МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение  
высшего профессионального образования

**«Южно-Уральский государственный университет»**

**(национальный исследовательский университет)**

**Факультет Вычислительной математики и информатики**

**Кафедра системного программирования**

|  |  |
| --- | --- |
| РАБОТА ПРОВЕРЕНА  Рецензент  <ученая степень, ученое звание>  \_\_\_\_\_\_\_\_\_\_ <И.О. Фамилия рецензента>  “\_\_\_”\_\_\_\_\_\_\_\_\_\_\_ 2015 г. | ДОПУСТИТЬ К ЗАЩИТЕ  Заведующий кафедрой, д.ф.-м.н., профессор  \_\_\_\_\_\_\_\_\_\_ Л.Б. Соколинский  “\_\_\_”\_\_\_\_\_\_\_\_\_\_\_ 2015 г. |

**Интеллектуальная система мониторинга и продвижения веб-сайтов в поисковых системах**

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ЮУрГУ – 010400.62.2015.11-027-19091.ВКР

|  |  |
| --- | --- |
|  | Научный руководитель  кандидат физико-математических наук, доцент каф. СП  \_\_\_\_\_\_\_\_\_\_Г.И. Радченко  Автор работы, студент группы ВМИ-411  \_\_\_\_\_\_\_\_\_\_Е.А. Неповинных  Ученый секретарь  (нормоконтролер)  \_\_\_\_\_\_\_\_\_\_\_\_\_ О.Н. Иванова  “\_\_\_”\_\_\_\_\_\_\_\_\_\_\_ 2015 г. |

Челябинск-2015

# Оглавление

[Оглавление 2](#_Toc418624402)

[Глоссарий 3](#_Toc418624403)

[Введение 4](#_Toc418624404)

[Структура и объем работы 5](#_Toc418624405)

[1 Обзор литературы 7](#_Toc418624406)

[1.1 Обзор алгоритмов 8](#_Toc418624407)

[1.2 Анализ существующих решений 10](#_Toc418624408)

[2 Требования к системе мониторинга и продвижения веб-сайтов 12](#_Toc418624409)

[2.1 Интеллектуальная система мониторинга и продвижения веб-сайтов 12](#_Toc418624410)

[2.2 Варианты использования системы 13](#_Toc418624411)

[2.3 Нефункциональные требования к системе 18](#_Toc418624412)

[2.4 Вывод 19](#_Toc418624413)

[3 Архитектура системы мониторинга и продвижения веб-сайтов 20](#_Toc418624414)

[3.1 Компоненты системы мониторинга 20](#_Toc418624415)

[3.2 Схема базы данных 22](#_Toc418624416)

[3.3 Применение алгоритма кластеризации для запросов 23](#_Toc418624417)

[3.3.1 Алгоритм четких k-средних применительно к решаемой задачи 25](#_Toc418624418)

[3.4 Вывод 27](#_Toc418624419)

[4 Реализация системы – дописать. 28](#_Toc418624420)

[4.1 База данных 28](#_Toc418624421)

[4.2 Реализация компонентов системы 28](#_Toc418624422)

[4.3 Организация взаимодействия компонентов системы 30](#_Toc418624423)

[4.4 Веб-интерфейс – доработать 31](#_Toc418624424)

[4.5 Отображение результатов 32](#_Toc418624425)

[5 Тестирование - сделать 33](#_Toc418624426)

[5.1 Реализация алгоритма 33](#_Toc418624427)

[6 Заключение 34](#_Toc418624428)

# Глоссарий

*Пользователь* – конкретное лицо или представитель организации, использующий систему для того, чтобы выполнять конкретные задачи или функции.

*Сбор позиций* — проверка позиций сайтов на странице результатов поиска в поисковой системе по заранее заданным параметрам.

*Ресурс* — веб-сайт, для которого осуществляется мониторинг позиций по заданному набору поисковых запросов.

*Мониторинг* —анализ позиции ресурса в результатах поиска.

*Позиция ресурса* в результатах поиска (далее: позиция) — порядковый номер ссылки на страницу определенного ресурса в результатах поиска по определенному поисковому запросу.

*Поисковый запрос (далее: запрос)* – строка для осуществления поиска с помощью поисковой системы.

*Регион мониторинга* – это параметр поискового запроса, осуществляющий территориальную специализацию результатов поиска.

*Результаты поиска* — упорядоченный список ссылок на различные ресурсы, которые пользователь видит в результате запроса к поисковой системе Яндекс.

*Продвижение интернет-сайта* — это комплекс мер по обеспечению [посещаемости сайта](https://ru.wikipedia.org/w/index.php?title=%D0%9F%D0%BE%D1%81%D0%B5%D1%89%D0%B0%D0%B5%D0%BC%D0%BE%D1%81%D1%82%D1%8C_%D1%81%D0%B0%D0%B9%D1%82%D0%B0&action=edit&redlink=1) потенциальными потребителями, которые заинтересованы в приобретении товаров или услуг, представленных на продвигаемом сайте.

*Анализ конкурентов* — метод определения компаний, работающих на рынке и предлагающих сходную продукцию, задачей которого является выявление сильных и слабых мест компании для последующей разработки правильной стратегии поведения компании на рынке и диверсификации рисков.

*Семантическое ядро сайта* — это набор поисковых слов, их морфологических форм и словосочетаний, которые наиболее точно характеризуют вид деятельности, товар или услугу, предлагаемые сайтом.

# Введение

Продвижение интернет-сайта в поисковых системах позволяет значительно увеличить количество просмотров и качество активных пользователей сайта. На сегодняшний день продвижение сайтов в русско-язычном сегменте сети Интернет осуществляется в 5-ти наиболее популярных поисковых системах [4]: Яндекс, Google, поиск Mail.ru, Rambler, Bing. Существует целый ряд различных способов продвижения интернет-сайта в поисковых системах: генерация и размещение на сайте уникального и релевантного контента, продвижение сайта в каталогах и социальных сетях, использование определенных ключевых слов, покупка внешних ссылок на сайт и др. Для скорейшего достижения высокой позиции сайта и привлечения большего числа клиентов, необходимо постоянно отслеживать положение продвигаемого веб-ресурса в поисковых системах. Для решения этой задачи существуют системы мониторинга [5]. Система мониторинга отслеживает и позволяет сохранять для дальнейшего анализа позиции веб-сайта в поисковых системах за определённый период времени. В настоящий момент наиболее развиты системы мониторинга в виде веб-приложений. Их популярность объясняется, главным образом, тем, что клиенты не зависят от конкретной операционной системы пользователя, поэтому веб-приложения являются межплатформенными сервисами.

Актуальность данной темы обусловлена необходимостью постоянного анализа позиций коммерческих сайтов в поисковой выдаче для увеличения их посещаемости в процессе поисковой оптимизации (Search Engine Optimization, SEO) сайта. Оперативное определение неэффективных запросов, по которым позиции сайта долгое время остаются низкими, позволяет скорректировать стратегию и оптимизировать расходы продвижения сайта, а данные о состоянии конкурентов помогают адекватно оценить перспективы продвигаемого проекта.

*Целью данной работы* является разработка интеллектуальной системы анализа данных для мониторинга позиций сайтов по запросам в поисковой системе Яндекс.

