

КЛАСТЕРНАЯ ИНДЕКСАЦИЯ ФАЙЛОВ ДЛЯ ОПТИМИЗАЦИИ ПОИСКА ИНФОРМАЦИИ В РАСПРЕДЕЛЕННОЙ ФАЙЛОВОЙ СИСТЕМЕ

Кушвид Евгений Сергеевич
ст. гр. СШИМ-15-1

Руководитель:
к.т.н., доцент
Чалая Лариса Эрнестовна

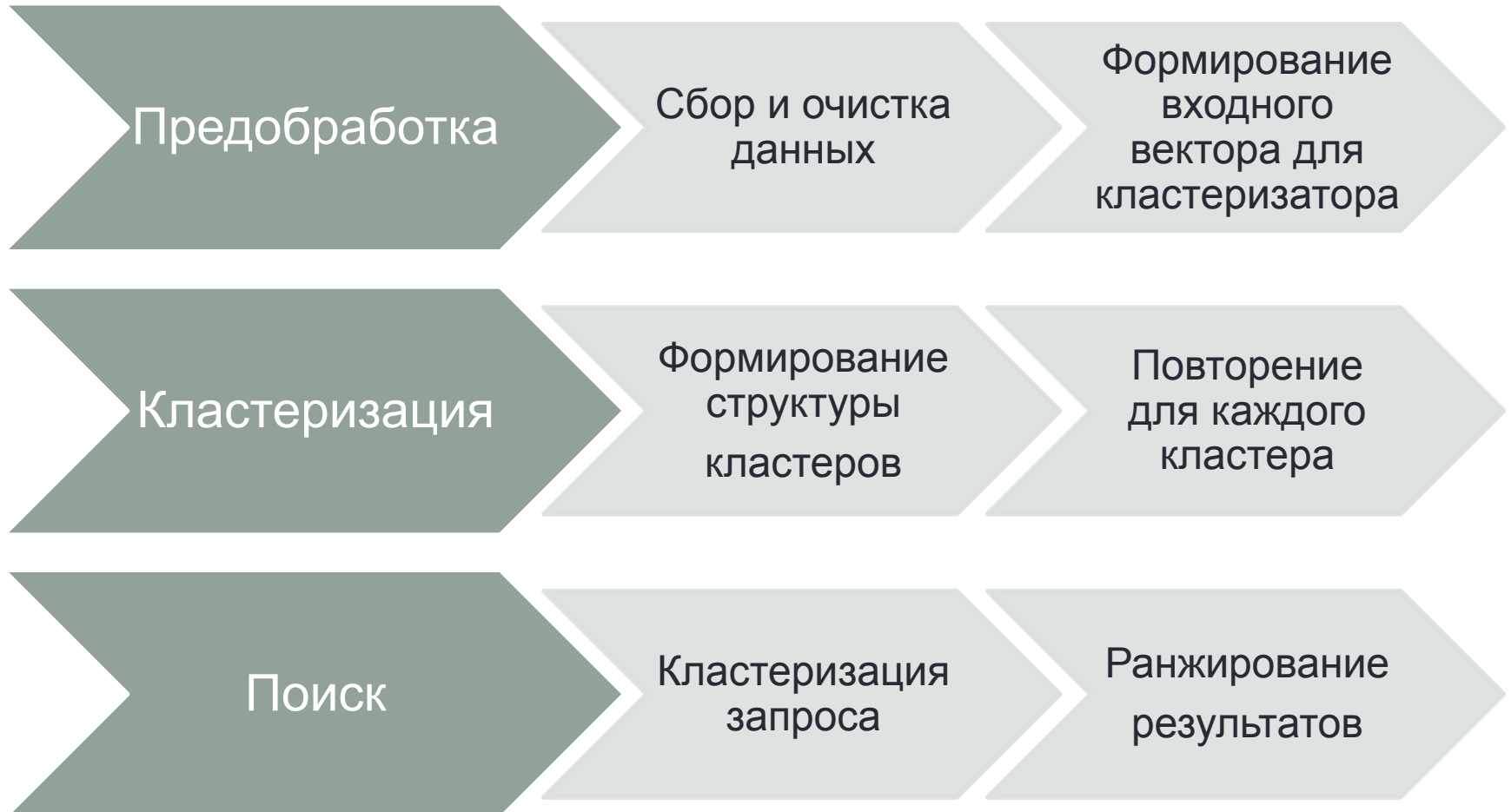
Актуальность:

- Высокие темпы роста объема текстовой информации
- Накопление неклассифицированных данных в распределенной структуре
- Отсутствие возможности оптимального смыслового определения архитектуры классов
- Необходимость в высококачественном и быстром поиске по большому массиву документов

Цель:

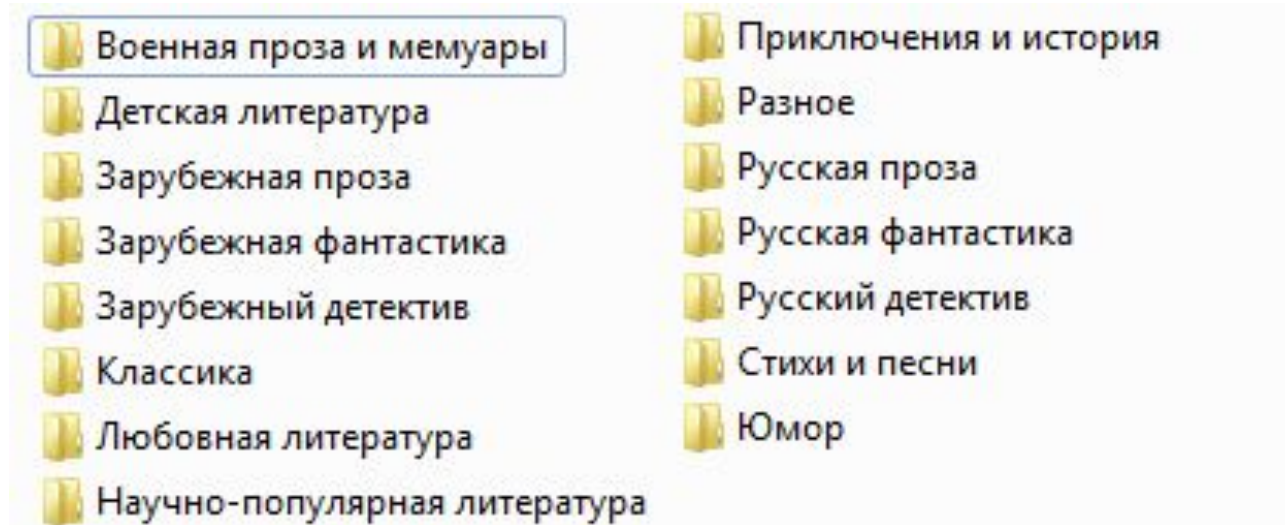
Целью работы является разработка метода эффективного поиска текстовой информации в распределенной файловой системе с высокой производительностью и качеством при малых ресурсных затратах приложения

Общая архитектура



Выборка:

Иерархическая структурированная библиотека
открытая для скачивания объемом 21гб



Сбор и очистка данных:



Формирование входного вектора для кластеризатора



Существующие подходы к кластеризации:

08 / 24

Алгоритм
k-средних

Нейронная
сеть
Кохонена

FOREL

SOINN

Существующие подходы к индексации

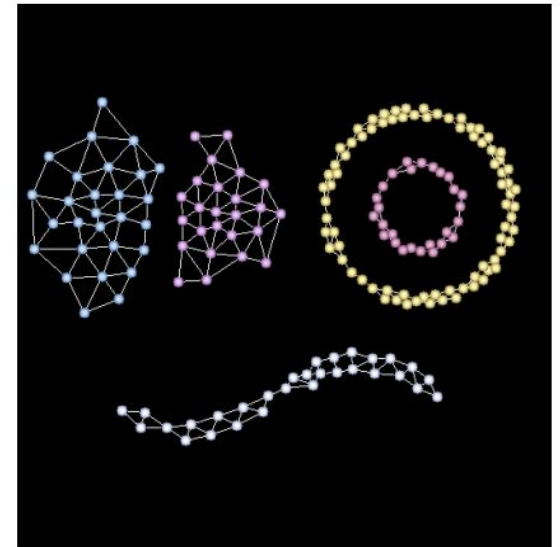
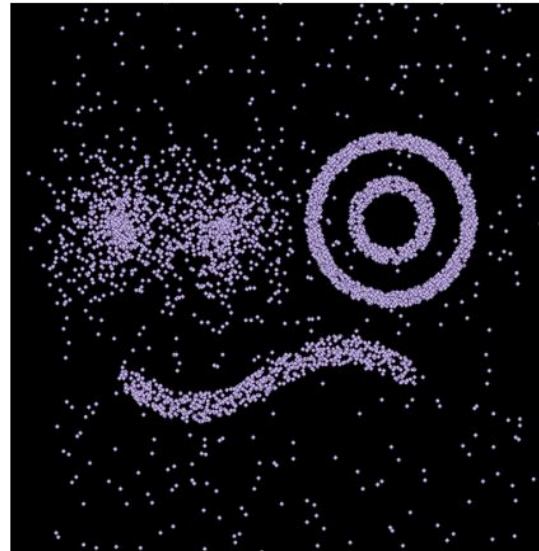
09 / 24

