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. |

**Интеллектуальный анализ данных системы 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](#_Toc403661147)

[Глоссарий 3](#_Toc403661148)

[Введение 4](#_Toc403661149)

[Структура и объем работы 5](#_Toc403661150)

[1. Обзор литературы 7](#_Toc403661151)

[2. Требования к системе интеллектуального анализа данных 13](#_Toc403661152)

[2.1. Система интеллектуального анализа данных мониторинга и продвижения веб-сайтов 13](#_Toc403661153)

[2.2. Варианты использования системы 13](#_Toc403661154)

[2.3. 15](#_Toc403661155)

[2.4. Функциональные требования к системе 16](#_Toc403661156)

[2.5. Нефункциональные требования к системе 17](#_Toc403661157)

[2.6. Вывод 18](#_Toc403661158)

[3. Архитектура системы интеллектуального анализа 19](#_Toc403661159)

[3.1. Алгоритм 20](#_Toc403661160)

[3.2. Веб-интерфейс 21](#_Toc403661161)

[3.3. Вывод 22](#_Toc403661162)

[4. Реализация системы 23](#_Toc403661163)

[4.1. База данных 24](#_Toc403661164)

[4.2. Реализация алгоритма 25](#_Toc403661165)

[4.3. Отображение результатов 26](#_Toc403661166)

[5. Тестирование 27](#_Toc403661167)

[6. Заключение 28](#_Toc403661168)

# Глоссарий

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

*Организация* — группа людей, деятельность которых сознательно координируется для достижения общих целей.

*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.

# 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].

Можно выделить следующих основных актеров, взаимодействующих с платформой ВИС:

1. *пользователь ВИС* – это актер, которому предоставляется удаленный доступ к ресурсам платформы ВИС посредством открытого API, обеспечивающего постановку и запуск виртуальных экспериментов;
2. *конечный пользователь* – это пользователь ВИС, которому обеспечивается удаленный веб-доступ к ресурсам платформы ВИС для проведения виртуальных экспериментов и удаленной визуализации результатов ВЭ (в соответствии с моделью SaaS);
3. *разработчик ВИС* – это актер, которому предоставляется удаленный доступ к ресурсам платформы ВИС в соответствии с моделью PaaS, для разработки ВИС;
4. *разработчик ПВС* – это актер, которому предоставляется удаленный доступ к ресурсам платформы ВИС в соответствии с моделью PaaS, для разработки внедрения прикладных программных пакетов в платформу в виде прикладных вычислительных сервисов;
5. *поставщик прикладного ПО –* этоактер, предоставляющий лицензионное прикладное ПО, обеспечивающее выполнение прикладных вычислительных сервисов;
6. *поставщик вычислительных ресурсов –* это актер, являющийся владельцем вычислительных ресурсов, на базе которых производится развертывание платформы ВИС;
7. *администратор платформы* – это актер, обеспечивающий развертывание, настройку и поддержку работы платформы ВИС на базе доступных вычислительных ресурсов.

Можно определить следующие основные варианты использования платформы ВИС (рис. 12).

1. Пользователь ВИСдолжен *зарегистрироваться и авторизоваться* в платформе для получения доступа к ресурсам платформы. Процесс регистрации может заключать в себе верификацию пользователя, для определения его прав по доступу к тем или иным группам виртуальных экспериментов.
2. Пользователь ВИС должен иметь доступ к каталогу доступных виртуальных экспериментов и иметь возможность *выбора виртуального эксперимента* для запуска. Доступ к тем или иным виртуальным экспериментам должен регулироваться правами доступа конкретного пользователя или группы пользователей.
3. Пользователь ВИС должен иметь возможность *запустить выбранный виртуальный эксперимент.* Для запуска эксперимента, пользователь ВИС должен установить значения всех входных параметров выбранного ВИС, загрузить исходные данные для проведения эксперимента (если таковые требуются) и инициировать процесс проведения ВЭ. Вычислительные ресурсы для проведения виртуального эксперимента должны быть выделены платформой автоматически из пула доступных вычислительных ресурсов.
4. Пользователь ВИС должен иметь возможность *получения результатов вычислительного эксперимента*. Потребитель сервисов может загрузить результаты эксперимента на локальный компьютер, либо запросить удаленный доступ к ним для удаленной визуализации и анализа результатов.
5. Конечный пользователь, в рамках процесса получения результатов вычислительного эксперимента, может запросить предоставление *удаленной визуализации результатов эксперимента*. Пользователю будет предоставлен интерфейс для интерактивного анализа и визуализации результатов, хранящихся на удаленных ресурсах платформы ВИС, без необходимости их загрузки на локальный компьютер.