Для достижения указанной цели необходимо решить следующие задачи:

1. изучить особенности существующих систем мониторинга позиций сайтов в поисковых системах;
2. изучить существующие алгоритмы интеллектуального анализа данных;
3. спроектировать архитектуру системы интеллектуального анализа данных для мониторинга позиций сайтов в поисковых системах;
4. реализовать и протестировать интеллектуальную систему для мониторинга позиций сайтов по запросам в поисковой системе Яндекс.

## Структура и объем работы

Работа состоит из введения, пяти разделов, заключения, библиографии и приложения. Объем работы составляет 37 страниц, объем библиографии – 9 источников.

В первой главе дается обзор технологий работы поисковых систем. В этом разделе будут рассмотрены алгоритмы ранжирования наиболее популярных поисковых систем и выявлены наиболее перспективные сервисы для мониторинга позиций сайтов.

Вторая глава содержит описание и анализ требований к системе мониторинга позиций сайтов в поисковой системе Яндекс.

В третьей главе представлена архитектура системы мониторинга позиций сайтов в поисковой системе Яндекс. В этом разделе подробно рассмотрена общая архитектура системы, интерфейсы компонентов и детали их реализации.

Четвертая глава посвящена реализации системы мониторинга позиций сайтов в поисковой системе Яндекс с использованием технологий django, AJAX и MVC.

В пятой главе приводятся результаты тестирования системы мониторинга позиций сайтов в поисковой системе Яндекс.

В заключении сделаны выводы о проделанной работе.

Приложение содержит детальную реализацию основных прецедентов.

# Обзор литературы

Для успешного продвижения сайта и определения его конкурентов необходимо понять алгоритмы работы поисковых систем, в которых планируется продвигать сайт. Поисковая система получает и обрабатывает запрос, сортирует все имеющиеся в своей базе данные о сайтах, к которым может иметь отношение данный запрос, то есть производит ранжирование, и выдает результат. На странице результатов поиска в определенном порядке располагаются [ссылки](http://www.webeffector.ru/wiki/Ссылка) на веб-ресурсы, которые, по мнению поисковой системы, наиболее соответствуют указанному запросу  [1, 11].

Мониторинг позиций сайта дает возможность осуществлять контроль за рейтингом сайтов в поисковых системах по целому или частичному запросу и решает следующие задачи:

* контроль действий конкурентов;
* своевременная корректировка позиции сайта в соответствии с целевыми запросами;
* изменение текстового блока, находящегося на страницах сайта для улучшения позиции в поисковых машинах [8].

Анализ конкурентов способствует достижению целого спектра маркетинговых и производственных целей организации. На его основе можно спрогнозировать исход конкурентной борьбы или выбрать оптимальную модель продвижения и позиционирования продукции компании.

Качественный анализ ближайших конкурентов позволяет смоделировать все многообразие вариантов развития событий и своевременно вносить необходимые изменения в деятельность компании [9].

Сравнительный анализ конкурентов по временным и другим параметрам позволяет выявить динамику изменений маркетинговой стратегии, провести оценку наиболее эффективных ходов лидера рынка, обеспечивающих наилучшие результаты. Внедрение этих знаний в практику работы компании позволит минимизировать ошибки в планировании и реализации собственной маркетинговой стратегии, а также добиться увеличения доходов и доли рынка [8,9].

Наиболее эффективным методом анализа текущего состояния конкурентов компании является интеллектуальный анализ данных. Интеллектуальный анализ данных представляет собой процесс обнаружения пригодных к использованию сведений в крупных наборах данных [4]. В интеллектуальном анализе данных применяется математический анализ для выявления закономерностей и тенденций, существующих в данных. Обычно такие закономерности нельзя обнаружить при традиционном просмотре данных, поскольку связи слишком сложны, или из-за чрезмерного объема данных. Эти закономерности и тренды можно собрать вместе и определить как модель интеллектуального анализа данных [2]. Модели интеллектуального анализа данных могут применяться к конкретным сценариям, а именно:

* прогноз: оценка продаж, прогнозирование нагрузки сервера или времени простоя сервера;
* риски и вероятности: выбор наиболее подходящих заказчиков для целевой рассылки, определение точки равновесия для рискованных сценариев, назначение вероятностей диагнозам или другим результатам;
* рекомендации: определение продуктов, которые с высокой долей вероятности могут быть проданы вместе, создание рекомендаций;
* определение последовательностей: анализ выбора заказчиков во время совершения покупок, прогнозирование следующего возможного события;
* группирование: разделение заказчиков или событий на кластеры связанных элементов, анализ и прогнозирование общих черт [8,10].

## Обзор алгоритмов

Алгоритм интеллектуального анализа данных — это набор эвристики и вычислений, который создает модель интеллектуального анализа данных из данных. Чтобы создать модель, алгоритм сначала анализирует предоставленные данные, осуществляя поиск определенных закономерностей и тенденций. Алгоритм использует результаты этого анализа для выбора оптимальных параметров создания модели интеллектуального анализа данных. Затем эти параметры применяются ко всему набору данных, чтобы выявить пригодные к использованию закономерности и получить подробную статистику [2, 7].

Кластеризация представляет собой объединение объектов в группы (кластеры) на основе схожести признаков для объектов одной группы и отличий между группами. Большинство алгоритмов кластеризации не опираются на традиционные для статистических методов допущения; они могут использоваться в условиях почти полного отсутствия информации о законах распределения данных. Кластеризацию проводят для объектов с количественными (числовыми), качественными или смешанными признаками. В данной работе рассматривается кластеризация только для объектов с количественными признаками. Исходной информацией для кластеризации является матрица (Х), каждая строчка которой представляет собой значения n признаков одного из Х объектов кластеризации.



Задача кластеризации состоит в разбиении объектов из X на несколько подмножеств (кластеров), в которых объекты более схожи между собой, чем с объектами из других кластеров [2, 4].

Методы кластеризации можно классифицировать на четкие и нечеткие. Четкие методы кластеризации разбивают исходное множество объектов X на несколько непересекающихся подмножеств. При этом любой объект из X принадлежит только одному кластеру. Нечеткие методы кластеризации позволяют одному и тому же объекту принадлежать одновременно нескольким (или даже всем) кластерам, но с различной степенью. Нечеткая кластеризация во многих ситуациях более "естественна", чем четкая, например, для объектов, расположенных на границе кластеров [3, 5, 15].

Методы кластеризации также классифицируются по тому, определено ли количество кластеров заранее или нет. В последнем случае количество кластеров определяется в ходе выполнения алгоритма на основе распределения исходных данных [3, 5].

В данной работе для решения задачи разбиения запросов используется алгоритм четкой кластеризации k-средних [3, 4].

## Анализ существующих решений

В настоящее время существует множество различных сервисов, которые занимаются интеллектуальным анализом больших объемов данных. Наиболее схожим с разрабатываемой системой является платный сервис SpyWords.

