**Слайд 0: Титульный лист**

Добрый день, уважаемые коллеги! Меня зовут Соломенников Николай, и сегодня мы рассмотрим тему “извлечение знаний о предметных областях для создания интеллектуальных помощников”.

**Слайд 1: Введение**

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.

**Слайд 2: Актуальность**

Актуальность темы можно рассмотреть с нескольких аспектов:

1. Современные помощники становятся незаменимой частью многих сфер жизни — от повседневных задач до сложных профессиональных областей.
2. Современный мир порождает огромные объемы информации, преобразование которой в структурированную форму требует эффективных методов извлечения знаний.
3. Пользователи все чаще ожидают, что помощники будут обладать глубокими знаниями в узкоспециализированных областях.
4. 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 помощником, точность его ответов и их релевантность.

**Слайд 3: Цели и задачи**

Основной целью былоисследование методов извлечения знаний о предметных областях для создания интеллектуальных помощников.

А задачи включали:

1. Анализ современных научных и практических источников.
2. Выявление основных концепции и теоретических основ, лежащих в основе технологий извлечения знаний.
3. Систематизация существующих подходов и технологии для создания баз знаний.
4. Обозначение направлений, требующих дальнейшего изучения.

**Слайд 4: Хранение знаний**

Прежде чем говорить о методах извлечения знаний, важно понимать, что такое знания, как они структурируются и где хранятся.

**Знания** — представляют собой результат мыслительной деятельности человека, основанный на обработке данных и обобщении его опыта. Знания выражают закономерности, принципы, связи и законы, характерные для определенной предметной области.

**Интеллектуальные помощники** являются частным случаем интеллектуальных систем, то есть систем искусственного интеллекта, ядром которых является база знаний. В свою очередь важным классом интеллектуальных систем являются **экспертные системы**, которые отражают опыт специалистов в областях.

Ключевым элементом этих систем является **база знаний**. Она предоставляет контекст, термины, правила и данные. Можно понимать базу знаний как хранилище всех знаний, необходимых для работы системы

Для формализации структуры знаний существуют различные **языки представления знаний**. Это формальные системы, которые используются для описания знаний, их хранения, обработки и использования. Именно они и заложены в базах знаний. Рассмотрим самые популярные из них.

**Слайд 5: Продукционные системы**

**Продукционные системы** представляютзнания в виде продукций (то есть правил) вида "Если <условие>, то <действие>".

Продукционные системы включает:

* базу данных (множество фактов)
* базу правил (набор продукций)
* интерпретатор (механизм логического вывода).

Среди достоинств данной системы: наглядность, высокая модульностью, лёгкостью внесения дополнений и изменений и простотой механизма логического вывода.

**Слайд 6: Семантические сети**

*Семантические сети* отображают информацию в виде графа, где узлы представляют понятия, а ребра — отношения между ними. Также уточняются атрибуты узлов и типы связей.

Семантических сетей позволяют хранить большие объёмы данных в структурированном и понятном для человека и системы виде. Они достаточно гибкие и универсальные. На их основе также можно делать логические выводы и находить скрытые связи между сущностями.

**Слайд 7: Фреймы**

Фреймы отражают структуру знаний в виде объектов с атрибутами и связями. Каждый фрэйм имеет имя и слоты, которые описывают его свойства. Каждый слот имеет имя, значение и демоны (процедуры, выполняемые при изменении значения).

Посредством слотов фреймы могут образовывать иерархии. Фреймовая модель наглядна, легко расширяется и модифицируется. Также она предоставляет большую свободу при описании знаний.

**Слайд 8: Онтологии**

*Онтологии* — описывают концепции, их свойства и связи в понятном для человека и системе виде. Основными компонентами онтологии являются: классы, экземпляры, свойства, отношения, а также аксиомы и правила, задающие ограничения и закономерности. Онтологии структурируют знания в иерархическом виде, где классы и их подклассы связаны отношениями наследования.

Преимущества онтологий:

* Универсальность и стандартизация описания знаний.
* Возможность представления сложных взаимосвязей между объектами.
* Логический вывод
* Простота интеграции данных из разных источников.

**Слайд 9: Методы извлечения знаний**

Для работы интеллектуальных систем необходимо наполнять их базу знаний, поэтому процесс извлечения знаний — это ключевой этап, включающий сбор, анализ и преобразование информации. В настоящее время процесс извлечения знаний остается «узким» местом при построении промышленных экспертных систем.

Существует множество методов извлечения знаний, которые можно условно разделить на две категории: *с участием человека* и *автоматические*. Эти подходы часто сочетаются, чтобы обеспечить полноту знаний и эффективность их извлечения.

**Слайд 10: 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. Тем не менее они полезны в случаях, когда необходимо работать со сложными данными.

