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

УДК 004.4`242+004.94

Мыц
Сергей Игоревич

Автоматизированный анализ расхождений значений параметров на больших совокупностях временных данных

АВТОРЕФЕРАТ
на соискание академической степени
магистра технических наук

по специальности 1-40 80 04 – Математическое моделирование, численные методы и комплексы программ

Подпись магистранта

Научный руководитель
Волорова Н. А.
к.т.н., доцент

Подпись руководителя

Минск 2015

КРАТКОЕ ВВЕДЕНИЕ

Объемы доступной рядовому человеку информации за последние десятилетия увеличились на много порядков. Это произошло благодаря развитию технологий по её передаче, хранению и обработке. Интернет является самым мощным источником информации для простого обывателя.

В современном мире перед человеком стоит задача уметь ориентироваться в огромном количестве существующих в глобальной информационной сети данных. Только различных веб-сайтов насчитывается уже несколько миллиардов. И это лишь информация в текстовой форме, кроме неё большое значение имеет мультимедиа: изображения, видео, аудио.

Эту задачу призваны помогать решать поисковые системы, и делать они должны это быстро и качественно. Информационный поиск такого рода и таких масштабов является технически и технологически сложной, наукоемкой деятельностью. В этом направлении работают сотни тысяч ученых-исследователей, программистов, аналитиков и инженеров по всему миру.

В процессе исследования способов улучшения качества поисковой системы используется изучение пользовательского поведения на страницах показа результата запросов. Из-за того, что объемы соответствующих данных велики, они требуют для работы с собой специальных навыков, подходов и инструментов.

Данная работа посвящена выбору подхода и созданию инструмента, призванного упростить и автоматизировать задачу исследования причин измерения числовых величин, характеризующих состояние поисковой системы.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цели и задачи исследования

Целью диссертационной работы является создание инструмента для анализа особенностей поведения отдельных категорий пользователей информационной системы.

Для достижения поставленной цели необходимо решить следующие задачи:

- Исследовать подходы к анализу причин изменений значений параметров.
- Разработать методы и алгоритмы на основе найденных подходов.
- Разработать архитектуру программного средства анализа причин изменений значений параметров.
- Реализовать программное средство, включая реализацию и тестирование корректности работы созданного инструмента.

Объектом данного исследования является характеристики процессов взаимодействия агентов.

Предметом исследования являются связи и зависимости между значениями характеристик целого и отдельных частей.

Личный вклад соискателя

Личный вклад автора диссертации состоит в предложении модифицированного метода анализа причин расхождений параметров, предложения алгоритмов для эффективной реализации математической модели, а также построении программного средства, использующего математическую модель для расчёта необходимый научных результатов. Вклад научного руководителя Н.А. Волоровой заключается в помощи в формулировке целей и задач исследования, содействии в подборе и анализе научной литературы по исследуемой проблеме, консультации и обсуждению возникавших проблем и вопросов.

Апробация результатов диссертации

Основные положения диссертационной работы докладывались и обсуждались на 51-ой научной конференции аспирантов, магистрантов и студентов БГУИР (Минск, Республика Беларусь, 2015), а также на

международной научно-практической конференции BIG DATA and Predictive Analytics (Минск, Республика Беларусь, 16-19 июня 2015 года).

Опубликованность результатов диссертации

По теме диссертации опубликована 2 печатных работы: в сборнике трудов и материалов 51-ой научной конференции аспирантов, магистрантов и студентов БГУИР и в сборнике материалов международной научно-практической конференции BIG DATA and Predictive Analytics (Беларусь, Минск, 2015).

Структура и объем диссертации

Диссертация состоит из общей характеристики работы, введения, шести глав, заключения, списка используемых источников, списка публикаций автора. В первой главе представлено описание предметной области. Вторая глава посвящена описанию проблемы и постановке задачи. Третья глава содержит теоретические детали в основе разрабатываемого инструмента. В четвертой главе дано описание библиотек и инструментов положенных в основу практической реализации. Пятая глава содержит описание основных частей реализованного инструмента. В шестой главе описан способ проверки корректности работы созданного инструмента.

