Система хранения данных с группировкой по категории KAPLARO

Тверьянович М.А. 661 группа

Научный руководитель
Старший преподаватель Луцив Д.В.

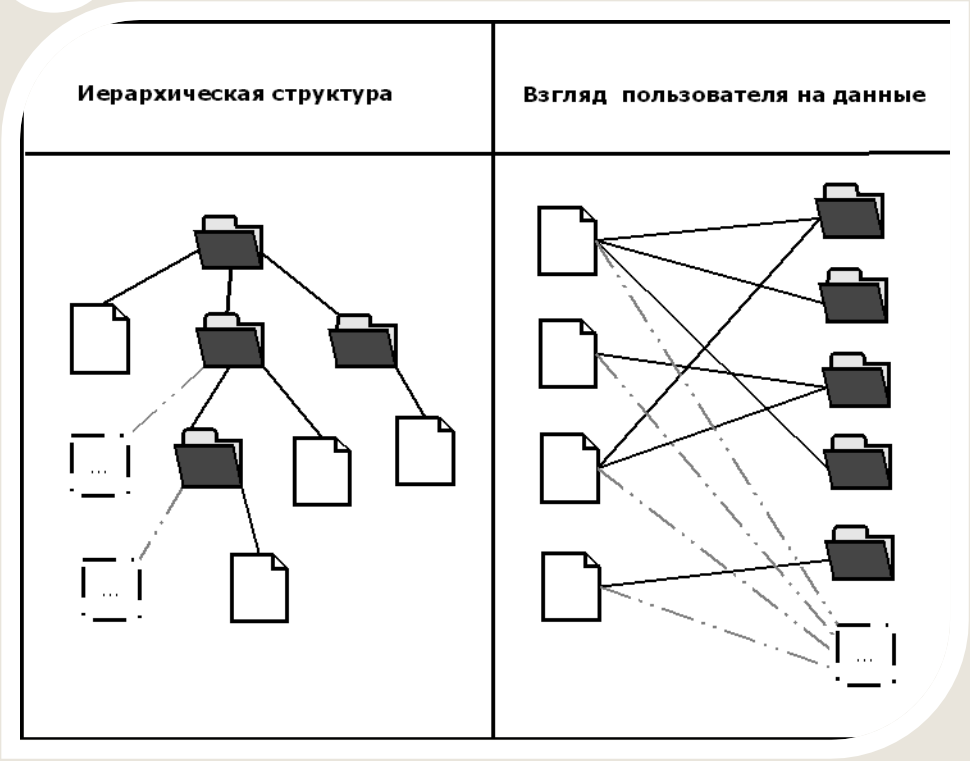
Рецензент
Доцент Графеева Н.Г.

Проблематика

- ***Традиционные способы описания данных ограниченны***
  - во основном название и путь
- =>
  - Сложности при поиске
  - Затраты времени на структурирование данных

Структурирование данных

- ***Иерархия***
  - отношение «один ко многим»
- ***Человек мыслит ассоциациями***
  - «многие ко многим»
- ***Добавление меток***
  - расширяет возможности для описания
  - разделяет данные на категории



Постановка задачи

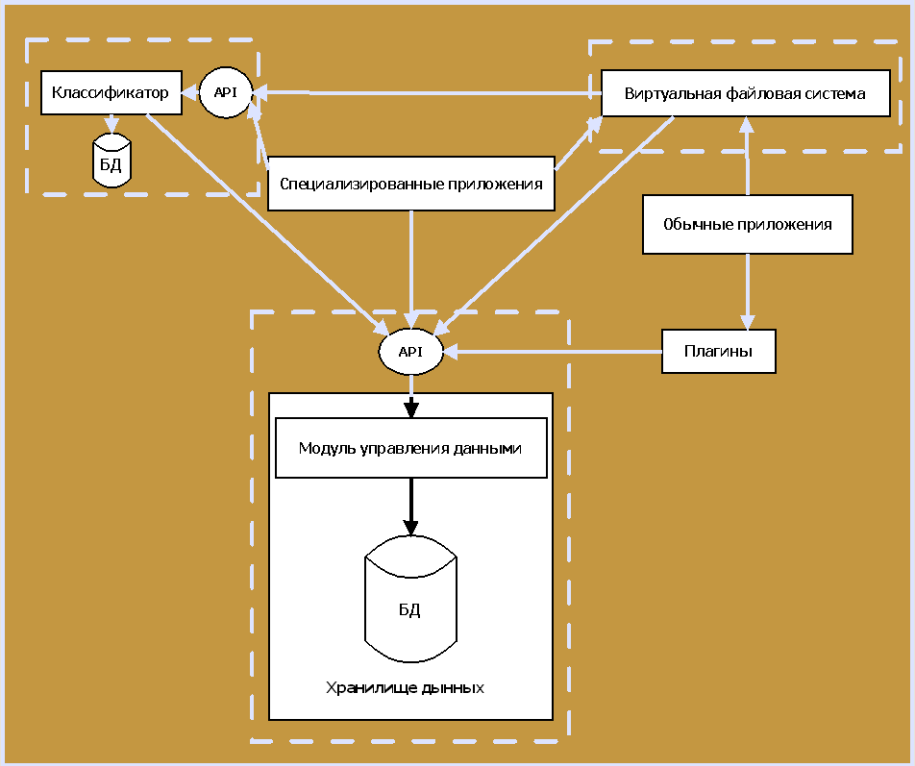
В результате данной работы предполагается:

- рассмотреть существующие модели хранения данных мультимедиа;
- разработать модульную архитектуру системы хранения данных, структурирующей файлы по метаданным;
- реализовать прототип системы, обладающей следующими свойствами
  - Возможность поиска файлов: по метаданным, по поисковому запросу
  - Проектирование и реализация полуавтоматической классификации файлов
  - Проектирование интерфейсов: API, виртуальной файловой системы

KAPLARO

Архитектура

- Хранилище данных
  - Модуль управления данными
  - База данных
  - API
- Специализированные приложения
  - Пользовательский интерфейс



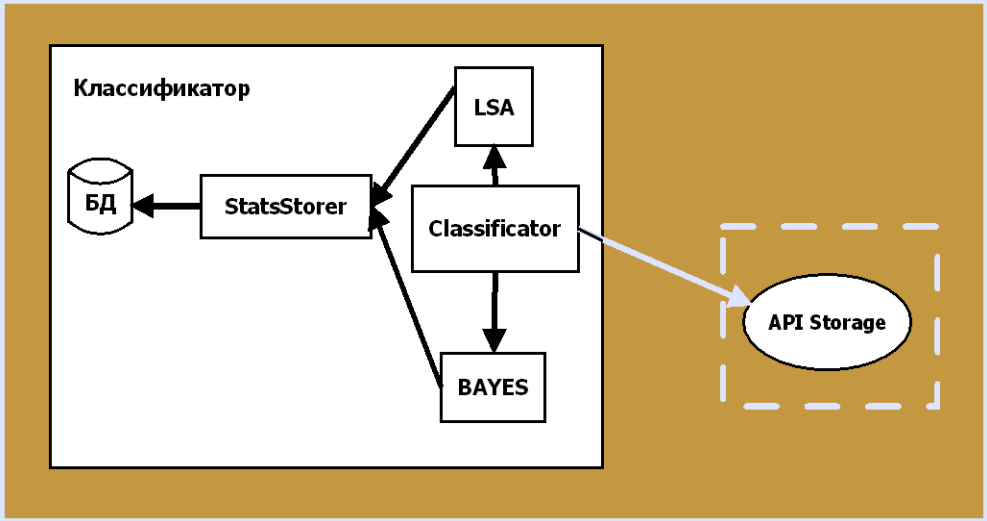
Классификатор

Методы

- LSA (ЛСА)– латентно-семантический анализ
  - метод обработки информации на естественном языке
  - анализирует взаимосвязь между набором документов и словами
- Наивная классификация Байеса
  - метод классификации на основе применения теоремы Байеса

Статистика

- Сохраняется в базе для последующего использования



Классификация нетекстовых данных

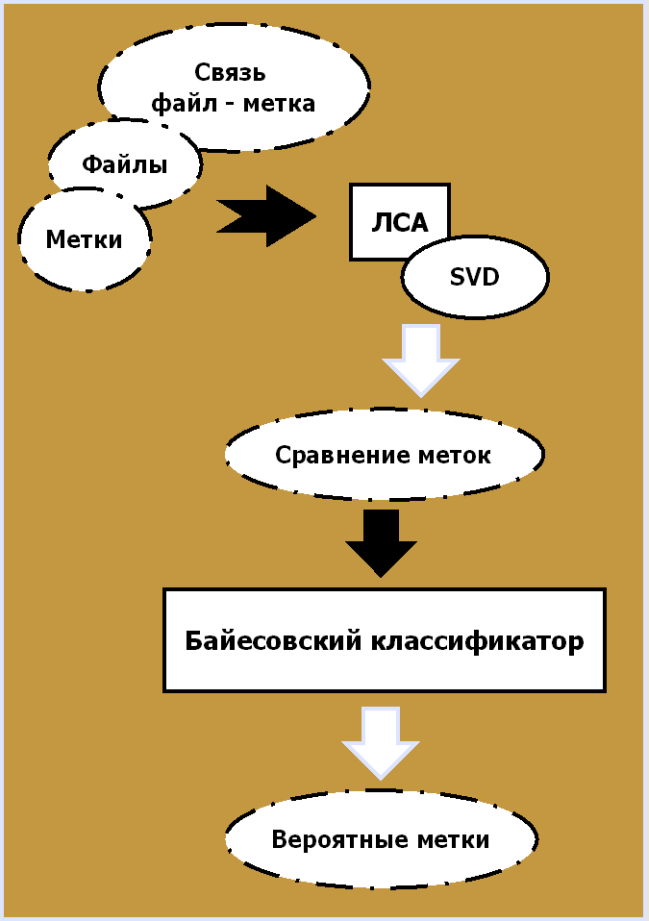
Данные:

- Метки , относящиеся к файлам

Алгоритм:

- ЛСА дает скрытые семантические связи между метками на основе корреляции
- Эти данные используются наивным Байесовским классификатором
  - для подсчета вероятности

    P(Метка | Метка)
- Результат наиболее вероятные метки для конкретного файла



Классификация текстовых данных

Данные:

- Слова из текстов, с исключением шумовых слов
- Слова более универсальный признак для классификации нежели метки

Для классификации используется:

- наивный Байесовский классификатор
  - Не требует затрат производительности при больших объемах данных
  - Дает быстрый результат
  - Постоянно обучается

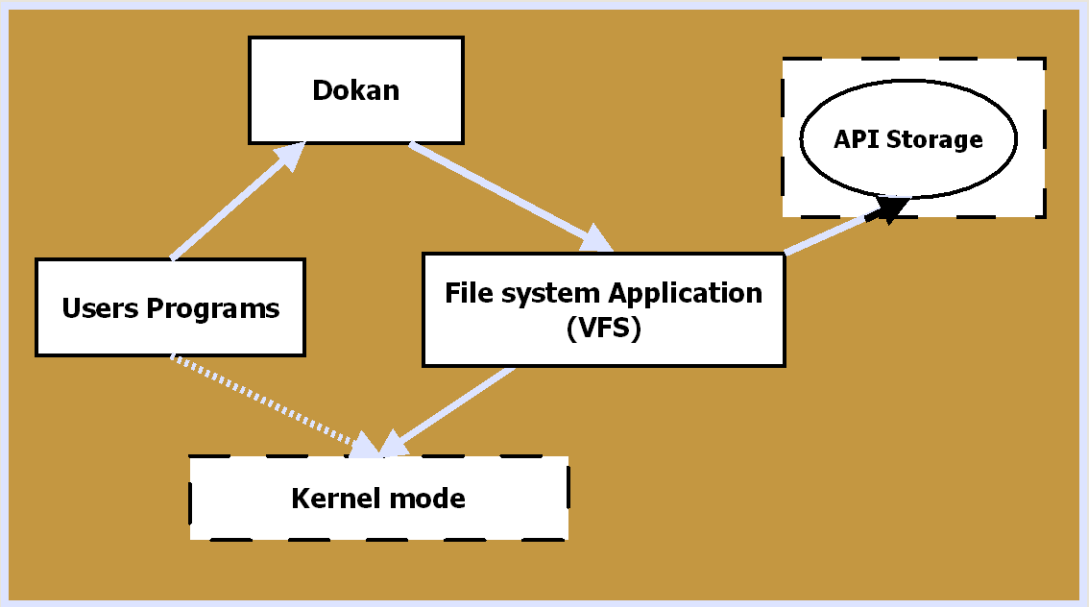
Виртуальная файловая система

ВФС:

- Реализована на основе проекта Dokan - user mode file system for windows

Путь в иерархии

- запрос к базе данных на основе меток
- Метка – каталог



Тестирование

ЛСА + Байесовский классификатор

- на неполных наборах меток, аналогичных наборам меток файлов, сохраненных в базе хранилища

СТАТИСТИКА	
150	файлов в базе
58	меток в базе
626	связей файл – метка
141	не полный набор меток для тестирования

РЕЗУЛЬТАТЫ ТЕСТИРОВАНИЯ	
127	Найдены все предполагаемые метки
6	Найдены все метки, но с вероятностью меньше максимальной
6	Найдена одна из двух меток
2	Метка не найдена
Результат в процентах	
Найдены	90,07%
Найдены с вероятностью меньше максимальной	4,26%
Найдена один из двух	4,26%
Не найдены	1,42%

Результаты

- Проведен обзор подходов к структуризации данных и подходов к классификации данных
- Подход и архитектура системы были представлены на конференции СПИСОК-2011, а также опубликованы в материалах конференции
- Разработан прототип системы на платформе Microsoft .NET на языке C# с использованием Microsoft SQL Server
- Спроектированы и разработаны компоненты системы: виртуальная файловая система, классификатор
- На базе Байесовской классификации, латентно-семантического анализа создан механизм полуавтоматического добавления меток к файлам в системе
- Реализована функциональность по поиску файлов: по метаданным, по поисковому запросу
- Функции системы проверены на тестовом наборе данных - набор мультимедийных файлов, описанных небольшим количеством меток