Новосибирский Государственный Университет

**Выпускная квалификационная работа бакалавра**

на тему:

# Разработка средств автоматизации построения таксономического ядра онтологий на основе корпуса текстов

Выполнила: Ахмадеева Ирина, ФИТ НГУ гр.9201

Научный руководитель: к.т.н., зав. лаб. ИСИ СО РАН, Загорулько Ю.А.

# Понятие онтологии

- В философии отология – это учение о бытии, о сущем, о его формах и фундаментальных принципах, о наиболее общих определениях и категориях бытия
- В ИИ **онтология** – точная спецификация концептуализации
- Основа онтологии:
  - Классы.
  - Отношения.
- Таксономия – скелет онтологии.

# Цель работы

- Разработать и реализовать методы и средства автоматизации построения таксономического ядра онтологии на основе корпуса текстов

# Задачи

- **Анализ** существующих **методов** автоматического построения онтологии.
- **Выбор методов**, подходящих для данной цели.
- **Разработка алгоритма** автоматического построения таксономического ядра онтологии на основе корпуса текстов.
- **Реализация** данного **алгоритма**.
- **Создание** графического **интерфейса** пользователя.

# Методы построения онтологий

- Статистические
- На основе продукций
- Анализ формальных понятий

# Статистические методы

- Основываются на частоте встречаемости слов в тексте.
- Достоинства:
  - Простота.
  - Универсальность.
  - Независимость от языка текстов.
- Недостатки:
  - Игнорирование семантической связи слов.

# Методы на основе продукций

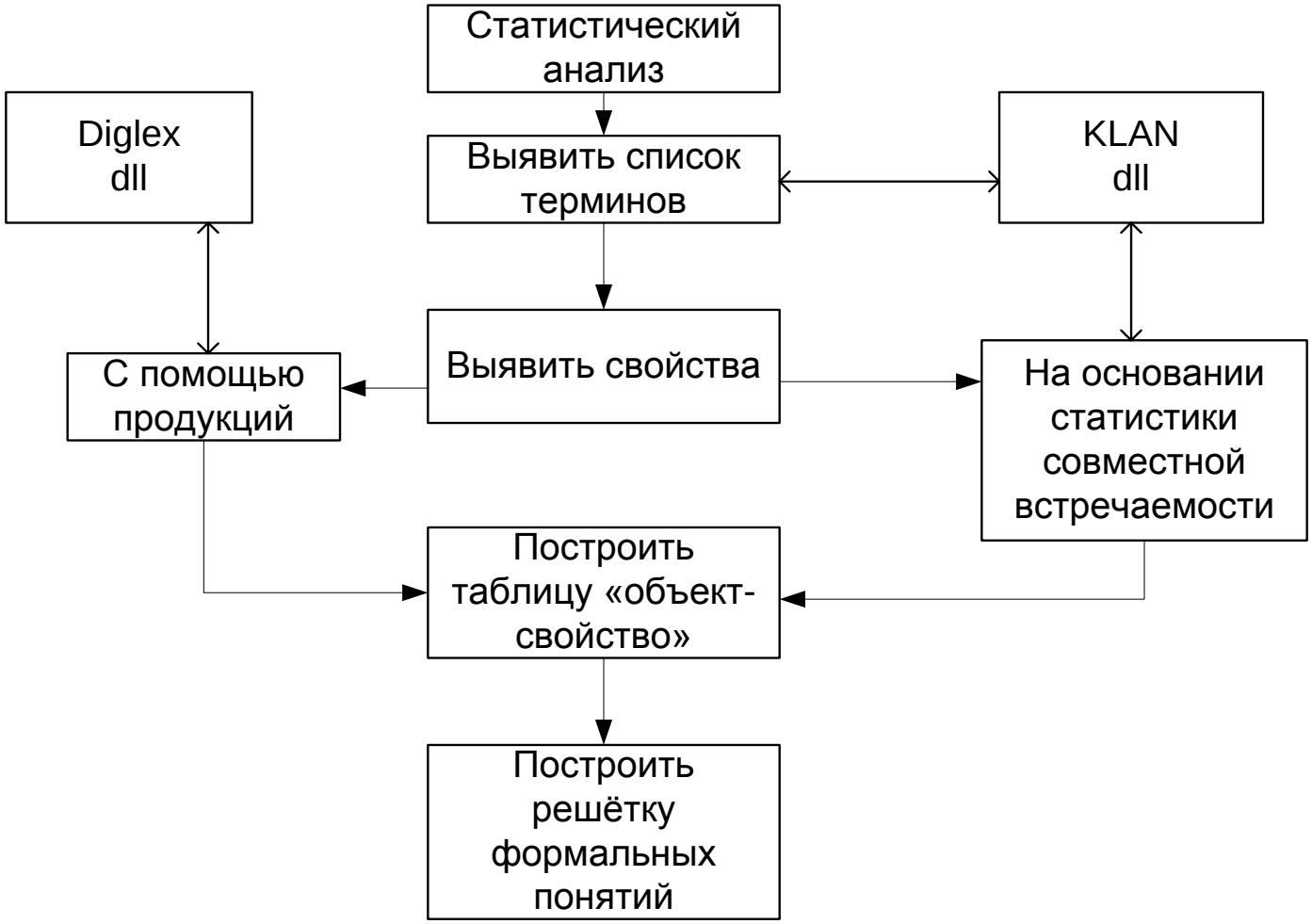
- Продукционное правило (продукция) – это пара «условие-действие».
  - Условная часть – шаблон, который определяет, когда правило может быть применено.
  - Часть действия определяет соответствующий шаг в решении задачи.