Прямой индекс

Инвертированный индекс

Обоснование выбранного решение:

SOINN



- Устойчивость к шумам
- Скорость
- Точность
- Адаптивность
- Отсутствие необходимости в эвристиках

Формирование структуры кластеров

11 / 24

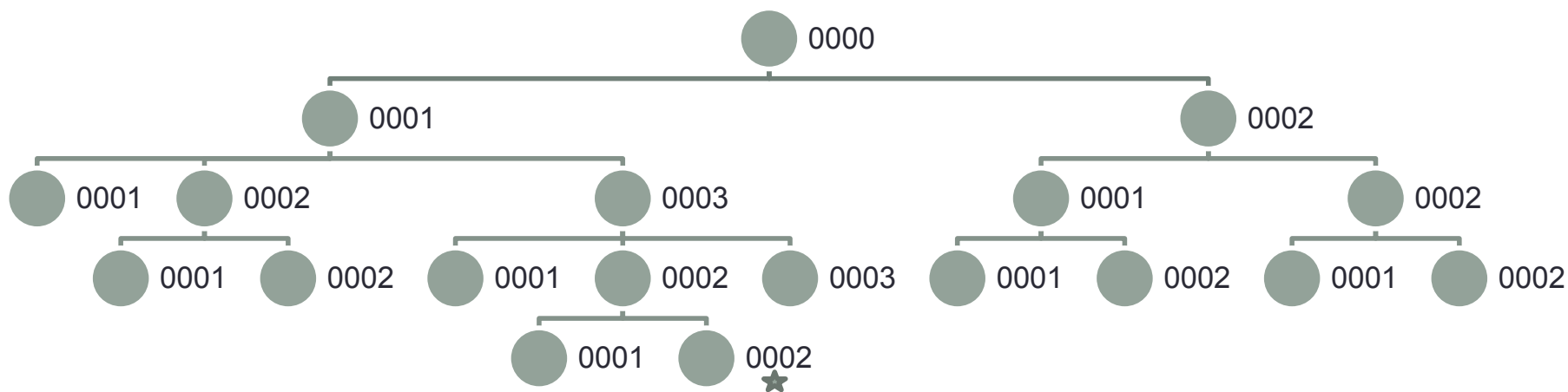


Индексация

Каждому кластеру присваивается уникальный индекс в порядке вложенности

Пределом кластеризации является сведение к один кластер это один файл и таким образом можно определить вложенность кластеров как уникальный
HASH

Пример иерархической индексированной кластерной структуры файлов



По окончании иерархической кластеризации каждый файл отделяется в отдельный персональный кластер (исключение: файлы дубликаты/копии)

Индекс файла со звездочкой: 0001000300020002

Пример вида метаданных в документе

Общие | Безопасность | **Особые** | Подробно

Имя: ▼

Тип: **Текстовый** ▼

Значение:

Свойства:

Имя	Значение	Тип
document_word_...	- 2- - fails - 49- - coke - ...	Т
index_datetime	2017-01-11 22:17	Т
index	0001_0001_0003_0002...	Т

