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>  “\_\_\_”\_\_\_\_\_\_\_\_\_\_\_ 2014 г. | ДОПУСТИТЬ К ЗАЩИТЕ  Заведующий кафедрой, д.ф.-м.н., профессор  \_\_\_\_\_\_\_\_\_\_ Л.Б. Соколинский  “\_\_\_”\_\_\_\_\_\_\_\_\_\_\_ 2014 г. |

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

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ЮУрГУ – 010400.62.2014.ХХХХХ1.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  “\_\_\_”\_\_\_\_\_\_\_\_\_\_\_ 2014 г. |

Челябинск-2014

# Оглавление

[Оглавление 2](#_Toc403728954)

[Глоссарий 3](#_Toc403728955)

[Введение 4](#_Toc403728956)

[Структура и объем работы 5](#_Toc403728957)

[1 Обзор литературы 6](#_Toc403728958)

[2 Требования к системе мониторинга и продвижения веб-сайтов 11](#_Toc403728959)

[2.1 Варианты использования системы 11](#_Toc403728960)

[2.2 Функциональные требования к системе 12](#_Toc403728961)

[2.3 Нефункциональные требования к системе 12](#_Toc403728962)

[2.4 Вывод 12](#_Toc403728963)

[3 Архитектура системы интеллектуального анализа 13](#_Toc403728964)

[3.1 Алгоритм 13](#_Toc403728965)

[3.2 Веб-интерфейс 13](#_Toc403728966)

[3.3 Вывод 13](#_Toc403728967)

[4 Реализация системы 14](#_Toc403728968)

[4.1 База данных 14](#_Toc403728969)

[4.2 Реализация алгоритма 14](#_Toc403728970)

[4.3 Отображение результатов 14](#_Toc403728971)

[5 Тестирование 15](#_Toc403728972)

[6 Заключение 16](#_Toc403728973)

# 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 особенности работы с платформами Django и Twitter Bootstrap;
3. изучить существующие алгоритмы интеллектуального анализа данных;
4. спроектировать архитектуру системы интеллектуального анализа данных для мониторинга позиций сайтов в поисковых системах;
5. 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, 6].

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.

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.

В зависимости от задач исследования анализ конкурентного окружения проводится по нескольким направлениям. К основным целям оценки относятся:

* мониторинг долей основных участников рынка;
* составление рейтинга компаний по определенным параметрам;
* составление карт позиционирования (карт занятых конкурентами рыночных ниш);
* анализ цен на товары конкурентов;
* ассортиментный анализ конкурентов .

На основе этих данных можно просчитать успех любой компании. Результаты оценки дают возможность оперативно реагировать на изменения предпочтений потребителей и принимать меры по устранению недостатков стратегий позиционирования и продвижения.

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.

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. 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. 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;

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.

* Модель интеллектуального анализа данных, создаваемая алгоритмом из предоставленных данных, может иметь различные формы, включая следующие.
* Набор кластеров, описывающих связи вариантов в наборе данных.
* Дерево решений, которое предсказывает результат и описывает, какое влияние на этот результат оказывают различные критерии.
* Математическую модель, прогнозирующую продажи.
* Набор правил, описывающих группирование продуктов в транзакции, а также вероятности одновременной покупки продуктов [2].

Для решения бизнес-задач требуется такой анализ данных, который позволяет построить модель для описания информации и в конечном итоге приводит к созданию результирующего отчета.

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 [7].

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»[8].

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 [9].

# 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 длительный период времени (от месяца до года) 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**).

