут обладать знаниями в узких областях, например, в медицинской диагностике или консультировании по юридическим вопросам. Создание специализированных помощников требует построения предметно-ориентированных баз знаний.

1. *Автоматизация процессов*:

В условиях ускоряющейся цифровизации бизнесов и процессов извлечение знаний позволяет автоматизировать рутинные задачи, такие как поиск информации, анализ данных и генерация отчетов.

1. *Проблема качества взаимодействия*:

Качество работы помощников напрямую зависит от того, насколько хорошо они понимают контекст и особенности предметной области. Извлечение знаний надлежащими методами позволяет повысить удобство взаимодействия с пользователем, точность ответов и/или их релевантность.

**Цели**

Проанализировать и систематизировать известные подходы к извлечению знаний о предметных областях для создания интеллектуальных помощников, выявить основные методы, их преимущества и ограничения, а также определить направления развития технологий в данной области.

**Задачи**

1. Выявить основные концепции и теоретические основы, лежащие в основе технологий извлечения знаний.
2. Систематизировать существующих методы представления знаний.
3. Проанализировать современные научные и практические источники, посвящённые извлечению знаний и их применению в интеллектуальных системах.
4. Систематизировать существующие подходы и технологии, обеспечивающие создание баз знаний.
5. Обозначить направления, требующие дальнейшего изучения.

**Место прохождения практики**

Практика проходилась в ФГБУН Институт математики им. С. Л. Соболева СО РАН, Лаборатория теории вычислимости и прикладной логики, 630090, Новосибирская обл., Новосибирск, пр. Академика Коптюга, 4.

Ожидаемыми результатами прохождения практики являются получение первичных навыков выполнения научно-исследовательской работы, изучение методов представления и извлечения знаний о предметных областях, а также составление отчёта по практике.

## **1 Представление знаний**

Разработка интеллектуальных систем (в частности, интеллектуальных помощников) невозможна без чёткого понимания роли и структуры знаний. Поскольку данная работа посвящена исследованию методов извлечения знаний для таких систем, необходимо сначала определить, что такое знания в контексте искусственного интеллекта, как они соотносятся с когнитивными моделями человека и каким образом представлены внутри интеллектуальных систем.

### **1.1 Интеллектуальные системы**

*Интеллектуальная система* — это система, основанная на методах искусственного интеллекта (ИИ), способная к восприятию, интерпретации, принятию решений и адаптации на основе поступающей информации. Ключевая особенность интеллектуальных систем заключается в имитации когнитивных способностей человека — таких как анализ, логическое мышление, обучение и обобщение.

*Интеллектуальный помощник* является частным случаем интеллектуальной системы, ориентированным преимущественно на взаимодействие с человеком. Основная цель интеллектуального помощника — оказание поддержки в решении задач, предоставление информации или рекомендаций на основе знаний о предметной области. Независимо от архитектурных различий, все интеллектуальные помощники используют знания в качестве основы своего функционирования.

### **1.2 Понятие знаний в контексте ИИ**

С позиции когнитивной модели человека, знания — это результат осмысления, обобщения опыта и обработки информации, позволяющий формировать представления о мире, делать выводы и принимать решения. Эти знания включают как декларативную информацию (факты, утверждения), так и процедурную (навыки, правила действий).

В контексте интеллектуальных систем знания рассматриваются как формализованные структуры, которые позволяют машине интерпретировать входную информацию, осуществлять логические операции, адаптироваться и обеспечивать осмысленное поведение. Таким образом, формальные знания в ИИ являются отражением когнитивных знаний человека, перенесённых в машиночитаемую форму. Это обеспечивает основу для построения reasoning-процессов, автоматического вывода и генерации решений.

### **1.3 База знаний как центральный компонент**

*База знаний* — это организованное хранилище структурированных знаний, на основе которых интеллектуальная система формирует поведение, строит рассуждения и взаимодействует с пользователем. В типичной архитектуре интеллектуальной системы база знаний может включать:

* Факты о предметной области;
* Правила вывода и логики принятия решений;
* Онтологические структуры и терминологические связи;
* Векторные или нейросетевые представления смыслов (в гибридных системах).

