Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.42

Шарков
Дмитрий Сергеевич

Быстрый полнотекстовый поиск в проектах с высокой нагрузкой и большим объемом данных

АВТОРЕФЕРАТ
магистерской диссертации на соискание степени
магистра технических наук

по специальности 1-40 81 01 – Информатика и технологии разработки программного обеспечения

Научный руководитель
Лихачев Д.С.
кандидат технических наук, доцент

Минск 2017

ВВЕДЕНИЕ

Высоконагруженные сайты и корпоративные системы, использующие полнотекстовый поиск, получили широкое распространение. В наше время уже много чего придумано и сделано для поиска. Например, поисковые системы, как google, yandex, yahoo!, bing, работают очень быстро даже несмотря на то, что они обрабатывают неисчисляемое количество сайтов и данных пользователей. Несмотря на большое количество информации, они должны также обрабатывать миллионы запросов одновременно от разных пользователей. Вследствии таких высоких требований на протяжении нескольких лет создавали и улучшали алгоритмы, позволяющие быстро и эффективно искать данные на основании сложного, а чаще всего составного, запроса.

Не новость также, что некоторые системы используют поиск не только по ключевым словам, а использоваться внутри программного обеспечения для своих нужд. Необходимо не забывать о рациональном использовании поиска, потому как внедрить его не всегда и легко.

Таким образом в данной диссертации рассмотрены уже существующие алгоритмы поиска, приведен сравнительный анализ существующих поисковых движков, реализовано программное обеспечение, характеризующее для выбранных конкурирующих алгоритмов качество, скорость, удобство использования, а также легкость в настройке и сопровождении. Из полученных результатов выбран наилучший алгоритм, указаны его индивидуальные характеристики, используя которые можно повысить скорость работы поиска. На основании алгоритма рассмотрено решение поставленной задачи, а также усовершенствован алгоритм с учетом индивидуальных характеристик поставленной задачи, приведен сравнительный анализ усовершенствованного алгоритма.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

В разработке программного обеспечения одним из важнейших требований является обеспечение качества и скорости программы, решающей поставленные бизнес задачи. Полнотекстовый поиск используется очень часто, в следствии чего необходимо осмысленно подходить к выбору того или иного алгоритма поиска и, таким образом, поискового движка. Стандартные алгоритмы поиска по базе данных не обладают высокой скоростью, из-за чего не подходят для высоконагруженных фреймворков. Таким образом, приходится искать что-то новое или реализовывать свое.

На момент написания диссертационной работы рассматриваемые поисковые движки являются наиболее востребовательными в сфере информационных технологий. Приведенная характеристика и подробное описание существующих поисковых движков может помочь при решении задачи. Более подробно приведено описание работы ElasticSearch и его анализаторов, которые влияют на производительность поиска.

Степень разработанности проблемы

В настоящий момент существует достаточно много методов поиска и уже разработанных поисковых движков. Разработан достаточно мощный инструмент, получивший широкое распространение – ElasticSearch. На его основе было разработано достаточно много программ, реализующие поиск для конкретных продуктов. Уже разработано достаточно большое количество инструментов и библиотек для поиска, однако единицы пытаются улучшить алгоритм поиска для конкретной задачи.

Цель и задачи исследования

Целью диссертации является разработка программного обеспечения на основе поискового движка, а также его усовершенствование.

Для выполнения поставленной цели в работе были сформулированы следующие задачи:

- изучить предметную область поисковых методов быстрого полнотекстового поиска
- изучить и проанализировать существующие актуальные методы поиска
- привести сравнительный анализ существующих методов
- для выбранного поискового движка подробно описать индивидуальные характеристики
- реализовать поэлементный поиск
- реализовать поиск при помощи ElasticSearch и привести сравнительный анализ с поэлементным поиском
- улучшить алгоритм поиска ElasticSearch и привести сравнительный анализ со стандартным поиском ElasticSearch

Объектом исследования является программное обеспечение полнотекстового поиска, применяемого на проектах с высокой нагрузкой и большим объемом данных.

Предметом работы выступают методы быстрого полнотекстового поиска на проектах с высокой нагрузкой и большим объемом данных.

Область исследования. Содержание диссертационной работы соответствует образовательному стандарту высшего образования второй ступени (магистратуры) специальности специальности 1-40 81 01 – «Информатика и технологии разработки программного обеспечения».

Теоретическая и методологическая основа исследования

В основу диссертации легли результаты известных программистов и инженеров по программированию в области поисковых систем и методов поиска.

Для получения теоретических результатов исследования рассматривались

наиболее быстрые и широко используемые среди разработчиков поисковые движки, такие как Sphinx, Apache Solr, Apache Lucene, Xapian, MySQL fulltext, PostgreSQL Textsearch.

Были проведены исследования по изучению поисковых алгоритмов и их реализации, на базе которого был разработан инструмент ElasticSearch.

Информационная база исследования сформирована на основе существующих решений в области поисковых алгоритмов.

Научная новизна диссертационной работы заключается в разработке и анализе самых быстрых на сегодняшний день поисковых движков, которые необходимо использовать при больших объемах данных или высокой нагрузке, так как обычные поисковые алгоритмы будут работать очень долго.