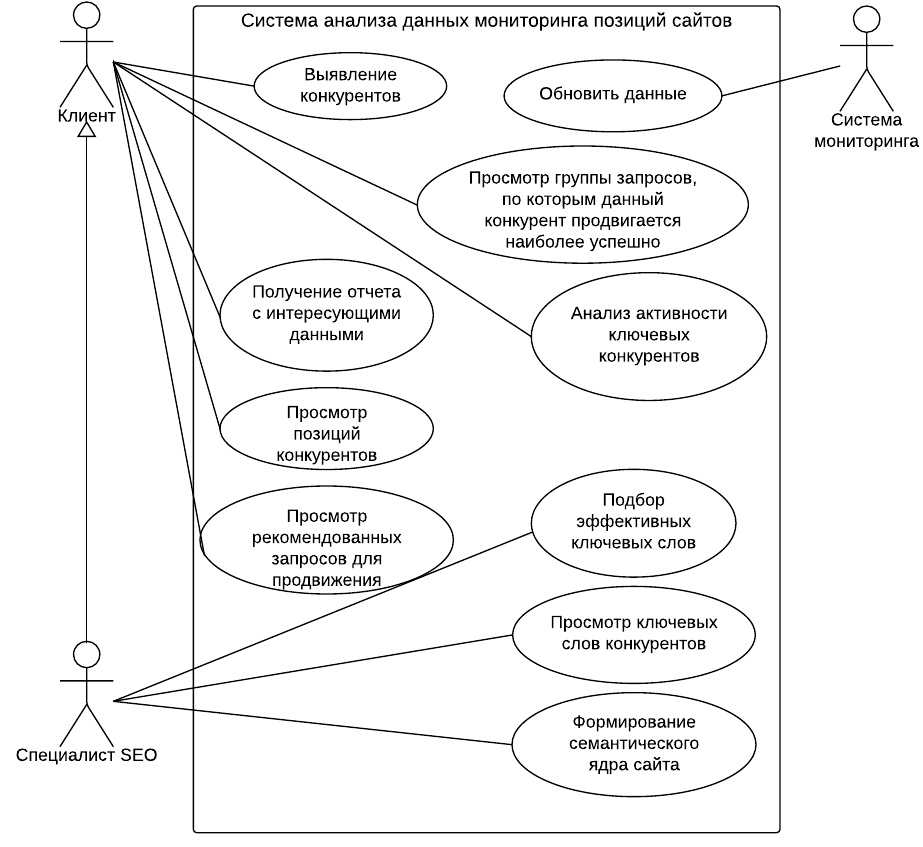


Рис. 2. Варианты использования системы мониторинга и продвижения веб-сайтов

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.

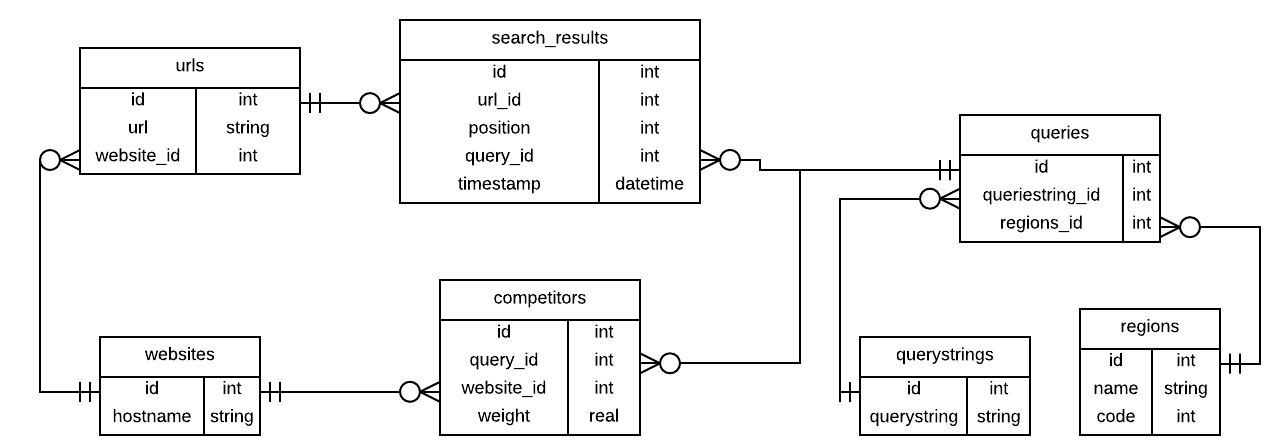
## Вывод

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-cспециалистам.

# Архитектура системы интеллектуального анализа

## База данных



**Рис. 2.** База данных системы

## Алгоритм

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 наблюдений за позициями веб-сайтов (M) в конкретный период времени, каждая строчка которой представляет собой значения n признаков одного из M 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.

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.

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.

### Базовый алгоритм нечетких c-средних

Нечеткие кластера опишем следующей матрицей нечеткого разбиения:

http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image020.gif

в которой k-ая строчка содержит степени принадлежности объекта http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image012.gif к кластерам http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image013.gif. Условия для матрицы нечеткого разбиения записываются так:

|  |  |
| --- | --- |
| http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image022.gif; | (1) |

|  |  |
| --- | --- |
| http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image023.gif. | (2) |

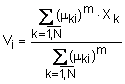
Нечеткое разбиение позволяет просто решить проблему объектов, расположенных на границе двух кластеров - им назначают степени принадлежностей равные 0.5. Недостаток нечеткого разбиения проявляется при работе с объектами, удаленными от центров всех кластеров. Удаленные объекты имеют мало общего с любым из кластеров, поэтому интуитивно хочется назначить для них малые степени принадлежности. Однако, по условию (1) сумма их степеней принадлежностей такая же, как и для объектов, близких к центрам кластеров, т.е. равна единице. Для устранения этого недостатка можно использовать возможностное разбиение, которое требует, только чтобы произвольный объект из X принадлежал хотя бы одному кластеру. Возможностное разбиение получается следующим ослаблением условия (1):

http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image024.gif.