- Достоинства:
  - Удобство применения.
  - Понятность экспертам.
  - Способность системы объяснить решение.
- Недостатки:
  - Недостаточная семантическая связность между правилами.
  - Сложность в формировании правил.

# Анализ формальных понятий

- Исходные данные: формальный контекст K = (G, M, I), где
  - *G* – множество объектов.
  - *M* – множество признаков.
  - $I \subseteq G \times M$ (отношение обладания признаком).
- Результат: решетка формальных понятий.
- Достоинства:
  - Выявление таксономических отношений.
- Недостатки:
  - Построение формального контекста осуществляется экспертом.

# Алгоритм построения онтологии



# Построение таблицы «объект-свойство»

1. Предварительный этап
2. Извлечение терминов
    1. На основе статистики встречаемости
3. Извлечение свойств
    1. На основе статистики совместной встречаемости
    2. С использованием продукций
        - <термин> обладает свойством | признаком | параметром <свойство>
        - <термин> зависит от признаков | свойств | параметров <список свойств>
        - <термин> имеет признаки | свойства | параметры <список свойств>
        - <свойство> характерно для <термин>
        - <термин> обладает <свойство>
        - свойством | признаком | параметром <термин> является <признак>
        - <термин> характеризуется наличием <свойство>
        - <термин> характерно <свойство>
4. Формирование таблицы «объект-свойство»