Наличие базы знаний позволяет системе не только интерпретировать ввод пользователя, но и строить выводы, проверять гипотезы, формулировать объяснения и адаптироваться к изменяющимся условиям. Без базы знаний система была бы ограничена обученной моделью и не могла бы адекватно реагировать в новых или нестандартных ситуациях.

### **1.4 Методы представления знаний**

Разнообразие методов представления знаний обусловлено необходимостью решать фундаментальную проблему выбора способа описания знаний — удобного и осмысленного для человека и одновременно эффективного для обработки на ЭВМ. Каждый метод адаптирован к определённым типам знаний (например, декларативным или сенсомоторным), уровню интерпретируемости (для человека или машины) и специфике предметной области.

В данной работе представлена концептуальная карта, отражающая классификацию ключевых методов представления знаний. Несмотря на множество существующих подходов, карта фокусируется на базовых, фундаментальных методах, из которых прочие производны — в виде гибридов, расширений или частных случаев.

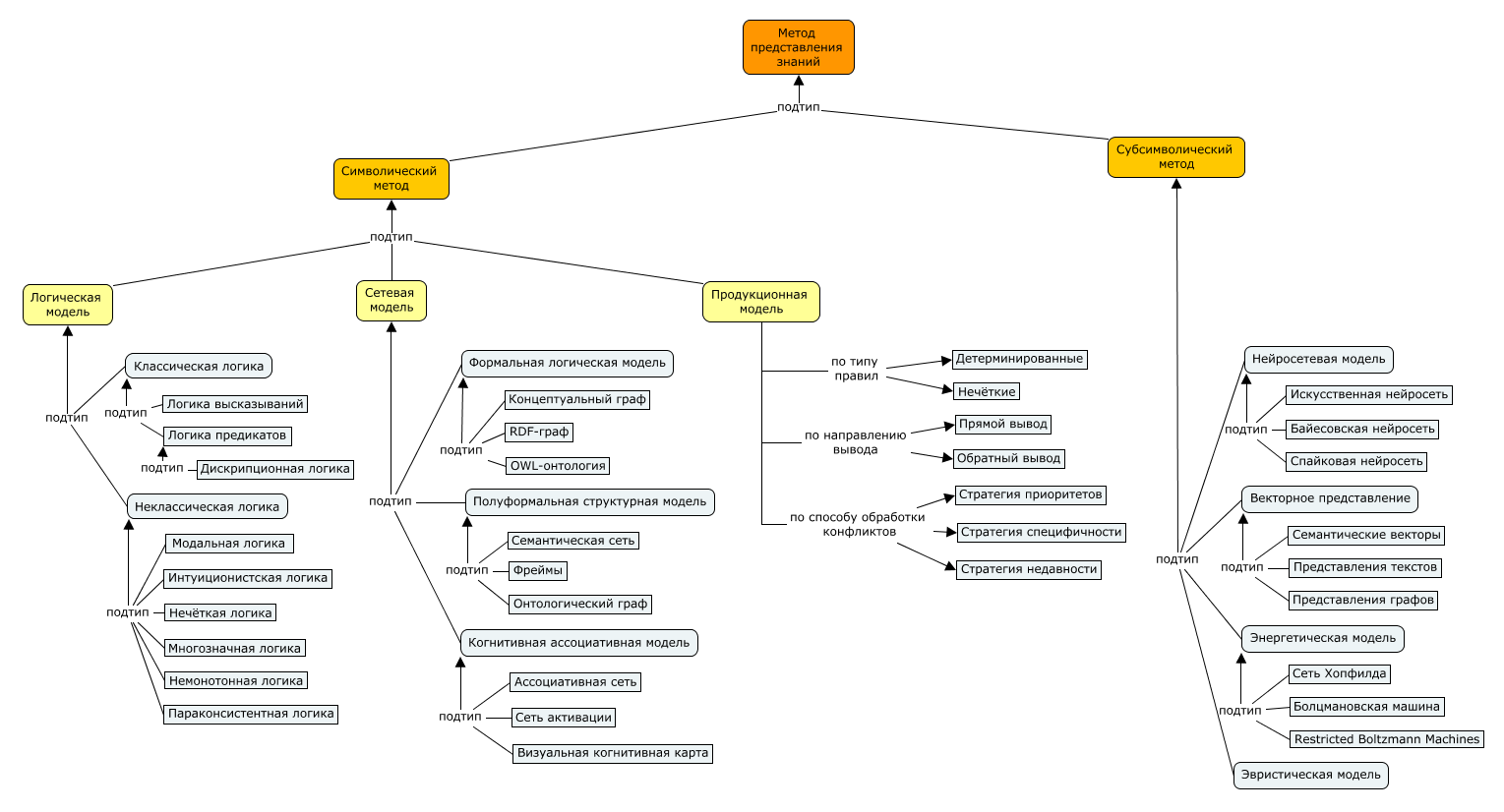


Рисунок 1.1 Концептуальная карта методов представления знаний

Методы представления знаний делятся на символические и субсимволические. Это разделение связано с двумя подходами к моделированию интеллекта:

* *Символический подход:* этот подход рассматривает интеллект как манипуляции с символами и понятиями. В рамках символического подхода строятся формальные модели знаний, представляющие информацию в виде логических выражений, правил и онтологий, а также соответствующие механизмы рассуждений и вывода. Такой подход обеспечивает высокую интерпретируемость и объяснимость знаний и выводов.