Концепция SpyWords – это функциональная платформа по анализу запросов конкурентов в контексте и поиске. Сервис позволяет специалистам по интернет-маркетингу и seo отслеживать полезную и уникальную информацию о сайте конкурентов: ключевые слова в контекстной рекламе и органическом поиске, тексты объявлений, позиции в контексте и органике, дневной бюджет и так далее. Функционал сервиса полезен как на стадии начального формирования семантического ядра для контекста/продвижения, так и на этапе оптимизации действующей компании (поиска новых эффективных слов/анализа активности ключевых конкурентов). Недостатком данного сервиса является то, что в нем отсутствует разделение запросов на регионы и пользователь получает отчеты в неудобном для себя виде [12].

Сервис Quicksprout проводит анализ веб-сайтов на предмет того, как они продвигаются в популярных социальных сетях и насколько активно посетители делятся их контентом. На основе этих данных можно узнать, какие страницы конкурентов пользуются наибольшей популярностью в социальных сетях, какая социальная сеть наиболее популярна в каждой тематике, и делать акцент на продвижение именно в ней. На сервисе можно анализировать свой сайт в сравнении с тремя сайтами конкурентов, причем как главные страницы, так и внутренние. Недостатком сервиса является то, что он англоязычный, и как следствие не поддерживает работу с популярными в России социальными сетями, такими как «Вконтакте» и «Одноклассники»[13].

Также полезным при продвижении сайтов является сервис Ahrefs. С помощью него можно узнать, какие позиции и по каким поисковым запросам занимает та или иная внутренняя страница любого сайта, причем будет отображаться изменение позиций и дата последнего изменения. Ahrefs имеет свою собственную базу данных о сайтах и внешних ссылках , которая ежедневно пополняется. Список появившихся и удаленных внешних ссылок, данные можно просматривать по дням, а также получать ежедневную статистику в удобной для пользователя форме. К недостаткам можно отнести отсутствие русско-язычного интерфейса [14].

# Требования к системе мониторинга и продвижения веб-сайтов

## Интеллектуальная система мониторинга и продвижения веб-сайтов

Интеллектуальная система мониторинга и продвижения веб-сайтов в поисковых системах является расширением системы мониторинга позиций сайтов в поисковых системах и направлена на ускорение и оптимизацию процесса продвижения веб-ресурса, а также анализа его состояния на фоне конкурентов.

Для корректной работы системы необходимы данные о сборе позиций интересующего веб-ресурса за любой период период времени (от одного дня) и список запросов, по которым эта информация собиралась. Запросы могут быть добавлены в систему как самим пользователем, так и специалистом по продвижению сайтов. Сборщик позиций ежедневно обрабатывает данные наиболее популярных поисковых систем, таких, как Яндекс и Google, до глубины 50, 200 или 500 позиций, после чего система анализирует полученную информацию и предоставляет ее в удобном для пользователя виде.

Благодаря системе анализа данных, на вкладке «Конкуренты» пользователю будет доступна информация о других ресурсах, которые занимали наиболее высокие позиции на страницах результатов поиска по интересующим его запросам в течение длительного времени. Также пользователь может увидеть список запросов, по которым каждый конкурент встречается наиболее часто (рис. 1).



**Рис. 1.** Выявление конкурентов

Система мониторинга и продвижения веб-сайтов позволит специалистам по продвижению веб-ресурсов сформировать новые запросы для продвижения и семантическое ядро сайта, на основе которого осуществляется ранжирование большинства поисковых систем.

Для подбора семантического ядра достаточно ввести ключевое слово, и сервис подберет информацию о том, как продвигаются конкуренты. Система обеспечивает анализ уже введенных пользователем запросов, отбирает наиболее эффективные и качественные из них, а также предлагает однородные запросы, по которым конкуренты осуществляют продвижение наиболее успешно.

Все табличные данные можно легко сортировать по возрастанию / убыванию. А содержимое любых отчетов легко экспортируется в популярный формат .odt для дальнейшего анализа.

## Варианты использования системы

Можно выделить следующих основных актеров, взаимодействующих с системой мониторинга и продвижения веб-сайтов:

1. *клиент* – это актер, использующий систему для мониторинга состояния интересующего его веб-ресурса;
2. *специалист SEO* – это клиент, занимающийся продвижением веб-ресурсов, и использующий систему для оптимизации процесса продвижения;
3. *Система мониторинга* – это актер, который обновляет данные, которые обрабатывает система с течением времени;

Можно определить следующие основные варианты использования системы мониторинга и продвижения веб-сайтов (рис. **2**).