Пример сохраненной метаданных в текстовом файле

15 / 24

```
Файл  Правка  Формат  Вид  Справка
|--- !ruby/object:DocumentSettingsindex: 0001_0004_0002_0003_0001_0008_0002_0005index_datetime: ^
2017-01-11 20:02:27 Zdocument_word_vector:- - sunflower - 2- - laziest - 1- - synagogues -
1- - recourse - 1- - sulphide - 2- - literary - 17- - loons - 1- - tumbleweeds - 3- -
firework - 8- - admiration - 18- - butch - 1- - signaled - 9- - deliveryman - 1- -
herpetic - 2- - objectification - 1- - underfunded - 18- - playing - 285- - multitudes -
1- - mutiny - 8- - disassociate - 4- - headbangers - 1- - disgust - 11- - communiqué - 1-
- swapping - 9- - skips - 2- - destroyed - 108- - chicks - 1- - antagonising - 1- -
confuses - 1- - ginned - 1- - evolves - 1- - notching - 2- - doorway - 3- - murky - 8- -
gaudy - 1- - dwellers - 6- - fancies - 2- - blacklisted - 5- - optic - 3- - bottlenecks
- 3- - suspicion - 82- - peek - 6- - monoclonal - 2- - crab - 3- - obvioulsy - 2- -
taxes - 89- - boards - 30- - alcoholics - 1- - seeped - 2- - arched - 2- - knucklehead -
1- - disks - 1- - blanch - 2- - fortress - 2- - pituitary - 1- - intrinsic - 3- - preys
- 1- - steered - 11- - unfaithful - 4- - tubal - 1- - blot - 9- - sheds - 11- - needing
- 28- - bonding - 4- - outrage - 83- - broody - 2- - enshrined - 14- - pans - 1- -
exhibitors - 2- - commuter - 25- - defamation - 11- - sheet - 20- - inaugura - 2- -
berate - 3- - finance - 64- - campus - 58- - elevation - 5- - supplanting - 2- -
assuring - 5- - plupart - 2- - watertight - 1- - pledged - 116- - genial - 1- - hey - 5-
- pirogue - 2- - westward - 1- - secret - 205- - bureaucratic - 17- - polarising - 2- -
commodities - 5- - rays - 3- - explains - 93- - arduous - 5- - lethargic - 5- - thorax -
1- - overstayers - 2- - unlocked - 10- - swarming - 2- - postmen - 1- - intended - 141- -
spreader - 1- - rapping - 2- - guideline - 2- - sported - 4- - gloomiest - 1- - prize -
49- - exhalation - 2- - swansong - 3- - researchers - 135- - outgrown - 1- - producing -
53- - grin - 8- - delivered - 161- - reconsiders - 4- - craftsman - 2- - showmanship - 1-
- advisories - 3- - seeded - 4- - publicly - 175- - descendants - 6- - reptile - 5- -
deindustrialisation - 2- - misuses - 2- - nonpublic - 5- - pun - 4- - baffle - 3- -
```

Поиск



Имплементация:



Визуальный интерфейс

Librarian

search for...

| type and press Enter

more.

< KNOWLEDGE IS POWER >

Поисковый запрос

Librarian

search for... ерь мы хотим, чтобы жил сверхчеловек | **more.**

Так говорил Заратустра.txt

D:/data/Научно-популярная литература/Философия/Ницше Фридрих/Так говорил
Заратустра.txt

Расширение поискового запроса

Librarian

search for...

Холера

more.