* *Нейросетевой (нейрокибернетический) подход:* противоположен символическому и ориентирован на воспроизведение структуры и принципов работы человеческого мозга. Основной акцент делается на создании программно-аппаратных моделей, имитирующих нейроны и их соединения — нейронных сетей. Эти системы хранят знания в виде распределённых числовых параметров (весов) и учатся выявлять закономерности из данных. Подход характеризуется высокой адаптивностью и способностью к обобщению, однако внутренние представления плохо интерпретируются человеком.

**Символические методы**

Символические методы представляют знания в виде чётко заданных структур: понятий, отношений, логических формул и правил. Эти методы обладают высокой интерпретируемостью и объяснимостью, поскольку напрямую связаны с человеческими понятиями и языком рассуждений. В рамках символического подхода выделяются три основные группы моделей:

1. *Логические модели* — опираются на формальные логические системы (логика высказываний, предикатов, модальные логики и др.) и обеспечивают возможность строгого логического вывода. Применяются в онтологиях, экспертных системах, системах формальной верификации и служат основой для формализации знаний.
2. *Сетевые модели* — представляют знания в форме графов, где узлы обозначают объекты или понятия, а рёбра — отношения между ними. Сюда относятся семантические сети, концептуальные графы, фреймы, когнитивные карты, OWL-онтологии и другие модели.
3. *Продукционные модели* — основаны на правилах вида "если–то", с рабочей памятью и механизмом логического вывода. Применяются в системах с реактивным поведением, экспертных системах и когнитивных архитектурах.

Замечание: несмотря на наличие методов автоматического извлечения символических знаний (например, извлечение триплетов или rule mining), интерпретируемость и корректность таких знаний по-прежнему в значительной степени зависят от участия человека. Только человек может гарантированно создать осмысленные правила, потому что он понимает контекст, цели, и значения символов.