# Построение решетки формальных понятий

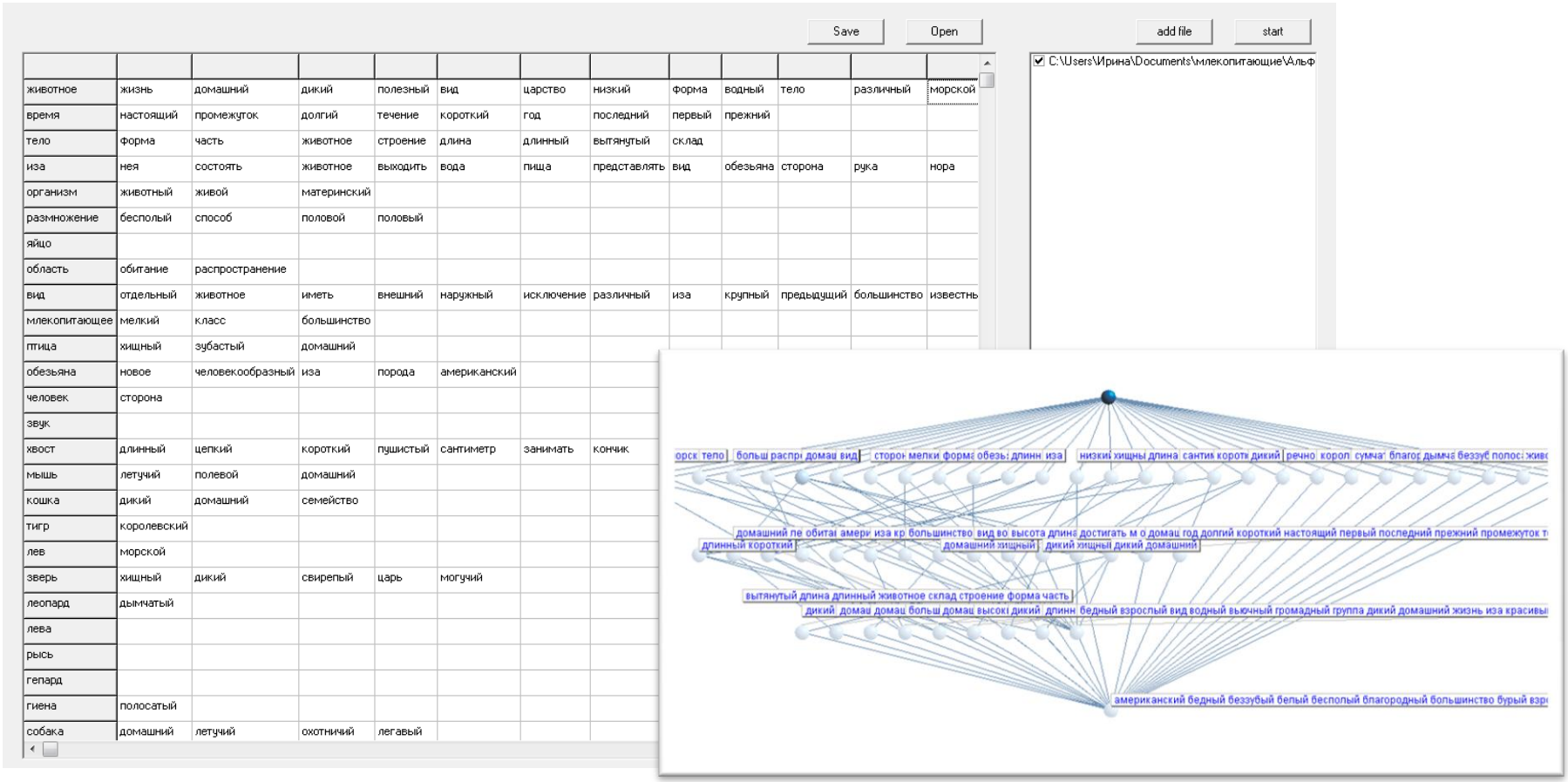
- **Операторы Галуа**:
  - Пусть K = (G, M, I) - формальный контекст, пусть $A \subseteq G$ и $B \subseteq M$, тогда:
    - $A' = \{m \in M \mid \forall g \in A : gIm\}$, т.е $A'$ – множество признаков, которыми обладают все объекты из множества $A$.
    - $B' = \{g \in G \mid \forall m \in B : gIm\}$, т.е. $B'$ – множество объектов, которые обладают всеми признаками из множества $B$.