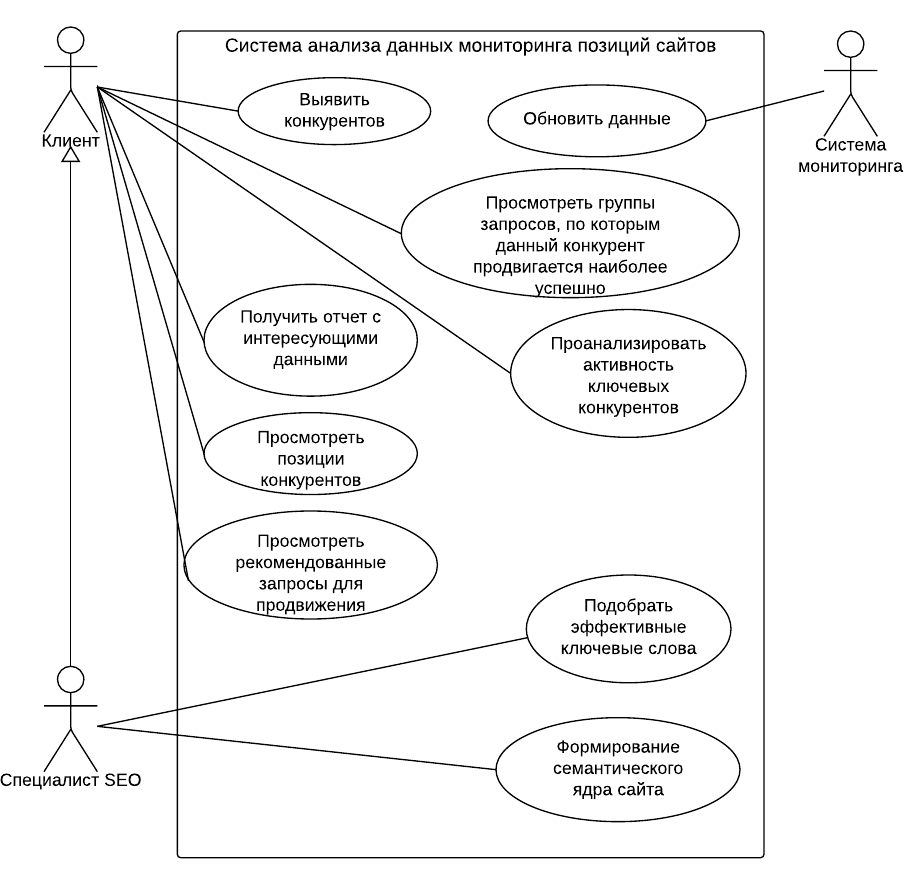


Рис. 2. Варианты использования системы мониторинга и продвижения веб-сайтов

1. Клиент может *выявить конкурентов своего веб-сайта,* то есть просмотреть веб-сайты, которые наиболее часто встречаются на странице результатов поиска по интересующим клиента запросам.
2. Клиент может *просмотреть группы интересующих его запросов,* то есть подмножество из заданных им запросов, по которым каждый из конкурентов занимает высокие позиции на странице результатов поиска в течение длительного времени.
3. Клиент может *проанализировать активность каждого из конкурентов* за заданный период времени, то есть увидеть, как часто и на каких позициях оказывался сайт конкурента в поисковой выдаче по каждому запросу за заданный период времени.
4. Клиент может *просмотреть позиции конкурентов* за заданный период времени по интересующим его запросам.
5. Клиент может *просмотреть запросы*, которые рекомендованы ему для продвижения. На вкладке «Рекомендованные запросы» пользователь видит список запросов, схожих с теми, которые клиент заранее задал в системе.
6. Клиент может *получить отчет о состоянии продвижения его сайта*, в котором также содержится информация о конкурентах. Все отчеты предоставляются клиенту в распространённом формате .odt.
7. Специалист SEO может *подобрать эффективные ключевые слова,* которые помогут в продвижении сайта клиента по заданным запросам в поисковых системах. Клиент вводит ключевое слово и получает список рекомендованных ему ключевых слов для формирования семантического ядра сайта.
8. Специалист SEO может *просмотреть ключевые слова конкурентов*, что значительно облегчит продвижение искомого сайта.
9. Специалист SEO может на основе полученных данных *сформировать семантическое ядро сайта*, которое напрямую влияет на результат ранжирования поисковых систем на странице результатов поиска. Система отображает наиболее успешных конкурентов, их семантическое ядро и предлагает специалисту новые ключевые слова и запросы.
10. Внешняя система мониторинга каждый день *собирает информацию* в поисковых системах *и записывает ее в базу данных*. В базе данных содержатся все веб-ресурсы, доменные имена которых встретились при сборе позиций по конкретному запросу на глубине 50, 200 или 500 позиций.

Перенести в приложение

|  |  |
| --- | --- |
| ***UseCase:* Выявление конкурентов** | ***UseCase:* Просмотр групп запросов** |
| *ID:* 1 | *ID:* 2 |
| *Аннотация:* Выявление конкурентов веб-сайта | *Аннотация:* Просмотр групп запросов, по которым конкуренты наиболее успешны |
| *Главный актер:* Клиент | *Главный актер:* Клиент |
| *Предусловия:* Клиентом в Системе мониторинга:   1. указан веб-сайт для анализа; 2. созданы запросы для анализа; 3. собраны данные результатах поиска по запросам в течение интервала времени | *Предусловия:* Выявлены конкуренты |
| *Описание:*  Клиент просматривает позиции для конкретного ресурса, система определяет часто встречающиеся веб-сайты в данных о сборе для каждого запроса, система выделяет основных конкурентов данного ресурса и группы запросов для каждого конкурента. | *Описание:*  Клиент запрашивает позиции по ресурсу, система на основе проанализированных данных сбора отображает пользователю запросы, по которым каждый конкретный конкурент занимал высокую позицию на странице результатов поиска длительный период времени. |
| *Постусловия:* Данные отображены Клиенту | *Постусловия:* Клиент просматривает группы запросов |

|  |  |
| --- | --- |
| **Просмотр позиций конкурентов** | **Просмотр запросов, рекомендованных для продвижения** |
| *ID:*4 | *ID:*5 |
| *Аннотация:*Просмотр позиций конкурента за заданный период времени по заданным запросам | *Аннотация:*Просмотр группы запросов, по которым рационально продвигать сайт клиента. |
| *Главный актер:* Клиент | *Главный актер:* Клиент |
| *Предусловия:* Выявлены конкуренты, выбран конкретный веб-ресурс для просмотра | *Предусловия:* Заданы интересующие запросы, есть данные о сборе не менее чем за месяц |
| *Описание:*  Клиент запрашивает данные о конкретном конкуренте, система запрашивает у базы данных информацию о сборе для конкретного веб-ресурса по заданным запросам и периоду времени, система отображает данные пользователю в табличном виде. | *Описание:*  Клиент просматривает позиции, система предлагает пользователю список из запросов, однородных уже заданным, по которым рекомендовано осуществлять продвижение. |
| *Постусловия:* Данные отображены пользователю | *Постусловия:* Данные отображены пользователю |

|  |  |
| --- | --- |
| **Получение отчета** | **Подбор ключевых слов** |
| *ID:*7 | *ID:*8 |
| *Аннотация:*Экспорт содержимое любых отчетов в формат .odt для дальнейшего анализа | *Аннотация:*Просмотр группы запросов, по которым рационально продвигать сайт клиента. |
| *Главный актер:* Клиент | *Главный актер:* Специалист по продвижению веб-сайтов |
| *Предусловия:* Выбраны данные для отчета | *Предусловия:* Выявлены конкуренты, заданы имеющиеся ключевые слова |
| *Описание:*  Клиент запрашивает отчет, система конвертирует данные предоставленные пользователю в документ формата .odt, затем создает документ. | *Описание:*  Клиент запрашивает поиск ключевых слов, на основе уже имеющихся ключевых слов и ключевых слов конкурентов, система предлагает пользователю список из ключевых слов, по которым рекомендовано осуществлять продвижение и формировать семантическое ядро сайта. |
| *Постусловия:* Отчет получен пользователем | *Постусловия:* Данные отображены пользователю |

|  |  |
| --- | --- |
| **Формирование семантического ядра сайта** | **Сбор данных** |
| *ID:*9 | *ID:*10 |
| *Аннотация:*На основе запросов пользователя, имеющихся ключевых слова и ключевых слов конкурентов сформировать семантическое ядро сайта | *Аннотация:*Ежедневный сбор данных о позициях веб-сайтов в поисковых системах и запись в БД |
| *Главный актер:* Специалист по продвижению веб-сайтов | *Главный актер:* Система мониторинга |
| *Предусловия:* Выявлены конкуренты, собраны данные | *Предусловия:* Заданы веб-сайты и запросы, по которым необходимо собирать данные в поисковых системах |
| *Описание:*  Клиент запрашивает формирование семантического ядра, на основе уже имеющихся ключевых слов и ключевых слов конкурентов, запросов с высокой частотностью и успешными позициями система предлагает пользователю список из ключевых слов, по которым рекомендовано формировать семантическое ядро сайта. | *Описание:*  Cистема сбора позиций в заданное время суток осуществляет запись в базу данных информации о позициях каждого веб-ресурса, который оказался в поисковой выдаче поисковой системы по конкретному запросу до заданной глубины поиска. |
| *Постусловия:* Данные отображены пользователю | *Постусловия:* Данные записаны в БД |

## Нефункциональные требования к системе

В результате интервью с заказчиком были сформулированы следующие нефункциональные требования.

Поскольку приложение ориентировано на клиентов, которые могут находится в географически удалённых точках, интерфейс системы мониторинга позиций сайтов в поисковой системе Яндекс должен быть доступен из сети Интернет. При этом все страницы веб-интерфейса должны корректно отображаться в следующих браузерах: Google Chrome 33 и выше, включая мобильную версию; Mozilla Firefox 29.0 и выше.

Для повышения удобства использования был выполнен анализ сценариев работы пользователей. Необходимо учесть результаты анализа для максимально удобного расположения основных элементов интерфейса.