Для оценки качества нечеткого разбиения используется такой критерий разброса:

|  |  |
| --- | --- |
| http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image025.gif, | (3) |

где  -



-центры нечетких кластеров;

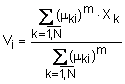
http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image027.gif - экспоненциальный вес, определяющий нечеткость кластеров.

### Алгоритм нечетких c-средних

Шаг 1. Установить параметры алгоритма: c - количество кластеров; m - экспоненциальный вес; http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image028.gif - параметр останова алгоритма.

Шаг 2. Случайным образом сгенерировать матрицу нечеткого разбиения http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image021.gif, удовлетворяющую условия (1) - (2).

Шаг 3. Рассчитать центры кластеров:

, http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image004.gif.

Шаг 4. Рассчитать расстояния между объектами из X и центрами кластеров: http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image029.gif, http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image030.gif, http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image004.gif.

Шаг 5. Пересчитать элементы матрицы нечеткого разбиения (http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image004.gif, http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image030.gif):

Если  http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image031.gif: ;

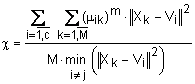
Если  http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image033.gif: http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image034.gif, http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image035.gif.

Шаг 6. Проверить условие http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image036.gif, где http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image037.gif - матрица нечеткого разбиения на предыдущей итерации алгоритма. Если "да", то перейти к шагу 7, иначе  - к шагу 3.

Шаг 7. Конец.

В приведенном алгоритме самым важным параметром является количество кластеров. Существует два формальных подхода к выбору числа кластеров.

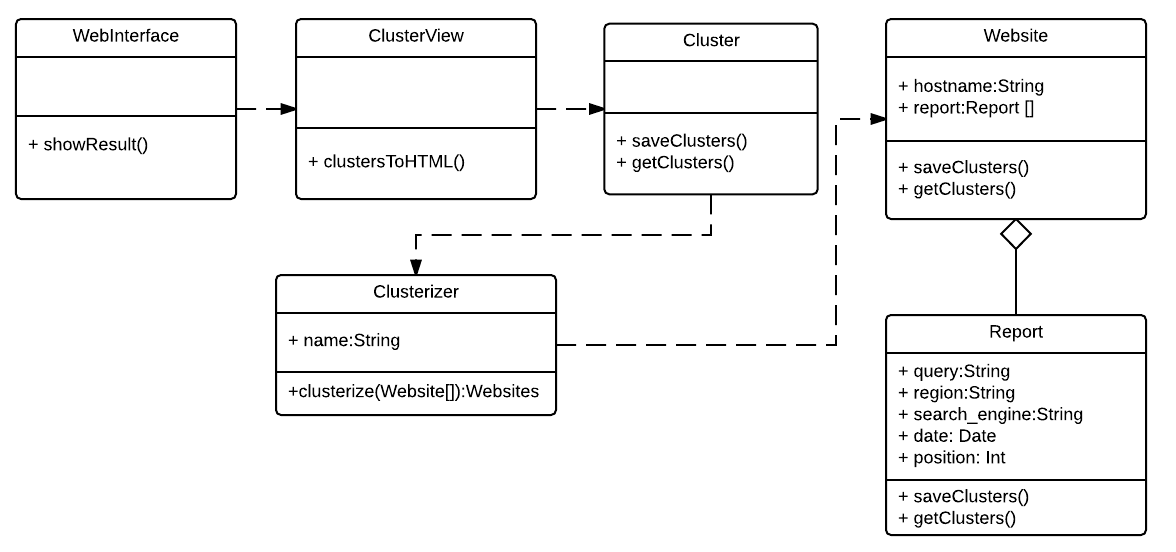
Первый подход основан на критерии компактности и разделимости полученных кластеров. Логично предположить, что при правильном выборе количества кластеров данные будут разбиты на компактные и хорошие отделимые друг от друга группы. Для алгоритма нечетких c-средних в рекомендуется использовать индекс Хие-Бени:

.

Второй подход предлагает начинать кластеризацию при достаточно большом числе кластеров, а затем последовательно объединять схожие смежные кластера. При этом используются различные формальные критерии схожести кластеров.