# Построение решетки формальных понятий

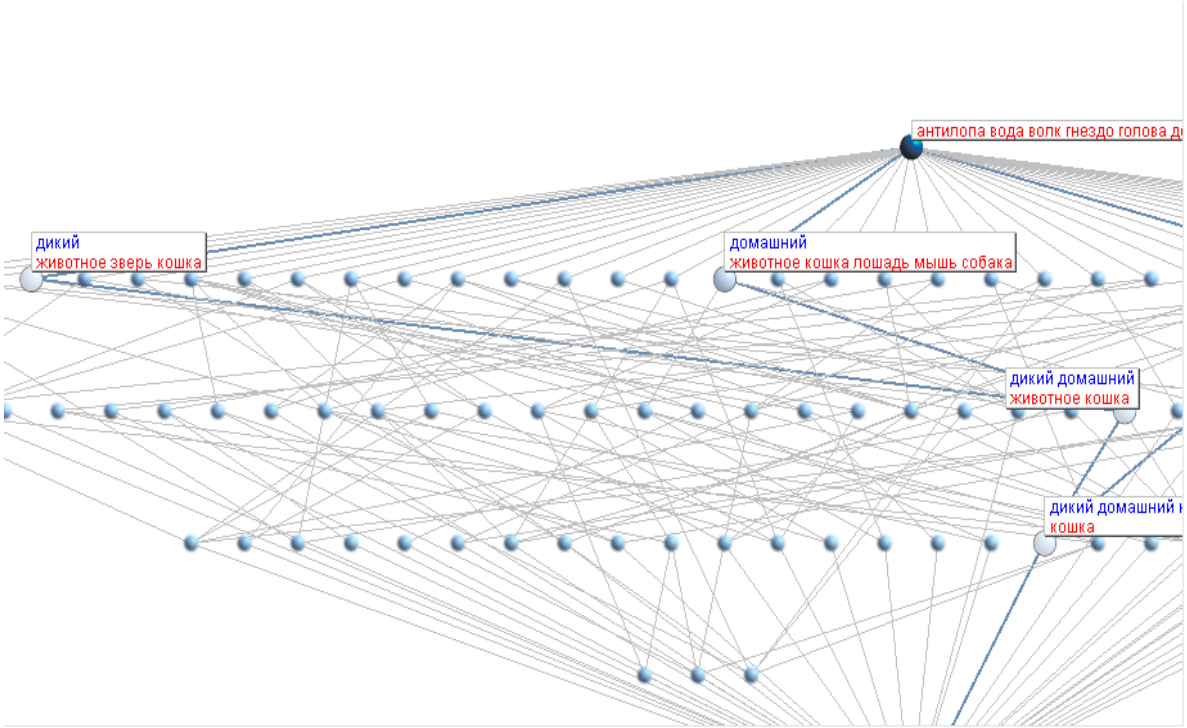
- **Формальное понятие** – это пара *(A, B)*, где $A \subseteq G$, $B \subseteq M$, такие что $A'=B$ и $B'=A$.
- *A* называется **объемом**, а *B* – **содержанием** понятия.
- Формальное понятие $(A_1, B_1)$ называется **подпонятием** $(A_2, B_2)$ если $A_1 \subseteq A_2$ (эквивалентно $B_2 \subseteq B_1$).

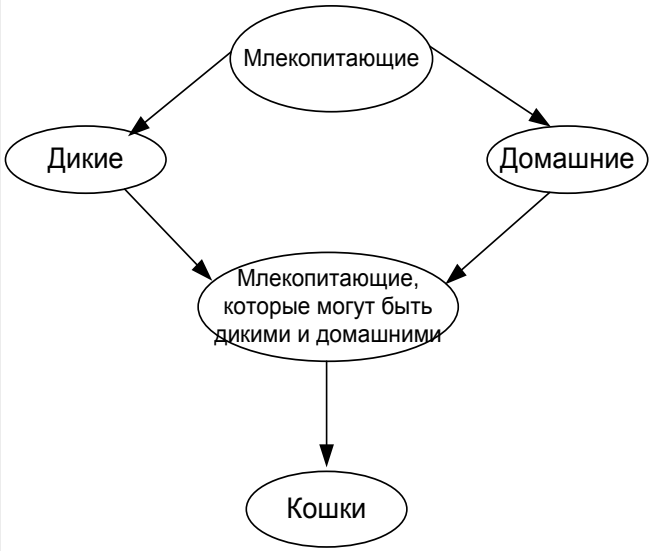
# Архитектура системы поддержки построения онтологий

# Программная реализация



# Пример


антилопа вода волк гнездо голова
дикий
животное зверь кошка
домашний
животное кошка лошадь мышь собака
дикий домашний
животное кошка
дикий домашний
кошка


Млекопитающие
Дикие
Домашние
Млекопитающие, которые могут быть дикими и домашними
Кошки

# Результаты

- **Выполнен анализ** существующих методов автоматического построения онтологии.
- **Выбраны методы**, подходящие для автоматизации построения таксономического ядра онтологии.
- **Разработан алгоритм** автоматического построения таксономического ядра онтологии на основе корпуса текстов.
- Предложенный **алгоритм реализован** в виде программы с графическим интерфейсом пользователя.