Студент: Иванов Иван Иванович, ВМИ-399

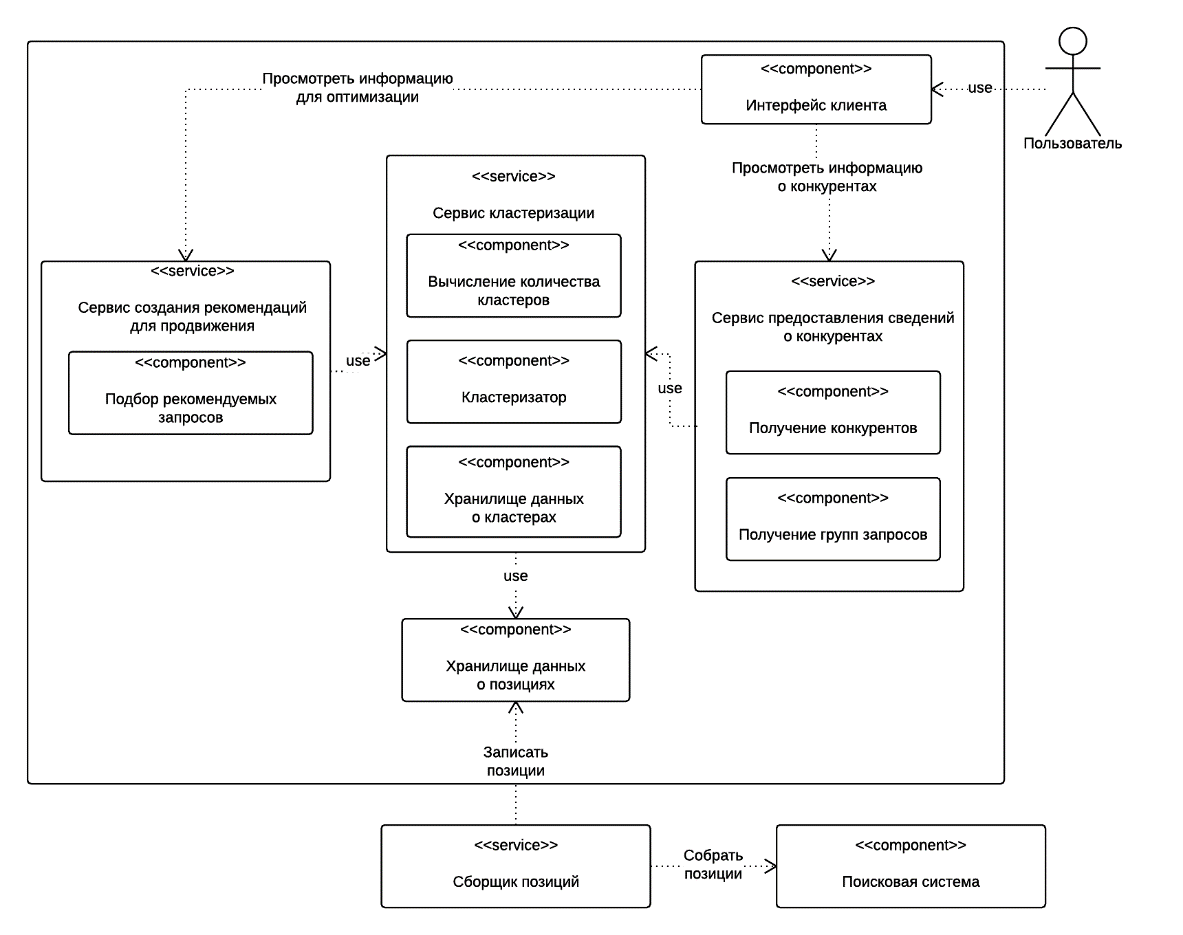
Руководитель: Петров Петр Петрович

Тема Разработка системы мониторинга и продвижения веб-сайтов.

# Архитектура системы мониторинга и продвижения веб-сайтов

## Компоненты системы мониторинга

На рисунке 1 изображена диаграмма компонентов интеллектуальной системы анализа данных мониторинга позиций сайтов в поисковых системах. Система состоит из следующих независимых сервисов: сервис кластеризации, сервис создания рекомендаций для продвижения, сервис предоставления сведений о конкурентах и сервис «хранилище данных о позициях». Сервис кластеризации является внутренним и не предоставляет открытого API для внешних пользователей. Сервис кластеризации использует данные о позициях из хранилища. Сервис создания рекомендаций для продвижения и сервис предоставления сведений о конкурентах используют данные о кластерах, предоставленные сервисом кластеризации*.*



**Рис. 1.** Компоненты системы мониторинга

Разберем более подробно компоненты, составляющие программную систему.

*Сборщик позиций* – это внешний сервис, обеспечивающий постоянный мониторинг позиций сайтов по запросам в поисковых системах и запись информации по позициям сайтов в хранилище.

*Сервис кластеризации* – это сервис, обеспечивающий разбиение запросов на кластеры. Объектом кластеризации является текст запроса к поисковой системе, привязанный к региону поиска. Реализация кластеризации в данной системе подробно рассматривается в п. 1.3 данной работы.

*Хранилище данных о позициях* – это база данных, содержащая информацию о результатах сбора позиций веб-сайтов по всем проектам за весь период времени.

*Компонент вычисления количества кластеров* – система, определяющая оптимальное количество кластеров, на которые следует разбить исходные данные.

*Кластеризатор –* система, разбивающая исходные данные на заранее заданное количество кластеров.