**Субсимволические методы**

Субсимволические методы представляют знания в неявной форме — как параметры, веса и активации в распределённых моделях. Они не обладают явной логической структурой и, как правило, не интерпретируемы человеком, но демонстрируют высокую способность к обобщению и адаптации. Основные типы субсимволических моделей:

1. *Нейросетевые модели* — глубокие и поверхностные нейронные сети (MLP, CNN, RNN, Transformers), байесовские и спайковые сети.
2. *Векторные представления* — эмбеддинги слов и текстов.
3. *Энергетические модели* — сети Хопфилда, ограниченные Болцмановские машины.
4. *Эвристические модели* — приближённые алгоритмы, использующие эвристики и эмпирические оценки.

Субсимволические знания формируются автоматически на основе данных и являются основой современных обучаемых ИИ-систем.

**Гибридные методы**

Гибридные методы сочетают символические и субсимволические подходы, объединяя объяснимость формальных моделей с адаптивностью нейросетевых представлений. В таких системах символические структуры (правила, онтологии) дополняются распределёнными векторами и нейросетями.

Пример — нейросимволическое обучение (neuro-symbolic AI), где логика интегрируется с глубоким обучением для улучшения качества вывода и интерпретируемости. Такие методы позволяют использовать сильные стороны обоих подходов.

### **1.5 Онтологии**

*Онтология* — это формализованная, явная спецификация концептуализации предметной области, реализованная на основе логического формализма, поддерживающая вывод и верификацию знаний. Проще говоря, онтология задаёт формальную модель понятий (классов), их отношений и ограничений, а также правила логического рассуждения над ними. Онтология де-факто является самой распространённой в интеллектуальных системах моделью представления знаний.

Онтология оп ределяет:

* Что существует в предметной области (классы, концепты);
* Как объекты связаны между собой (свойства, отношения);
* Какие существуют ограничения на свойства и иерархии;
* Как можно логически рассуждать о данных понятиях и объектах.

Онтология включает четыре компонента:

1. Классы — абстрактные категории объектов (например, Человек, Организация, Студент). Классы образуют иерархии с подклассами и могут иметь ограничения и эквивалентности.
2. Индивидуумы — конкретные объекты, относящиеся к классам (например, Иван, МГУ).
3. Свойства — описывают отношения между индивидуумами или связи индивидуумов с данными:
   1. Отношения между индивидуумами (например, учитсяВ, являетсяРодителем).
   2. Отношения между индивидуумом и значением (например, имеетВозраст, имеетИмя).
4. Аксиомы — формальные утверждения, задающие структуру и логику онтологии:
   1. TBox — описывает классы и их иерархии.
   2. RBox — описывает свойства и ограничения на них.
   3. ABox — содержит утверждения о конкретных индивидуумах.

Онтологии служат для:

* Создания единого словаря понятий для людей и машин;
* Обеспечения интероперабельности между системами;
* Поддержки логического вывода, проверки и валидации знаний;
* Повышения переносимости знаний между разными приложениями и платформами;
* Стандартизации знаний для Semantic Web и интеллектуальных систем.

Классификация онтологий:

* По логическому ядру:
  + Логика описаний (DL), например OWL — поддерживает вывод.
  + Первая порядок логики (FOL) — более выразительна, но вычислительно сложнее.
  + Без формализма — только визуальные или понятийные связи.
* По уровню абстракции:
  + Онтологии верхнего уровня — общие понятия (DOLCE, SUMO).
  + Онтологии домена — специализированные (медицина, право).
  + Онтологии задачи — описывают процессы (диагностика, планирование).
  + Онтологии приложений — специфичные для конкретных систем.
* По реализации:
  + RDF/OWL — стандарты Semantic Web.
  + XML, графовые БД (Neo4j), таблицы и словари.

Преимущества онтологий:

* Стандартизированы и поддерживаются W3C.
* Обеспечивают формальную и читаемую как человеком, так и машиной структуру знаний.
* Позволяют выполнять логический вывод и проверку консистентности.
* Широко поддерживаются инструментами (Protégé, GraphDB и др.).
* Устойчивы к изменениям и расширяемы.