Веб-интерфейс к интеллектуальной системе мониторинга и продвижения веб-сайтов в поисковых системах должен быть выполнен с использованием платформы Django, поскольку с ее использованием реализовано большинство проектов заказчика.

## Вывод

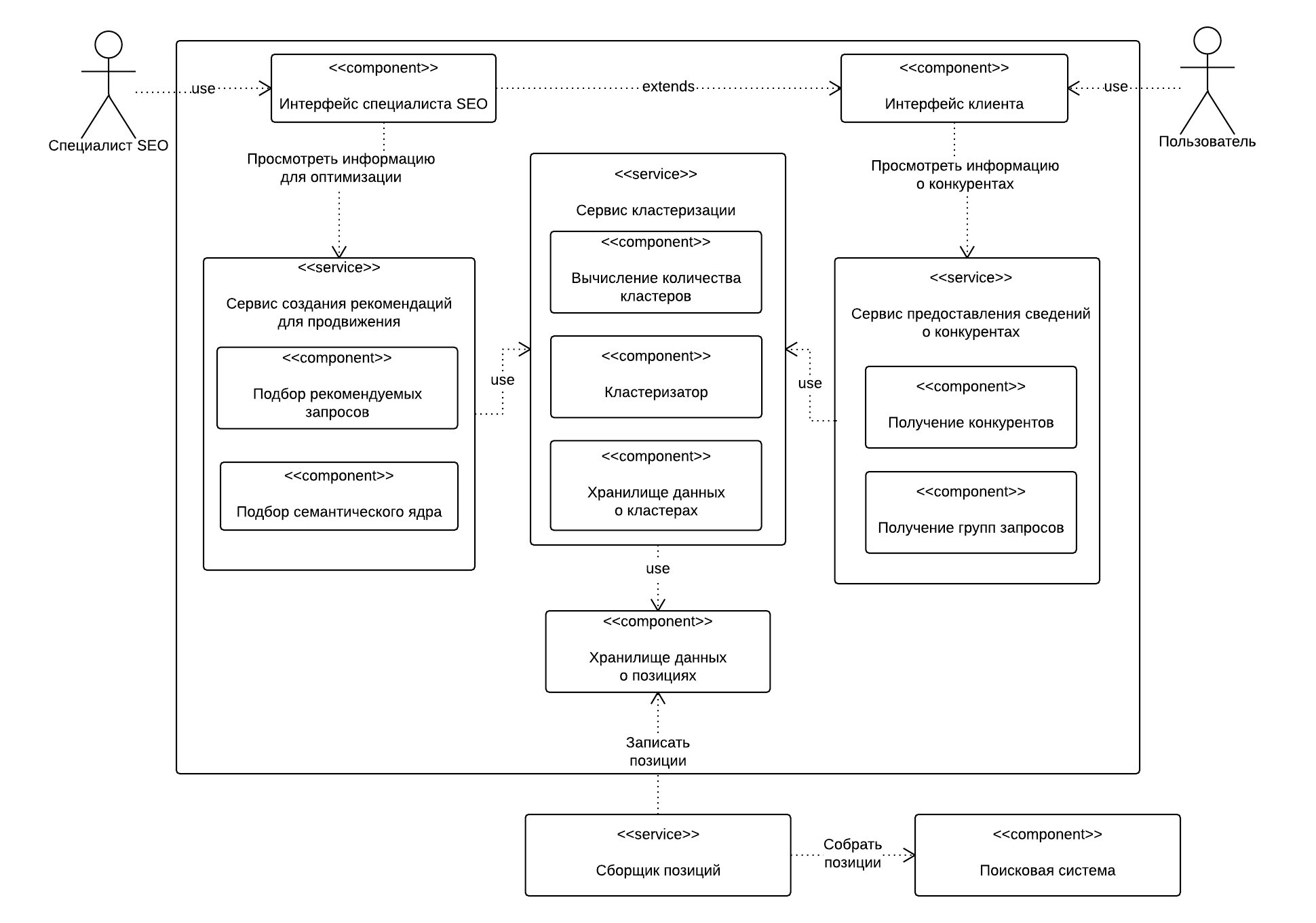
Таким образом, интеллектуальной системе мониторинга и продвижения веб-сайтов в поисковых системах должна быть ориентирована на удобство использования клиентами и легко доступна через сеть интернет на различных устройствах. Также система должна интегрироваться в имеющуюся инфраструктуру заказчика.

Система рассчитана на предоставление пользователям актуальной информации о текущем состоянии интересующих его веб-ресурсов и оказание помощи в продвижении сайтов SEO-специалистам.

# Архитектура системы мониторинга и продвижения веб-сайтов

## Компоненты системы мониторинга

На рисунке 3 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. 3.** 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. 3.3 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.

*Подбор ключевых слов* – система анализа данных хранилища, определяющая ключевые слова для сайта, запрошенного пользователем.

*Подбор семантического ядра* – система анализа данных хранилища, подбирающая семантическое ядро для сайта, запрошенного пользователем.

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

*Интерфейс специалиста SEO* – интерфейс клиента, позволяющий получить информацию для оптимизации сайта.

## 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.



**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;
* 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

Сборщик позиций производит постоянный мониторинг позиций ресурсов в результатах поиска по заданному набору запросов в день . Основной характеристикой отслеживания является максимальная глубина поиска , определяющая максимальное количество позиций, анализируемых сборщиком.

Исходной информацией для алгоритма кластеризации в рамках решаемой задачи является матрица , каждая строка которой представляет собой поисковый запрос, а столбец представляет определенный анализируемый веб-сайт.

Для заполнения ячейки матрицы с номером , отвечающей за усредненную позицию -того сайта по -тому запросу за дней, воспользуемся следующей метрикой:

*.*

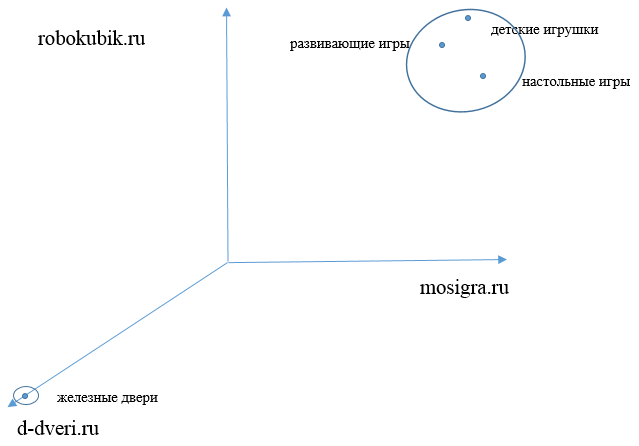
Таким образом, в ячейках матрицы будут располагаться величины, являющиеся разницей между максимальной глубиной поиска и текущим положением сайта по запросу в день . В этом случае, чем выше сайт находится в поисковой выдаче, тем больше будет его вес в матрице по данному запросу.

