Разработка универсальной платформы для идентификации объектов по минимально возможному числу признаков

Выпускная квалификационная работа студента 461 группы

Афанасенко Никиты Владимировича

Научный руководитель:
аспирант кафедры системного программирования Вахитов А.Т.

Рецензент:
аспирант кафедры системного программирования Гуревич Л.С.

Постановка задачи

- Проанализировать существующие средства идентификации в биологии и выявить их недостатки
- Разработать метод оптимизации процесса идентификации и сравнить его с наиболее часто употребляемым
- Спроектировать и разработать веб-сервис, на базе которого может быть построен произвольный определитель
- Разработать демонстрационное клиентское приложение

Выявленные недостатки

- Большое количество задаваемых исследователю вопросов
- Невысокая скорость работы
- Невозможность построения собственного определителя
- Жесткая привязка к интерфейсу пользователя
- Отсутствие API для доступа к функциям определителя
- Платформенные ограничения

Способы оптимизации процесса идентификации

- Минимизация средней длины пути идентификации
- Повышение надежности результата
  - Экспертные оценки
  - Учет сложных для идентификации признаков

Оптимизация процесса идентификации

- Подсчет минимального набора признаков, отделяющего класс от всех остальных
  - Сводится к поиску наименьешго набора переменных, обращающих булевскую формулу в истину
- Было отмечено, что в среднем 73% присваиваний обращают формулы в истину
- Используем алгоритм для формул с большим количеством удовлетворяющих присваиваний (Гирш 1998)
- По сравнению с энтропийным методом улучшение на 20-50%

Результаты экспериментов

- **35 классов, 38 признаков, 114 состояний**

  Энтропийный метод: 7.71 шагов

K = 1	K = 7	K = 15
10.5	9.54	6.98

- **130 классов, 24 признака, 103 состояния**

  Энтропийный метод: 14.52 шагов

K = 1	K = 7	K = 15
11.63	11.28	11.24

- **1039 классов, 331 признак, 909 состояний**

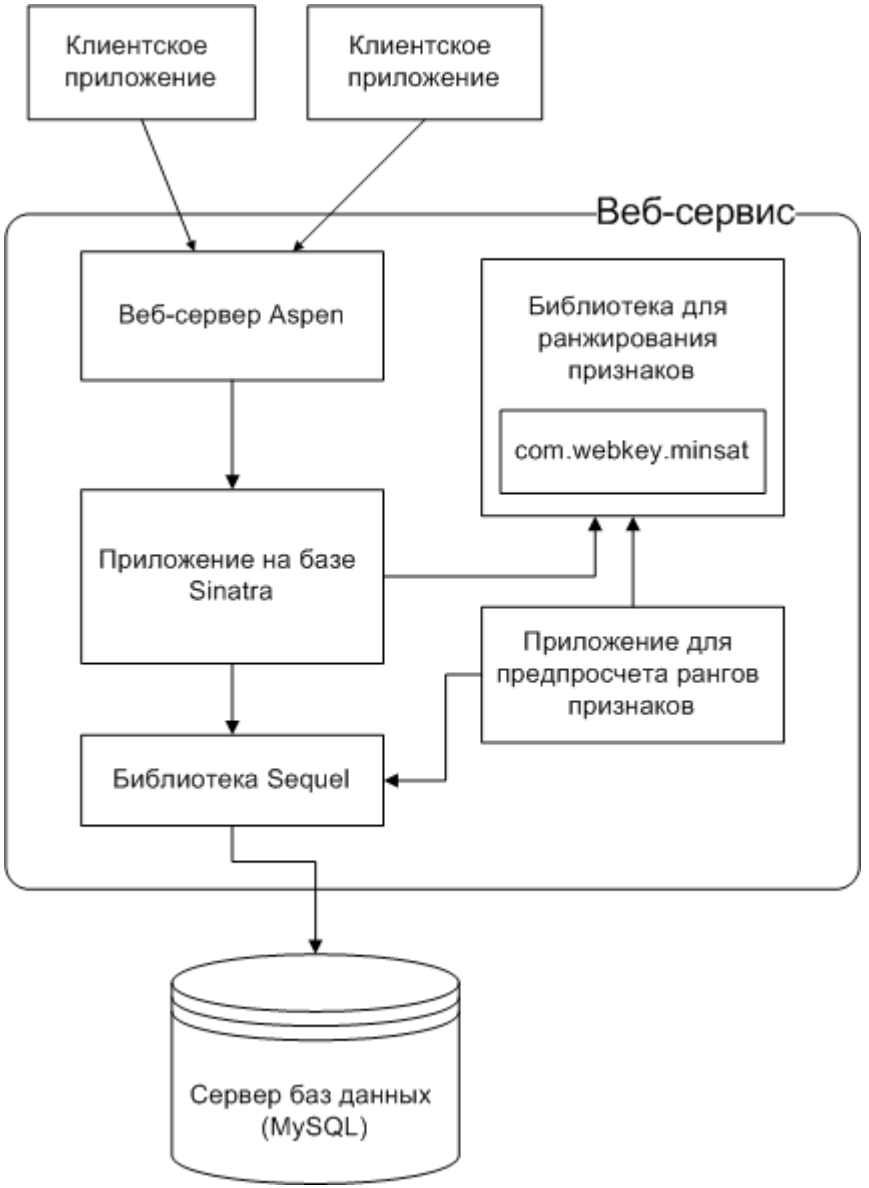
  Энтропийный метод: 53.17 шагов

  Метод минимальных разделяющих множеств: **24.3** шага

Реализация определителя

- Клиент-серверный подход позволит решить архитектурные проблемы существующих определителей
- Критерии выбора платформы:
  - Скорость разработки
  - Доступность технологии для всех платформ
  - Использование быстрых языков
  - Библиотека готовых решений
- Выбран JRuby

Архитектура веб-сервиса



Результаты

- Проведен анализ биологических определителей и выявлены их недостатки
- Найден метод оптимизации процесса идентификации
- Проведено имитационное моделирование, показавшее улучшение на 20-50% по сравнению с энтропийным методом
- Даны рекомендации по применению нового метода в биологических определителях
- Реализованы веб-сервис и демонстрационное клиентское приложение