Таблица 1.1 Онтологии в сравнении с другими методами представления знаний

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Метод | Поддержка вывода | Интерпретируемость | Стандартизация | Универсальность |
| Онтологии (OWL) | Полная | Высокая | Да (W3C) | Максимальная |
| Продукционные системы | Частичная | Высокая | Нет | Ограниченная |
| Фреймы | Нет | Высокая | Нет | Ограниченная |
| Семантические сети | Нет | Высокая | Частично | Ограниченная |
| Нечёткие системы | Частичная | Низкая | Нет | Ограниченная |

## **2 Методы извлечения знаний**

Для создания интеллектуальных систем необходимо наполнить их базу знаний. Процесс извлечения знаний — это ключевой этап, который включает в себя сбор, анализ и преобразование информации в структурированный вид, пригодный для работы системы. В настоящее время большинство разработчиков интеллектуальных систем отмечают, что процесс извлечения знаний остается самым «узким» местом при построении промышленных экспертных систем.

Существует множество методов извлечения знаний, которые можно условно разделить на две категории: *с участием человека* (например, экспертов или пользователей) и *автоматические* (основанные на алгоритмах обработки данных). Эти подходы часто сочетаются, чтобы обеспечить высокую точность и полноту знаний.

### **2.1 Методы извлечения знаний с участием человека**

Методы с участием человека требуют человеческого вмешательства на различных этапах извлечения знаний. Эти методы трудозатратны, занимают много времени и склонны к потере информации по мере её прохождения между разными этапами, но тем не менее они полезны в случаях, когда необходимо работать со сложными и неструктурированными данными, где автоматические методы могут быть неэффективными. К тому же они незаменимы при извлечении знаний, в которых критически важна точность и семантика (например, в медицине). Данные методы можно разделить на *коммуникативные* и *текстологические*.

#### **2.1.1 Коммуникативные методы**

Коммуникативные методы основаны на взаимодействии аналитиков (инженеров по знаниям, которые специализируются на заполнении баз знаний) с экспертами в предметной области. Также зачастую привлекаются обычные пользователи для тестирования полученной базы знаний. Данные методы делятся на: *пассивные* (эксперт не вовлекается напрямую в процесс передачи знаний) и *активные* (предполагают активное взаимодействие между аналитиком и экспертом для получения знаний). Активные методы можно дополнительно разделить на *индивидуальные* (когда задействуется один эксперт) и *групповые* (задействуется одновременно несколько экспертов)*.*

* **Пассивные методы**:
  + *Наблюдение* — аналитик наблюдает за работой эксперта.
  + *Протокол мыслей вслух* — аналитик пассивно слушает эксперта.
  + *Лекции* — эксперт делится своими знаниями в формате лекции.
* **Индивидуальные активные методы**:
  + *Интервью* —эксперт и аналитик общаются один на один.
  + *Анкетирование* — эксперт отвечает на вопросы письменно, за счёт чего ответы даются в более сжатой и лаконичной форме.
  + *Экспертные игры* — основаны на моделировании ситуаций или задач в игровой форме.
* **Групповые активные методы**:
  + *Мозговой штурм* — метод генерации идей, при котором группа экспертов предлагает как можно больше решений задачи без критики, а затем выбирает наиболее подходящие.
  + *Круглый стол* — метод обсуждения, где эксперты обмениваются мнениями и анализируют проблему, стремясь прийти к консенсусу или выявить разные точки зрения.
  + *Ролевые игры* — метод моделирования ситуаций, где эксперты принимают на себя роли участников системы, чтобы изучить её поведение и найти оптимальные решения.
  + *Краудсорсинг* — использование широкой аудитории людей для сбора и классификации данных. Может включать в себя анкетирование, голосования или редактирование данных. Например, в WikiData пользователи могут вносить информацию о фактах, связывая их с соответствующими понятиями.

#### **2.1.2 Текстологические методы**

Текстологические методы направлены на систематический анализ текстовых источников с целью выявления и формализации знаний. Эти методы применяются для работы с существующими документами и позволяют извлекать знания, которые трудно получить от экспертов или автоматическими средствами. Часто предполагают участие экспертов для интерпретации извлечённой информации. Требуют глубокого понимания контекста и терминологии.