Рассмотрим пример анализа. Пусть имеются 4 запроса: «детские игрушки», «настольные игры», «развивающие игры», «железные двери» и 3 веб-сайта: «robokubik.ru», «d-dveri.ru», «zapovednikrest.ru». В таблице 1 приведены результаты расчета усредненной метрики для соответствующих сайтов по указанным запросам при глубине поиска . Как видно из таблицы, сайты «robokubik.ru» и «zapovednikrest.ru» присутствуют в TOP-200 результатах по запросам «детские игрушки», «настольные игры», «развивающие игры». Сайт «d-dveri.ru» по данным запросам в 200 первых результатов выдачи не входил, зато он появляется в выдаче по запросу «железные двери».

Таблица 1. Результаты сбора позиций

|  |  |  |  |
| --- | --- | --- | --- |
|  | robokubik.ru | mosigra.ru | d-dveri.ru |
| детские игрушки | 199 | 190 | 0 |
| развивающие игры | 196 | 170 | 0 |
| настольные игры | 168 | 195 | 0 |
| железные двери | 0 | 0 | 180 |

Можно провести кластеризацию представленных данных, для выявления близких по смыслу запросов. Визуальное представление решения задачи кластеризации представлено на рисунке 4. В качестве осей координат на графике используются веса сайтов. Точками на графике отображаются соответствующие запросы.



**Рис. 4.** Пример работы алгоритма кластеризации

Таким образом, из анализа видно, что веб-сайты robokubik.ru и mosigra.ru конкурируют друг с другом по запросам: «детские игрушки», «настольные игры», «развивающие игры», а запрос «железные двери» не входит в их кластер.

Кроме того, в системе предусмотрен алгоритм выявления конкурентов для каждой группы запросов. Для каждого центра кластера упорядочим метрики кластеризации по значению центра в данной метрике. Поскольку каждой метрике соответствует один веб-сайт, мы сможем получить список сайтов, отсортированный по значимости в данном кластере. Выбираются 5 наиболее значимых веб-сайтов-конкурентов.

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 задачи

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

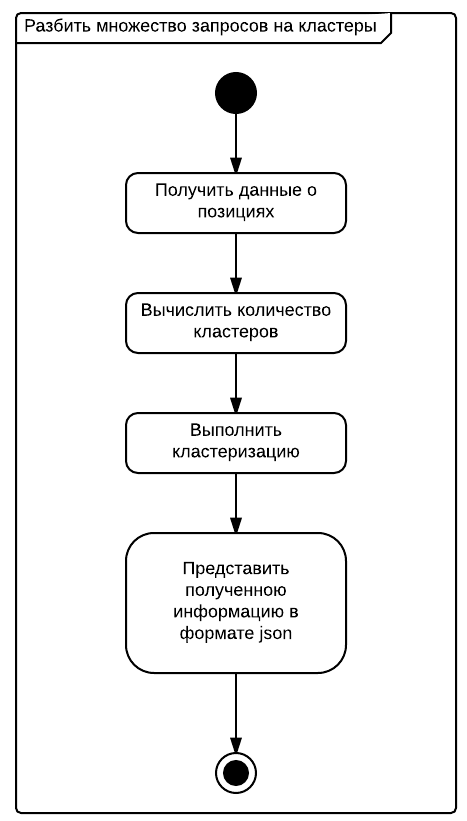
3. Сформировать кластеры. Для каждого центроида из множества отобрать подмножество точек с минимальным расстоянием до

4. Вычислить новые центроиды как среднее всех точек кластера :

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. 5.



**Рис. 5.** Диаграмма деятельности использования алгоритма кластеризации

## Вывод

# Реализация системы – дописать.

## База данных

## Реализация компонентов системы

1. *Сервис кластеризации*

Сервис обеспечивает работу со следующими ресурсами:

* 1. *Кластеры:* конечная точка /clusters. Допустимые действия:
     1. GET /clusters — получить список кластеров – возвращает полученные данные с информацией о принадлежности каждого элемента к определенному кластеру ф формате массива, каждый элемент которого словарь, где идентификатору кластера соответствует степень принадлежности каждого элемента множества этому кластера, например [{id: 1, points: {«1»: 0.1, «2»: 0.1, «3»: 0.9 …}}, {id: 2, points {…}}] .
     2. GET /clusters/<id> — получить информацию о кластере с id в формате словаря, состоящего из идентификатора кластера и степеней принадлежности каждого элемента исходного множества к этому кластеру . Например, GET /cluster/1 -> {id: 1, points: {«1»: 0.1, «2»: 0.1, «3»: 0.9 …}}
  2. *Запросы*, конечная точка /query. Допустимые действия: GET /query — получить список запросов. GET /query/1 — получить информацию о запросе с id 1. GET /query/1 ->{id: 1, positions: {«2015-02-10»: 5, «2015-02-11»: 4, «2015-02-12»: 4 …}}

1. *Сервис сведений о конкурентах*
   1. *Конкуренты*. Конечная точка /resource/1/competitors. Допустимые действия: GET /resource/1/get\_competitors — получить список конкурентов для ресурса с id 1.
   2. *Запросы*. Конечная точка /queries. Допустимые действия: GET /queries?competitor1=5&competitor2=10— получить список запросов для ресурса с id 1 и конкурентов с id=1 или 2.

Для определения оптимального количества кластеров используется CS индекс (рис. 6). Для его расчета вычисляется отношение максимального расстояния между элементами кластера к минимальному расстоянию между кластерами. Оптимальным считается то количество кластеров, при котором значение CS минимально.

def quality(clusters):

max\_clusters\_points = 0

min\_clusters\_dist = 0

for cluster\_center, points in clusters:

for point in points:

max\_point = 0

for point1 in points:

dif = abs\_clust(diff(point1, point))

if dif > max\_point:

max\_point = dif

max\_clusters\_points += max\_point / len(points)

min\_clust = 100000000000000000

for cluster\_center1, points1 in clusters:

dif\_clust = abs\_clust(diff(cluster\_center1, cluster\_center))

if not compare(cluster\_center, cluster\_center1) and (dif\_clust < min\_clust):

min\_clust = dif\_clust

min\_clusters\_dist += min\_clust

return max\_clusters\_points / min\_clusters\_dist

**Рис. 6.** Листингфункции для определения количества кластеров

Для выполнения кластеризации используется конструктор KMeans с параметром n\_clusters=n и метод fit, который разбивает искомый объект на кластеры. Для получения конкурентов каждого кластера метрики сортируются по убыванию относительно значения центра кластера в данной метрике. После этого результат приводится к виду, пригодному для обработки (рис. 7).

def clusterize(data, items, n):

kmeans = KMeans(n\_clusters=n)

kmeans.fit(data)

clusters = []

result = {}

clusters\_for\_quality = []

centers = kmeans.cluster\_centers\_

for c in centers:

srt = range(len(c))

srt.sort(key=lambda i: -c[i])

clusters.append(srt)

clusters\_for\_quality.append((c, []))

for i in range(0, len(kmeans.labels\_)):

if not kmeans.labels\_[i] in result:

result[kmeans.labels\_[i]] = set()

result[kmeans.labels\_[i]].add(items[i])

clusters\_for\_quality[kmeans.labels\_[i]][1].append(data[i])

value = quality(clusters\_for\_quality)