Холерные бунты.txt

Холерный вибрион.txt

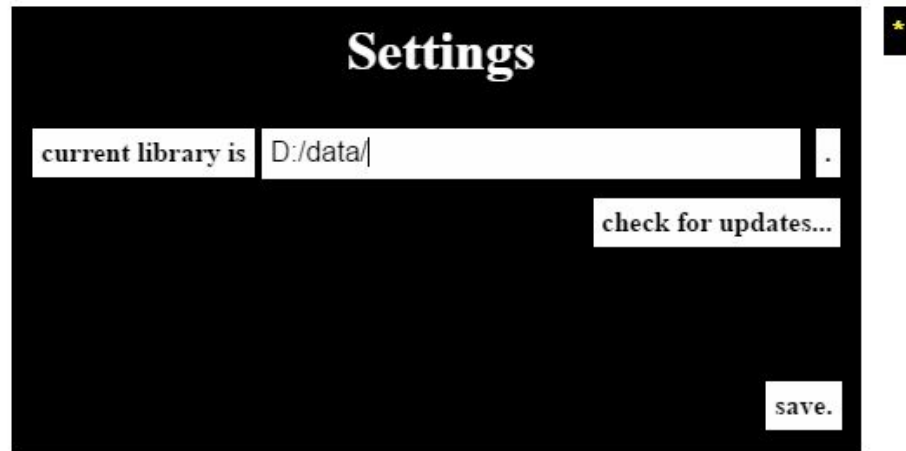
Холерик.txt

Холера.txt

Холевая кислота.txt

D:/data/Научно-популярная литература/Медицина/Бактериология/Холерный вибрион.txt

Экран настроек

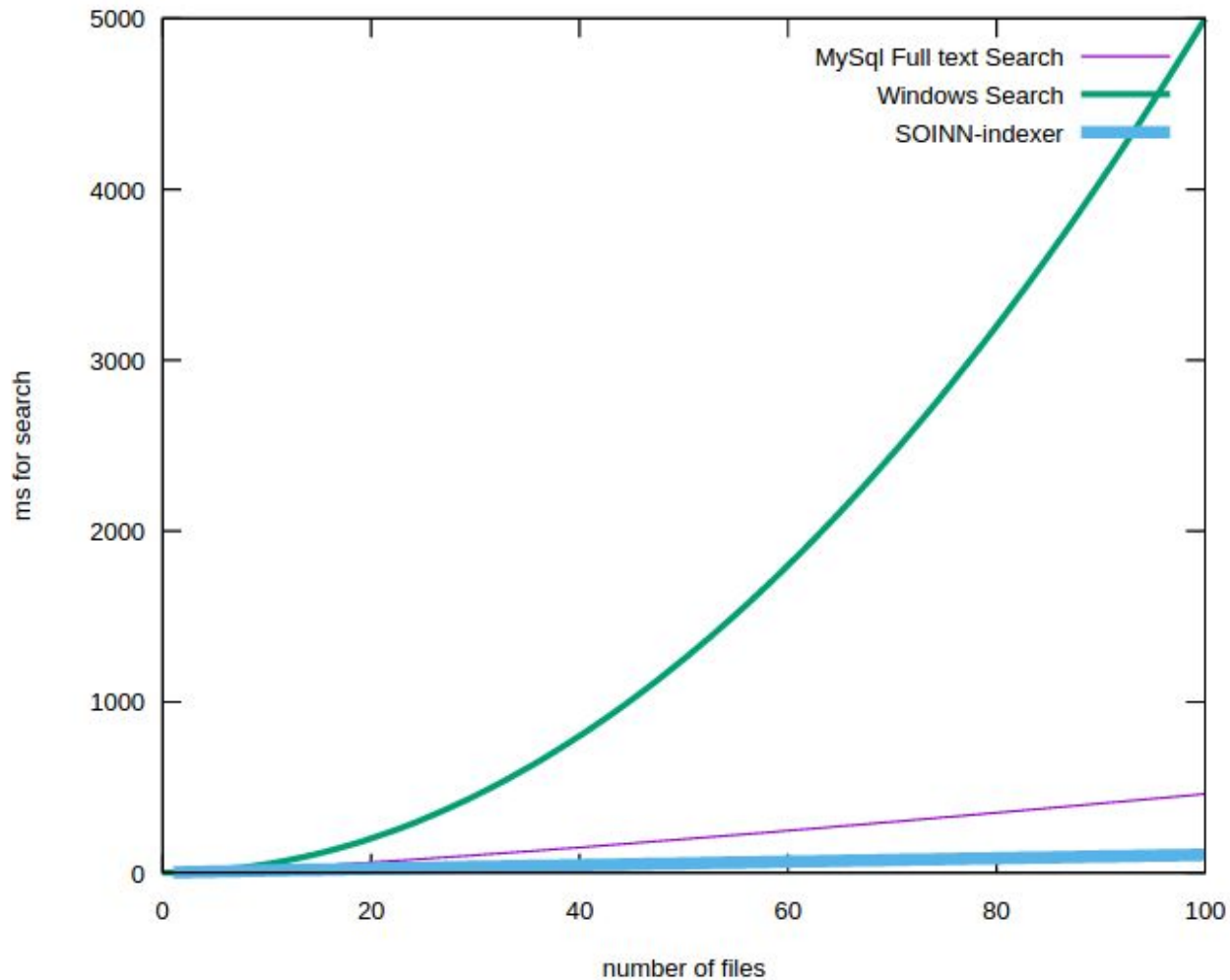


< KNOWLEDGE IS POWER >

Анализ эффективности:



Сравнительная характеристика алгоритмов поиска



Выводы

Недостатки

Достоинства

Привязка к Windows

Необходимость
иметь права
администратора

Необходимость
хранения структуры
сети

Быстрый

Точный

Распределенная
хранение структуры

Холодный старт