* *Анализ учебников* — изучение учебных пособий для выделения базовых и систематизированных знаний по предметной области.
* *Анализ литературы* — исследование научных статей, монографий и других материалов для выявления новых знаний, закономерностей и подходов.
* *Анализ документов* — обработка нормативных актов, инструкций, отчётов и других документов для извлечения формализованных знаний, правил и процессов.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок 2.1 Классификация методов извлечения знаний с участием человека

### **2.2 Автоматические методы извлечения знаний**

Автоматические методы извлечения знаний предполагают минимальное вмешательство человека и основаны на использовании алгоритмов и программных средств для обработки больших объёмов данных. Эти методы особенно эффективны при работе со структурированными и полуструктурированными данными, а также в случае необходимости быстрой обработки больших текстовых корпусов. Автоматические методы позволяют значительно ускорить процесс извлечения знаний и уменьшить трудозатраты, однако они требуют качественных исходных данных и могут быть менее точными при работе с неструктурированной информацией.

Автоматические методы можно классифицировать в зависимости от степени структурированности обрабатываемых данных. Данные могут быть *структурированными*, *полуструктурированными* или *неструктурированными*. Каждый тип требует своих подходов и методов обработки.

#### **2.2.1 Методы для структурированных данных**

*Структурированные данные* — это данные, организованные в виде фиксированных схем и таблиц (например, базы данных, таблицы, CSV-файлы). Методы для их обработки обычно используют строгие алгоритмы и математические модели. Ниже приведена классификация основных автоматических методов извлечения знаний из таких данных для создания интеллектуальных систем.

* **Методы извлечения ассоциативных правил и закономерностей**
  + Частые множества*:* Поиск комбинаций атрибутов или значений, встречающихся с частотой выше заданного порога.
  + *Правила ассоциаций:* Формирование правил вида «Если A, то B» с оценками поддержки (support) и доверия (confidence).
  + *Последовательные паттерны:* Извлечение частых упорядоченных последовательностей элементов.
  + *Временные и пространственные ассоциации:* Расширение правил с учётом временного и пространственного контекста.
* **Методы обучения правил и логических моделей**
  + *Обучение правил:* Автоматическая генерация интерпретируемых правил классификации и регрессии (например, CN2, RIPPER, FOIL).
  + *Индуктивное логическое программирование (ILP):* Обучение логических правил первого порядка с использованием фоновых знаний.
  + *Обучение с объяснениями:* Извлечение правил с акцентом на прозрачность и человекочитаемость.
* **Методы анализа структурированных графовых данных**
  + *Извлечение подграфов и шаблонов:* Поиск часто встречающихся подструктур в графах знаний.
  + *Семантический анализ и аннотирование графов:* Автоматическое обогащение графовых данных семантической информацией.
* **Методы формального анализа и концептуализации**
  + *Формальный концептуальный анализ (FCA):* Выделение концептов — пар объектов и общих атрибутов, построение концептуальных решёток.
  + *Построение иерархий и онтологий:* Автоматическое создание таксономий и онтологических структур.
* **Методы генерации гипотез и индуктивного вывода:** Автоматическое формирование новых предположений и логических выводов на основе данных.

#### **2.2.2 Методы для полуструктурированных данных**

*Полуструктурированные данные* — это данные, которые не имеют жёсткой структуры, но содержат определённые теги или маркеры (например, XML-документы, JSON, HTML-страницы). Методы извлечения знаний учитывают гибкую структуру и часто комбинируют синтаксический анализ, семантические технологии и машинное обучение.