*Хранилище данных о кластерах* – база данных, содержащая информацию о точках кластеризации и кластерах.

*Получение конкурентов* – система анализа данных хранилища кластеров, выявляющая конкурентов сайта, запрошенного пользователем.

*Получение групп запросов* – система анализа данных хранилища кластеров, выявляющая запросы, по которым сайт, заданный пользователем, является конкурентом для исходного сайта.

*Подбор рекомендуемых запросов* – система анализа данных хранилища, определяющая запросы, по которым рекомендуется начать продвижение по сайту, запрошенному пользователем.

*Интерфейс клиента* – пользовательский интерфейс, с помощью которого можно получить сведения о конкурентах.

## Схема базы данных

Для организации работы мониторинга и продвижения веб-сайтов была разработана схема баз данных, содержащих информацию о поисковых запросах, найденных веб-сайтах и степени связи между запросами и веб-сайтами. Схемы разработанных хранилищ данных представлены на рисунках 2 и 3.



**Рис. 2.** Хранилище данных о позициях



**Рис. 3.** Хранилище данных о кластерах

Рассмотрим структуру базы данных о позициях:

* queries – это таблица для связи строки запроса с конкретным регионом;
* urls – это таблица, содержащая URL веб-сайтов, которые использует система для мониторинга позиций;
* regions – это таблица, содержащая коды регионов, по которым производится мониторинг;
* subscriptions – это таблица, реализующая связь «многие-ко-многим» между таблицами queries и regions;
* search\_results – это таблица, содержащая информацию о позиции запроса для каждого исследуемого региона в поисковой системе Яндекс.

Рассмотрим структуру базы данных о кластерах:

* points – это таблица, содержащая информацию о кластеризуемых элементах;
* clusters – это таблица, содержащая информацию о полученных в ходе кластеризации подмножествах;
* clusres\_points – это таблица, содержащая информацию о принадлежности элементов кластерам.

## Применение алгоритма кластеризации для запросов

В результате работы алгоритма пользователь сможет увидеть конкурентов для каждого проекта, зарегистрированного системе, а также список запросов для каждого конкурента, по которым он соперничает с искомым сайтом.

*Алгоритм четких k-средних применительно к решаемой задаче*

Поставим задачу кластеризации.

1. В качестве объектов кластеризации возьмем множество запросов .
2. В качестве признаков кластеризации по запросу возьмем строку матрицы , содержащую усредненную позицию всех анализируемых запросу за дней.
3. В качестве признака сходства запросов и возьмем евклидову метрику расстояния между координатами запросов в мерном пространстве, определенных соответствующими строками матрицы :

Алгоритм реализуется следующим образом.

1. Случайным образом инициализировать центроиды кластеров запросами так, что количество совпадает со значением оптимального количества кластеров, вычисленным ранее.

2. Для каждого центроида вычислить расстояния до всех точек

3. Сформировать кластеры. Для каждого центроида из множества отобрать подмножество точек с минимальным расстоянием до

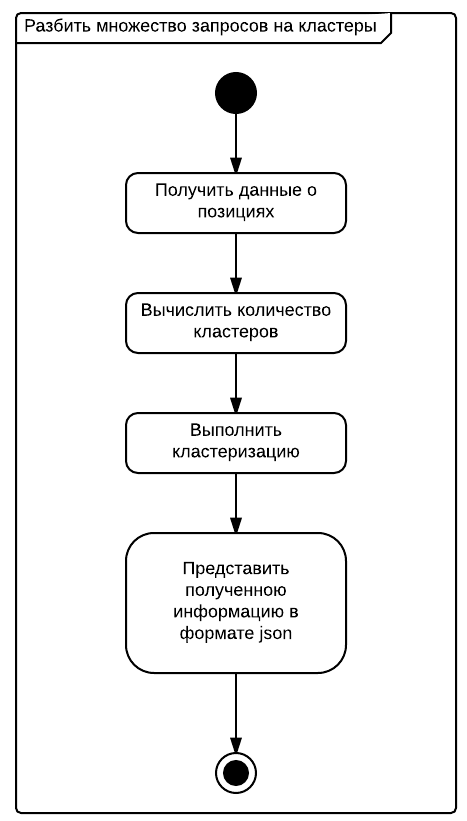
4. Вычислить новые центроиды как среднее всех точек кластера :

где *-* количество точек в кластере *.*

5. Сравнить значения с полученными на предыдущей итерации. В случае нахождения хотя бы одного несовпадения перейти к шагу 2.

6. Конец алгоритма.

Использование алгоритма k-means в разрабатываемой системе представлено на рис. 4.



**Рис. 4.** Диаграмма деятельности алгоритма кластеризации

## Вывод

В соответствии с системными требованиями была спроектирована архитектура системы мониторинга и продвижения сайтов в поисковых системах. Система имеет сервис-ориентированную архитектуру, состоящую из трех сервисов. Кроме того, в соответствии с функциональными требованиями были выделены компоненты, представляющие собой интерфейсы для взаимодействия с внешними сущностями. Для реализации основной функциональности был разработан алгоритм поиска конкурентов веб-сайта на основе алгоритма четкой кластеризации k-средних. Также для хранения результатов работы системы были спроектированы хранилище данных о кластерах и хранилище данных о позициях. В соответствии с нефункциональными требованиями была подробно разработана реализация наиболее важнейших прецедентов. В соответствии с анализом сценариев работы пользователя были разработаны доступные пользовательские интерфейсы системы мониторинга и продвижения сайтов в поисковых системах.