Для методов с участием человека требуется как минимум один:

* Эксперт в предметной области
* Инженер по знаниям, которые специализируются на заполнении баз знаний
* Программист – для разработки экспертных систем
* Обычный пользователь, который зачастую привлекается для тестирования полученной базы знаний

Данные методы можно разделить на *коммуникативные* и *текстологические*.

Коммуникативные методы основаны на взаимодействии инженеров по знаниям с экспертами. А текстологические направлены на анализ текстовых источников.

… показать примеры

К проблемам методов извлечения знаний с участием человека можно отнести

* Значительные трудозатраты на разработку баз знаний
* Психологические проблемы взаимодействия персонала
* Профессиональные. Например, недостаточная квалификация у инженера по знаниям в предметной области приводит к неполнотой извлекаемых знаний и некорректной их интерпретацией

Одним из способов решения данных проблем являются инструментальные средства, которые позволяют экспертам самостоятельно заполнять базу знаний. Эти средства могут включать интеллектуальных помощников, которые будут вести активный диалог с экспертом и давать ему подсказки. Это позволяет сократить трудозатраты инженера по знаниям и перенаправить его деятельность на анализ, верификацию и отладку уже полученной БЗ.

**Слайд 11: Автоматические**

Автоматические методы извлечения знаний предполагают минимальное вмешательство человека и основаны на использовании алгоритмов и программ для обработки больших объёмов данных. Эти методы позволяют значительно ускорить процесс извлечения знаний и уменьшить трудозатраты, но требуют качественных исходных данных.

Автоматические методы можно классифицировать по степени структурированности обрабатываемых данных. Каждый тип требует своих подходов и методов обработки.

Методы для структурированных данных (например, баз данных) включают:

* Метод извлечения закономерностей, который ищет скрытые связи и паттерны в базах данных (используются алгоритмы Apriori и FP-Growth).
* Классификация и кластеризация: это одни из методов машинного обучения, направленный на разделение объектов на группы. На основе кластеров можно создать базы знаний, где каждый кластер становится разделом или темой.
* Регрессионный анализ используется для изучения зависимостей между переменными.
* OLAP-анализ — это многомерный анализ для выявления трендов.

2. Методы для полуструктурированных данных (например, XML-документов, JSON, HTML-страниц):

* Веб-скрейпинг: автоматический сбор и обработка данных с веб-страниц.
* Автоматическое построение онтологий, которое включает следующие этапы:
  1. Сбор данных
  2. Подготовка данных (удаление шумов, фильтрация лишней информации, нормализация и разбиения текста на части)
  3. Извлечение сущностей и отношений (с помощью методов Named Entity Recognition и Semantic Role Labeling и семантический анализ)
  4. Формализация знаний
  5. Построение иерархий понятий
  6. Обогащение онтологии дополнительными связями и свойствами
  7. Оценка и валидация онтологии (проверка корректности и полноты)
* Анализ логов: выявление ошибок и аномалий в системных записях.

3. Для неструктурированных данных:

* Text Mining - процесс извлечения полезной информации, закономерностей и знаний из текстовых данных. Осуществляется за счёт большого количества методов (например, машинное обучение, deep learning, лингвистический анализ, семантический анализ, статические методы и т. д.). Основные задачи Text Mining:
  + Извлечение сущностей
  + Классификация и кластеризация текстов
  + Извлечение ключевых слов и тем
  + Анализ эмоциональной окраски
  + Выявление связей и отношений
  + Поиск информации
  + Обнаружение аномалий
* Обработка естественного языка (NLP) — это область искусственного интеллекта, связанная с пониманием, обработкой и генерацией текстов на естественном языке. Она направлена на "понимание" языка компьютерами и включает как методы машинного обучения, так и строгие алгоритмы. Например: грамматический анализ, выявление популярных паттернов или скрытых тем, классификация текстов, методы deep learning и т. д.

**Слайд 12: Точки роста**

* Требуются более совершенные алгоритмы для извлечения знаний из больших данных с минимальным участием человека.
* Современные NLP-модели нуждаются в улучшении для точного анализа сложных отношений и контекста.
* Необходимы инструменты для извлечения знаний на разных языках с учетом культурных особенностей.
* Важно создавать методы для актуализации знаний в реальном времени.
* Сочетание автоматических методов и ручной разметки может повысить точность.
* Развитие онтологий и семантических сетей способствует улучшению понимания предметных областей.
* Создание методов, обеспечивающих этичное извлечение знаний.

**Слайд 13: Результаты**

Тема извлечения знаний очень обширна, тем не менее была проделана большая исследовательская работа и проведён анализ методов извлечения знаний. Тем самым поставленные задачи были успешно выполнены.