* **Методы парсинга и структурного анализа**
  + Синтаксический разбор XML, JSON, HTML с использованием грамматик, XPath, XQuery.
  + Извлечение элементов, атрибутов, иерархий и связей.
* **Методы извлечения сущностей и отношений (Information Extraction, IE)**
  + Автоматическое выделение сущностей и их атрибутов из тегированных или частично тегированных данных.
  + Правила, шаблоны, статистические и машинно-обучаемые методы для извлечения семантических связей.
  + Обычно работают на основе структурированной информации в рамках документа.
* **Веб-скрейпинг и предварительная обработка**
  + Автоматический сбор и структурирование данных с веб-страниц и API.
  + Парсинг HTML, очистка и преобразование в полуструктурированный формат для дальнейшего анализа.
* **Методы нормализации и связывания сущностей**
  + Разрешение неоднозначностей, объединение информации об одних и тех же объектах из разных источников (Entity Resolution, Record Linkage).
  + Стандартизация и согласование данных.
* **Построение онтологий и схем**
  + Извлечение концептов и отношений из структур и метаданных.
  + Автоматическое построение иерархий, таксономий и онтологий.
* **Семантическое аннотирование и использование семантических веб-технологий**
  + Присвоение семантических меток (RDF-теги, OWL-онтологии) для улучшения машинной интерпретируемости.
  + Использование SPARQL для запросов и интеграции знаний.
* **Гибридные методы и языки запросов**
  + Комбинация правил, шаблонов и машинного обучения.
  + Применение языков XPath, XQuery, SPARQL для извлечения и агрегации знаний.

#### **2.2.3 Методы для неструктурированных данных**

*Неструктурированные данные* — это данные, не имеющие чёткой схемы или организации (например, тексты, изображения, аудио и видео). Методы извлечения знаний из таких данных основаны преимущественно на обработке естественного языка, анализе мультимодальных сигналов и машинном обучении. Основная сложность — неопределённость, многозначность и высокая вариативность данных. Зачастую требуется требуют предварительная обработка.

* **Методы обработки естественного языка (NLP)**
  + *Извлечение сущностей:* Автоматическое выделение из текста имен собственных (имена, организации, даты).
  + *Извлечение отношений и событий:* Определение семантических связей между сущностями и событий из текста.
  + *Синтаксический и семантический анализ:* Построение деревьев зависимостей, разбор структуры предложений, выделение смысловых ролей.
  + *Текстовая категоризация и тематическое моделирование:* Автоматическое присвоение текстам категорий, выделение скрытых тем (например, LDA).
  + *Кластеризация и обобщение текста:* Группировка документов по смыслу, автоматическое резюмирование.
* **Методы извлечения знаний из изображений и видео**
  + Выделение объектов, классификация сцен, распознавание действий.
  + Использование сверточных нейросетей (CNN), видеомоделей и мультимодальных подходов.
* **Методы анализа аудио и речи**
  + Распознавание речи (ASR), выделение ключевых фрагментов, анализ эмоций и интонаций.
* **Интеграция и мультимодальные методы**
  + Объединение текстовой, визуальной и аудиоинформации для комплексного извлечения знаний.
* **Методы обучения на основе представлений**
  + Использование эмбеддингов слов, предложений и документов (Word2Vec, BERT, GPT).
  + Извлечение скрытых закономерностей и знаний из распределённых представлений.
* **Гибридные методы и семантический веб**
  + Комбинация NLP, онтологий и правил для создания объяснимых моделей.

## **3 Точки роста в области извлечения знаний**

1. **Автоматизация и масштабируемость**

Методы автоматического извлечения знаний всё ещё несовершенны. Развитие алгоритмов, которые могут извлекать знания из больших массивов данных с минимальным вмешательством человека, является важным направлением.

1. **Улучшение качества обработки естественного языка**

Хотя современные модели NLP, такие как GPT или BERT, добились значительного прогресса, они не всегда точно извлекают сложные отношения или контекст. Требуются более точные методы анализа языка и разрешения неоднозначностей.