# Требования к системе интеллектуального анализа данных

# Система интеллектуального анализа данных мониторинга и продвижения веб-сайтов

# Варианты использования системы

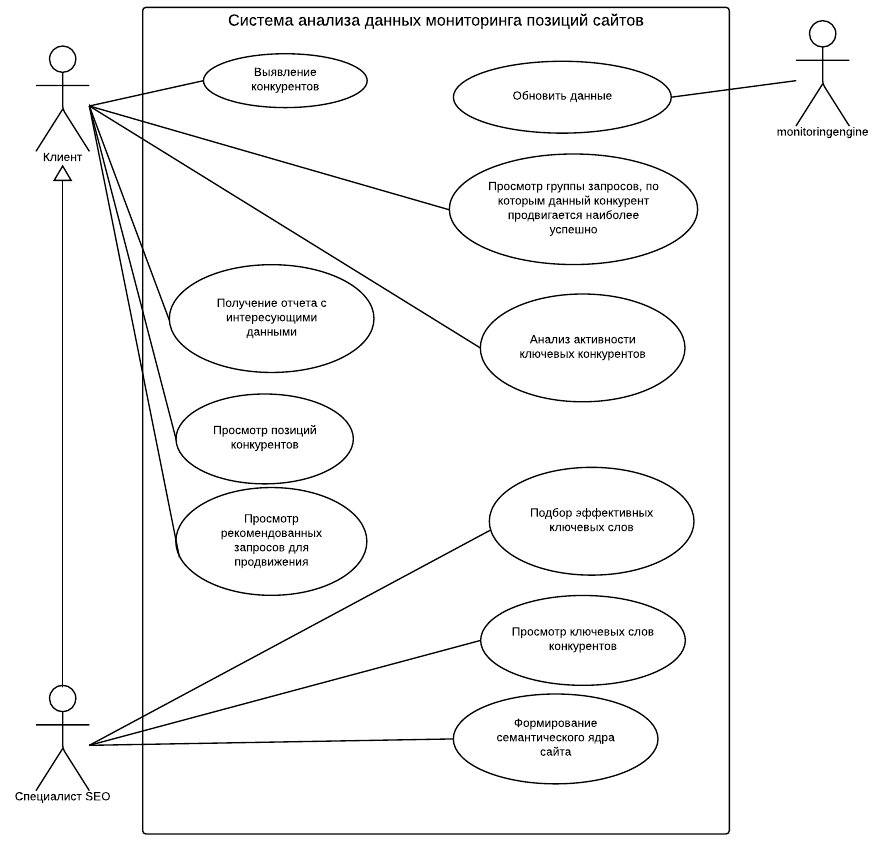


Рис. 1. Варианты использования системы интеллектуального анализа данных мониторинга позиций сайтов

1. Клиент может *выявить конкурентов,* наиболее успешно продвигающих свой продукт в схожей области.
2. Клиент может *просмотреть группы интересующих его запросов,* по которым каждый конкурент продвигается наиболее успешно.
3. Клиент может *проанализировать активность каждого из конкурентов* за заданный период времени.
4. Клиент может *просмотреть позиции конкурентов* за заданный период времени.
5. Клиент может *просмотреть запросы*, которые рекомендованы ему для продвижения.
6. Клиент может *получить отчет о состоянии продвижения его сайта*, в котором также содержится информация о конкурентах.
7. Специалист SEO может *подобрать эффективные ключевые слова,* которые помогут в продвижении сайта клиента по заданным запросам в поисковых системах.
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.
10. Внешняя система monitoringengine каждый день *собирает информацию* в поисковых системах *и записывает ее в базу данных*.

# 

# Функциональные требования к системе

# Нефункциональные требования к системе

# Вывод

# Архитектура системы интеллектуального анализа

# Алгоритм

# 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 поисковых систем . [Электронный ресурс] URL: http://www.liveinternet.ru/stat/ru/searches.html (дата обращения 28.05.2014).
      5. Определение позиций сайта в поисковиках Яндекс и Гугл . [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)