Основные положения, выносимые на защиту

1. Анализ существующих поисковых движков. Представлено подробное описание их возможностей, а также выбран наиболее быстрый и подходящий для решения задачи поисковый движок.

2. Представлены функциональные и нефункциональные требования к поисковому движку, подробно описаны достоинства, настройка, алгоритм работы ElasticSearch.

3. Приведена подробная характеристика приведенных поисковых движков. Разработан посимвольный поиск, а также поиск при помощи инструмента ElasticSearch. Усовершенствован алгоритм поиска с учетом индивидуальных характеристик поставленной задачи.

Теоретическая значимость диссертации заключается в том, что в ней рассмотрены наиболее востребованные существующие поисковые движки и приведено подробное сравнение наиболее важных для поиска и хранения данных характеристик.

Практическая значимость диссертации состоит в том, что на основе рассмотренных поисковых движков, приведенной характеристики разработано программное обеспечение, позволяющее наиболее эффективно и быстро искать необходимые данные.

Апробация и внедрение результатов исследования

Результаты исследования были представлены на III Международной научно-практической конференции “Современные технологии: актуальные вопросы, достижения и инновации”.

Отдельные положения диссертации, в частности разработанный поиск при помощи ElasticSearch был успешно внедрен в один из настоящих проектов.

Публикации

Основные положение работы и результаты диссертации изложены в одной опубликованной работе общим объемом 10 п.л. (авторский объем 10 п.л.).

Структура и объем работы. Структура диссертационной работы обусловлена целью, задачами и логикой исследования. Работа состоит из введения, трёх глав и заключения, библиографического списка и приложений. Общий объем магистерской диссертации − 76 страниц, включая 6 иллюстраций, 2 приложения и библиографический список из 31 наименования.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассмотрено современное состояние проблемы обеспечения качества и скорости программы, определены основные направления исследований, а также дается обоснование актуальности темы диссертационной работы.

В **общей характеристике работы** сформулированы ее цель и задачи, показана связь с научными программами и проектами, даны сведения об объекте исследования и обоснован его выбор, представлены положения, выносимые на защиту, приведены сведения о личном вкладе соискателя, апробации результатов диссертации и их опубликованность, а также, структура и объем диссертации.

В первой главе магистерской диссертации представлено описание наиболее быстрых и широко используемых поисковых движков, таких как Sphinx, Apache Solr, Apache Lucene, Xapian, MySQL fulltext, PostgreSQL Textsearch, описано их применение и краткое описание по настройке и использованию, а также критерии выбора подходящего движка для проекта с высокой нагрузкой и большим объемом данных. Были сделаны выводы о приведенных поисковых движках и сделан выбор в пользу ElasticSearch для решения поставленной задачи.

Во второй главе магистерской диссертации кратко описаны функциональные и нефункциональные требования поискового движка, подробно рассмотрено основные возможности ElasticSearch, основанном на Apache Lucene, описана его настройка, алгоритм работы, а также анализаторы, при помощи которых можно улучшить поиск для поставленной задачи.

Третья глава магистерской диссертации посвящена сравнительному анализу существующих поисковых движков, реализации програмного обеспечения, характеризующего для выбранных конкурирующих алгоритмов качество, скорость, удобство использования, а также легкость в настройке и сопровождении. На основании алгоритма рассмотрено решение поставленной задачи, а также

усовершенствован алгоритм с учетом индивидуальных характеристик поставленной задачи, приведен сравнительный анализ усовершенствованного алгоритма.

В **приложениях** приведена реализация индексирования и поиска при помощи ElasticSearch.

ЗАКЛЮЧЕНИЕ

1. Рассмотрены наиболее быстрые и широко используемые системы полнотекстового поиска, такие как Sphinx, Apache Solr, Apache Lucene, Xapian, MySQL fulltext, PostgreSQL Textsearch. Приведено их применение и краткое описание по настройке и использованию.

2. Проанализированы алгоритмы поиска, работа с ними, а также представлен подробный вывод об упомянутых выше поисковых системах, на основании которого был выбран наилучший из них для решения поставленной задачи.

3. Учитывая поставленную задачу кратко приведены агрегированные функциональные и нефункциональные требования для поискового алгоритма, представлены его ключевые возможности, а также рассмотрена работа анализаторов, при помощи которых можно улучшить поиск для поставленной задачи.

4. Приведен подробный сравнительный анализ существующих поисковых алгоритмов, характеризующий для выбранных конкурирующих алгоритмов качество, скорость, удобство использования.

5. Разработано программное обеспечение на основе поискового движка ElasticSearch, приведен сравнительный анализ решения с поэлементным перебором на примере базы данных MS SQL.

6. Учитывая индивидуальные характеристики поставленной задачи и анализатора, представлен усовершенствованный алгоритм для поиска. Приведен сравнительный анализ усовершенствованного алгоритма с предыдущими.

Список опубликованных работ

1. Шарков Д.С. Быстрый поиск на проектах с высокой нагрузкой и большим объемом данных. // Управление информационными ресурсами: материалы III Международной научно-практической конференции “Современные технологии: актуальные вопросы, достижения и инновации”. Пенза, 23 ноября 2016г. / Наука и Просвещение (ИП Гуляев Г.Ю.), Пенза 2016, ISBN: 978-5-9909306-7-4. с. 23-32