1. **Создание многоязычных и культурно-чувствительных помощников**

Многие интеллектуальные помощники работают преимущественно на английском языке. Развитие инструментов для извлечения знаний на других языках и с учетом культурных особенностей остаётся актуальной задачей.

1. **Интеграция с динамическими данными**

Современные системы должны уметь работать с постоянно обновляющимися данными, такими как новости, результаты исследований и изменения в законодательстве. Это требует разработки методов для обновления знаний в реальном времени.

1. **Гибридные подходы**

Комбинация автоматических методов извлечения знаний и экспертного подхода (ручная разметка) может повысить точность систем. Разработка таких гибридных методов может стать значимым шагом в этой области.

1. **Семантические технологии и онтологии**

Развитие методов создания семантических сетей и онтологий, которые обеспечивают глубокое понимание предметной области, является важным направлением для повышения эффективности интеллектуальных систем.

1. **Этические и правовые аспекты**

Работа с данными из различных источников требует учета вопросов приватности, авторского права и достоверности информации. Создание методов, обеспечивающих этичное извлечение знаний, становится всё более важным.

## **Заключение**

В рамках выполнения научно-исследовательской работы была достигнута поставленная цель — проведён анализ и систематизация современных подходов к извлечению знаний о предметных областях для создания интеллектуальных систем (в частности, интеллектуальных помощников). Также были выявлены основные методы и обозначены перспективные направления развития технологий извлечения знаний. В результате анализа современных научных и практических источников были изучены ключевые публикации и работы, посвящённые извлечению знаний. Было произведено обширное исследование методов представления знаний и дальнейшая их классификация. Проведена систематизация подходов к созданию баз знаний, включающая методы извлечения знаний с участием человека (коммуникативные и текстологические) и автоматические методы. Представлена классификация методов по степени структурированности данных, что позволило выделить их сильные и слабые стороны. Обозначены перспективы дальнейшего изучения.

Во время выполнения работы была приобретена компетенция анализа научной литературы и были улучшены навыки написания официальных документов. А также были приобретены знания о предметной области. Учитывая достигнутые результаты, я оцениваю выполнение работы положительно и считаю, что поставленные задачи были успешно решены.

## **Литература и интернет-ресурсы**

1. Гаврилова Т. А., Кудрявцев Д. В., Муромцев Д. И. “Инженерия знаний. Модели и методы”. [Электронный ресурс]. URL: <https://vk.com/wall-204281522_202> (дата обращения: 17.12.2024).
2. Муштак О. И. "Извлечение знаний — особенности коммуникативных и текстологических методов при проектировании интеллектуальных информационных систем". Студентка 5 курса направления «Философия» Департамента философии Уральского гуманитарного института Уральского федерального университета имени первого Президента России Б. Н. Ельцина, г. Екатеринбург. [Электронный ресурс]. URL: <https://elar.urfu.ru/bitstream/10995/80575/1/episteme_2017_06.pdf> (дата обращения: 17.12.2024).
3. Четвёркин И. И. “Автоматизированное формирование базы знаний для задачи анализа мнений”. [Автореферат диссертации]. М.: МГУ, 2013. URL: <https://cs.msu.ru/sites/cmc/files/theses/201311-iic.pdf> (дата обращения: 17.12.2024).
4. Чуприна С. И., Минин А. С. “Искусственный интеллект и машинное обучение. Методы и средства построения онтологически управляемых систем приобретения знаний”. [Электронный ресурс]. URL: <https://cyberleninka.ru/article/n/metody-i-sredstva-postroeniya-ontologicheski-upravlyaemyh-sistem-priobreteniya-znaniy> (дата обращения: 17.12.2024).
5. 3Blue1Brown. "Neural networks". [Электронный ресурс]. URL: <https://www.youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi> (дата обращения: 17.12.2024).
6. SkillFactory. "Гайд по работе языковых моделей для начинающих"**.** [Электронный ресурс]. URL: <https://habr.com/ru/companies/skillfactory/articles/837366/> (дата обращения: 17.12.2024).