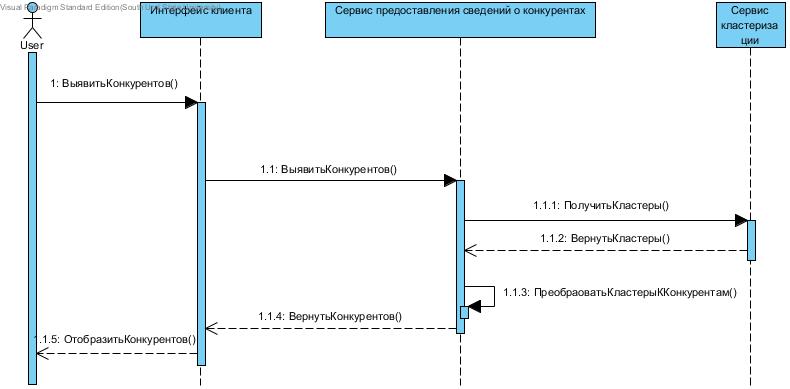
return clusters, result, value

**Рис. 7.** Листингфункции кластеризации

1. *Сервис создания рекомендаций для продвижения*
   1. Ресурсы /resources:
      1. GET /resources – выдает каталог доступных пользователю ресурсов
      2. GET /resources/<id> - выдает данные о ресурсе с id 1
   2. Рекомендуемые запросы:
      1. Конечная точка /resources/1/recommended\_queries. Допустимые действия: GET /resource/1/recommended\_queries — получить список рекомендуемых запросов для ресурса с id 1.
      2. Конечная точка /resource/1/recommended\_keywords. Допустимые действия: GET /resource/1/recommended\_keywords — получить список рекомендуемых ключевых слов для ресурса с id 1.
      3. Конечная точка /resource/1/ semantic\_core. Допустимые действия: GET /resource/1/ semantic\_core — получить список рекомендуемых слов для формирования семантического ядра для ресурса с id 1.

## Организация взаимодействия компонентов системы

Получение информации о конкурентах (рис. 6)

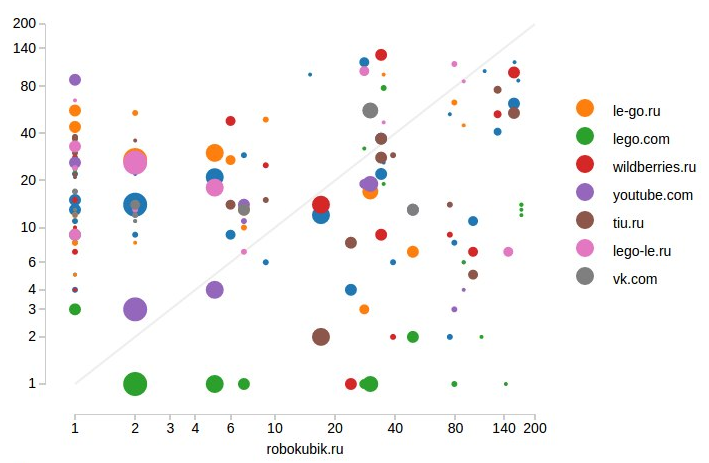


**Рис. 6.** Диаграмма последовательностей для получения конкурентов

## Веб-интерфейс – доработать

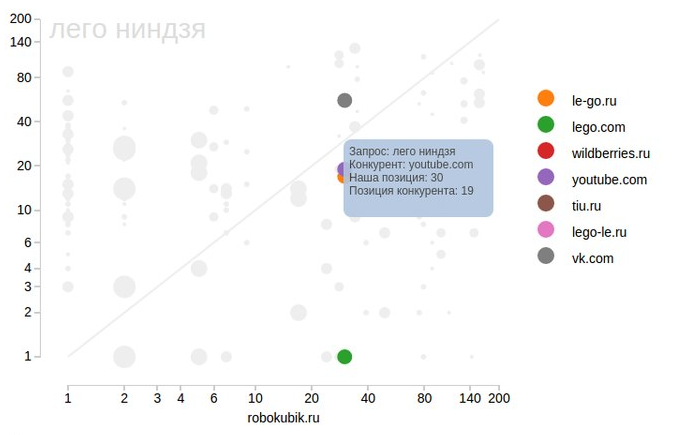
Веб-интерфейс системы поиска конкурентов по группам запросов было решено реализовать с помощью двумерного графика. На рисунке 7 приведен пример работы системы для веб-сайта robokubik.ru. На осях графика отмечены позиции сравниваемых веб-сайтов с использованием логарифмической шкалы. Горизонтальная ось представляет собой позиции искомого сайта, вертикальная – сравниваемых веб-сайтов конкурентов. Точки на графике представляют собой поисковые запросы. Их координаты определяются позицией искомого сайта в поисковой системе по горизонтальной оси и координатой сайта конкурента по вертикальной оси. Цвет точки определяется цветом сайта конкурента. Соответствие цветов веб-сайтам описано в легенде. Каждая точка на графике имеет диаметр, который определяется частотностью запроса в поисковой системе: чем выше частотность, тем больше диаметр точки.

Кроме того, для удобства восприятия, на график добавлена диагональная линия. Точки, оказавшиеся выше диагональной линии, представляют запросы, по которым позиция искомого сайта выше позиции конкурента, причем, чем дальше точка от диагональной линии, тем позиция искомого сайта выше позиции конкурента. То же касается точек, оказавшихся ниже диагональной линии. Они представляют собой запросы, по которым позиция конкурентов превышает позицию искомого сайта. Чем ближе запросы к диагональной линии, тем ближе находятся их позиции в поисковой выдаче.



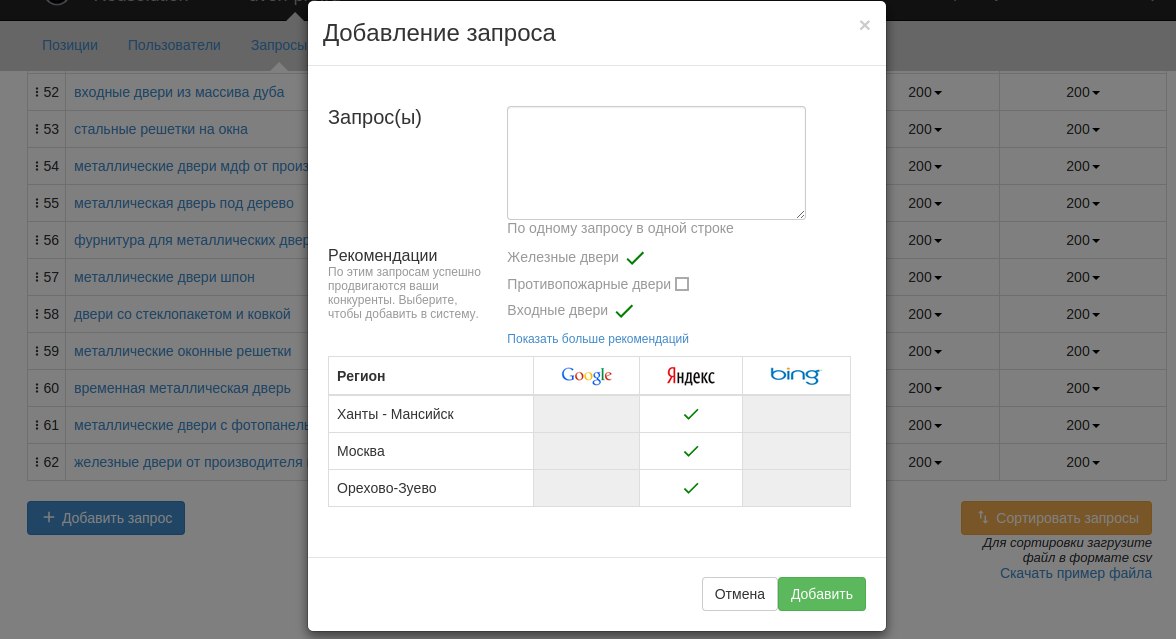
**Рис. 7.** Интерфейс поиска конкурентов