Вторым параметром алгоритма кластеризации является экспоненциальный вес (m). Чем больше http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image039.gif, тем конечная матрица нечеткого разбиения http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image021.gif становится более "размазанной", и при http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image040.gif она примет вид http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image041.gif, что является очень плохим решением, т. к. все объекты принадлежат ко всем кластерам с одной и той же степенью. Кроме того, экспоненциальный вес позволяет при формировании координат центров кластеров усилить влияние объектов с большими значениями степеней принадлежности и уменьшить влияние объектов с малыми значениями степеней принадлежности. На сегодня не существует теоретически обоснованного правила выбора значения экспоненциального веса. Обычно устанавливают http://matlab.exponenta.ru/fuzzylogic/book1/images_12/image042.gif.

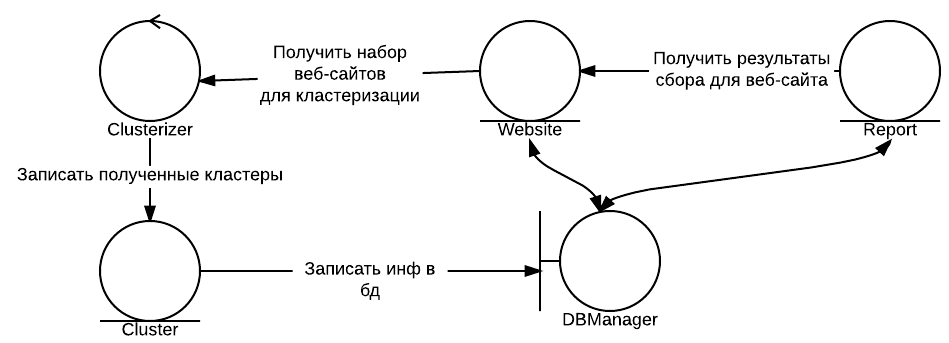
Диаграмма классов системы представлена на рисунке 2.



**Рис. 3.** Диаграмма классов интеллектуальной системы

* Website содержит информацию о собранных позициях, связанных с веб-сайтом, которую получает из базы данных
* Report представляет собой одну запись со страницы результатов поиска в конкретной поисковой системе по следующим параметрам: строка запроса, регион, дата сбора. Содержит доменное имя веб-сайта и позицию, занимаемую им на странице результатов поиска по данному запросу.
* Cluster содержит содержит набор веб-сйтов, которые интеллектуально сгруппированны по заранее заданным признакам
* ClusterView - класс, преобразовывающий данные в формат, отображаемый пользователю
* WebInterface - класс, обрабатывающий действия пользователя и возвращающий html-страницы в качестве ответа.

Диаграмма взаимодействия компонентов системы представлена на рисунке 3.



**Рис. 3.** Взаимодействие компонентов системы

Веб-интерфейс обрабатывает запрос, полученный от пользователя и запрашивает требуемую страницу у ClusterView. ClusterView запрашивает кластеры у модели Cluster. Для создания кластеров используется класс кластеризатора, который с помощью интеллектуального иерархического алгоритма кластеризации разбивает имеющиеся у него объекты класса веб-сайт на кластеры по заранее заданным параметрам. Данные для обработки класс получает из сущности Website. Website хранит в себе список объектов типа Report, которые содержат в себе информацию о соборе позиций по данному веб-сайту.

## Веб-интерфейс

## Вывод

# Реализация системы

## База данных

## Реализация алгоритма

## Отображение результатов

# Тестирование

# 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**

* + - 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. Microsoft Dveloper Network [Электронный ресурс] URL: http://msdn.microsoft.com/ru-ru/library/ms175595.aspx(дата обращения 01.11.2014)
      3. BCGroup. Маркетинговые исследованияи аналитика [Электронный ресурс] URL: http://www.bcgroup.su/analiz-konkurentov (дата обращения 01.11.2014)
      4. Статистика 01.roup.su/analiz-k [Электронный ресурс] URL: http://www.liveinternet.ru/stat/ru/searches.html (дата обращения 28.05.2014).
      5. Определениеtй ресурс] analiz-konkurentovarch engine resu[Электронный ресурс] URL: http://www.semonitor.ru/positioning.html (дата обращения 29.05.2014).
      6. Как работают поисковые системы - SearchEngines.ru . [Электронный ресурс] URL: http://www.searchengines.ru/articles/004556.html (дата обращения 30.05.2014).
      7. SpyWords . [Электронный ресурс] URL: http://spywords.ru/ (дата обращения 01.11.2014)
      8. QuickSpout . [Электронный ресурс] URL: <http://www.quicksprout.com/> (дата обращения 01.11.2014)
      9. Ahrefs. Site Explorer . [Электронный ресурс] URL: <https://ahrefs.com/> (дата обращения 01.11.2014)