Для сравнения состояния позиций искомого сайта с позициями конкурентов по конкретному запросу необходимо кликнуть на точку на графике, представляющую этот запрос. При этом на графике подсветятся точки, соответствующие выбранному запросу, и пользователь сможет увидеть все позиции конкурентов по этому запросу. Кроме того, пользователь увидит всплывающее окно с текстом запроса, информацией о конкуренте и его позиции по этому запросу в поисковой системе (рис.8).



**Рис. 8.** Интерфейс конкурентов для запроса

Интерфейс создания рекомендаций для продвижения встроен в интерфейс добавления запросов в систему мониторинга позиций сайтов. Пользователю доступно окно ввода, куда он может самостоятельно вписать интересующие запросы. Ниже представлены запросы, которые система определила как рекомендуемые на основе уже добавленных. Пользователь выбрать понравившиеся из видимых или раскрыть список и просмотреть все рекомендованные запросы (рис.9).

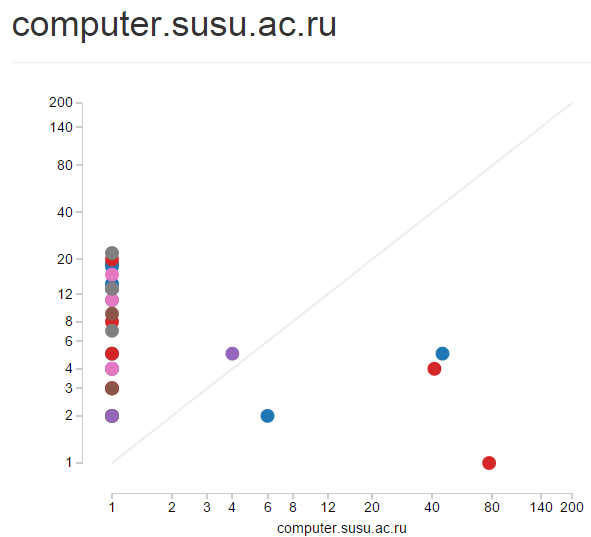


**Рис. 9.** Интерфейс предоставления рекомендаций для продвижения

## Отображение результатов

# Тестирование – сделать

Работа системы была протестирована на 50 проектах, имеющихся в системе. В каждом проекте содержится около 100 запросов. На всех запросах были получены корректные результаты при тестировании сервиса поиска конкурентов и сервиса создания рекомендаций для продвижения. На рисунке 10 представлен результат поиска конкурентов для веб-сайта computer.susu.ac.ru.



**Рис. 10.** Тестирование сервиса поиска конкурентов для сайта computer.susu.ac.ru

В ходе тестирования оказалось, что в результатах, полученных с помощью системы сбора позиций не учитывается разница в доменах второго уровня и ниже. В результате тестирования нами была произведена корректировка работы хранилища сборщика, чтобы не учитывались результаты домена более высокого уровня.

## Реализация алгоритма

Тестовый результат для проекта robokubik.ru  
<http://pastebin.com/Ub41aXRu>

# Заключение

<http://www.quicksprout.com/>

<http://www.ibm.com/developerworks/webservices/library/us-analysis.html>

<http://www.orbitmedia.com/blog/website-competitive-analysis-tools/>

Рекомендации ключевых слов:

http://www.google.com/patents/US20050198068

<http://www.adgooroo.com/>

https://adwords.google.com/KeywordPlanner

<https://www.google.ru/search?q=search+keyword+recommendation>

После изучения существующих решений было принято решение реализовать систему мониторинга позиций сайтов в поисковой системе Яндекс в виде отдельного django-приложения с использованием средств Twitter Bootstrap. Django [8] — это высокоуровневая веб-платформа для языка программирования Python, поощряющий быструю разработку и чистый, прагматичный дизайн. Twitter Bootstrap – это HTML, CSS и Javascript платформа для Web-разработки [9].

**Литература** *объем библиографии*: не менее 15 источников, хотя-бы 5 статей.

* + - 1. Bar-Ilan J., Mat-Hassan M., Levene M. Methods for comparing rankings of search engine results // Comput. Networks. 2006. Vol. 50, № 10. P. 1448–1463.
      2. T.Segaran. Programming Collective Intelligence by Toby Segaran. Copyright 2007 Toby Segaran, 978-0-596-52932-1, pp 29-38
      3. S. Ghosh, S.K.Dubey Comparative Analysis of K-Means and Fuzzy CMeans Algorithms. ((IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013
      4. M. E. Celebi A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications.Volume 40, Issue 1, January 2013, Pages 200–210
      5. M. V. Kiselev, S. M. Ananyan, S. B. Arseniev Principles of Data Mining and Knowledge Discovery Lecture Notes in Computer Science Volume 1704, 1999, pp 366-371
      6. K. Wagstaf, C. Cardie Constrained K-means Clustering with Background Knowledge, Proceedings of the Eighteenth International Conference on Machine Learning, 2001, p. 577–584.
      7. Microsoft Dveloper Network [Электронный ресурс] URL: http://msdn.microsoft.com/ru-ru/library/ms175595.aspx(дата обращения 01.11.2014)
      8. BCGroup. Маркетинговые исследованияи аналитика [Электронный ресурс] URL: http://www.bcgroup.su/analiz-konkurentov (дата обращения 01.11.2014)
      9. Статистика 01.roup.su/analiz-k [Электронный ресурс] URL: http://www.liveinternet.ru/stat/ru/searches.html (дата обращения 28.05.2014).
      10. Определениеtй ресурс] analiz-konkurentovarch engine resu[Электронный ресурс] URL: http://www.semonitor.ru/positioning.html (дата обращения 29.05.2014).
      11. Как работают поисковые системы - SearchEngines.ru . [Электронный ресурс] URL: http://www.searchengines.ru/articles/004556.html (дата обращения 30.05.2014).
      12. SpyWords . [Электронный ресурс] URL: http://spywords.ru/ (дата обращения 01.11.2014)
      13. QuickSpout . [Электронный ресурс] URL: <http://www.quicksprout.com/> (дата обращения 01.11.2014)
      14. Ahrefs. Site Explorer . [Электронный ресурс] URL: <https://ahrefs.com/> (дата обращения 01.11.2014)
      15. Алгоритм нечеткой кластеризации. [Электронный ресурс] URL: http://home.deib.polimi.it/matteucc/Clustering/tutorial\_html/cmeans.html (дата обращения 01.11.2014)