Общий объем работы составляет 32 страницы, из которых основного текста – 26 страниц, список используемых источников из 10 наименований на 1 странице.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

В **первой главе** дано описание основных понятий, необходимых для понимания предметной области и дальнейшего содержания работы. Введены понятия информационного и веб поиска. Описаны механизмы использующиеся для анализа состояния и качества работы поисковой системы. На основе этих механизмов далее будет построен теоретический подход к решению задачи

Вторая глава посвящена описанию проблемы и постановки практической задачи. Как в ходе проводимых AB экспериментов, так и во время обычной работы системы иногда возникают значимые изменения одной из наблюдаемых метрик. Это может быть вызвано рядом причин: технические неполадки, сезонные изменения, а также просто последствия применяемых в рамках эксперимента модификаций системы. В любом случае, после возникновения изменений существует необходимость поиска всестороннего анализа для поиска их причин: технические неполадки необходимо локализовать и устранять, а закономерные изменения — понять, исследовать и задокументировать. Перед исследующим проблему аналитиком предстает сложная, крупная и местами рутинная задача разбора конкретной ситуации.

В **третьей главе** приводится описание базового подхода, в основе которого лежит использование анализа чувствительности для исследования специальным образом выбранной функции. Затем на его основе вводится альтернативное решение лучше удовлетворяющее практическим требованиям. В качестве такого альтернативного метода к базовому подходу можно выбрать применение деревьев решений.

Они способны реализовать оба требуемых пункта:

1. Построить функцию
2. Численно охарактеризовать зависимость значения от аргументов

Одним из запланированных направлений дальнейшего развития является попытка реализации базового подхода и некоторых его вариаций для дальнейшего сравнения с альтернативным на открытом случайно сгенерированном наборе данных.

Четвертая глава содержит набор технологий и библиотек использованных для разработки инструмента, в т.ч. система распределенных вычислений *MapReduce,* инструмент организации цепочек вычислений с зависимостями *REM,* обертка над *REM* с добавленным слоем кеширования результатов, программа реализующая градиентный бустинг над решающими деревьями под названием *Matrixnet*.

В **пятой главе** дается описание основных структурных частей инструмента и детали их реализации. Структура итогового инструмента совпала с предполагаемой:

1. Периодический расчет и создание ежедневных выжимок данных.
2. Агрегация выжимок и извлечение информации необходимой для работы инструмента.
3. Непосредственный расчет необходимых результатов для ответа на запрос.
4. Интерфейс для выполнения запросов к инструменту и просмотра результатов.

Шестая глава посвящена описанию подхода к проверки корректности работы разработанного инструмента. Для проведения тестирования работы инструмента был применен следующий подход:

1. Выбрано несколько примеров расхождений метрик в прошлом, для которых известны причины.
2. На основе причин к каждому из примеров был составлен список срезов, которые наиболее точно, по мнению аналитика, отражали детали проблемы.
3. На выбранных примерах был запущен инструмент и получены результаты ранжирования срезов.
4. Проведено сравнение заранее выбранных срезов с топ-5 срезов из тех, что выдал инструмент

ЗАКЛЮЧЕНИЕ

В начале выполнения этой работы была поставлена задача по разработке вспомогательного инструмента для анализа причин изменений числовых параметров поисковой системы.

Первым шагом был выбор базового подхода к проблеме и изучение его особенностей. В результате этого был сделан вывод о необходимости внесения корректировок в метод исходя из требований практического применения.

Был разработан альтернативный вариант решения задачи, который, с одной стороны, использует основные идеи из базового подхода, а с другой — лучше соответствует требованиям практики и более приспособлен для использования в основе производственного инструмента.

Основываясь на результатах разработки альтернативного подхода был создан инструмент решающий поставленную в диссертации задачу. Он отказоустойчиво и распределённо выполняет ежедневную предобработку данных, а также, в ответ на запросы к нему, в течение часа предоставляет результаты анализа за большие промежутки времени.

Создание этого инструмента позволило автоматизировать работу аналитиков над отдельным классом исследовательских задач.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Мыц С.И. Автоматизированный анализ расхождений значений параметров на больших совокупностях временных данных / С.И. Мыц // Компьютерные системы и сети: материалы 51-ой научной конференции аспирантов, магистрантов и студентов.

2. Мыц С.И. Автоматизированный анализ изменений значений параметров на больших совокупностях данных / С.И. Мыц // BIG DATA and Predictive Analytics: сборник материалов международной научно-практической конференции (Минск, Республика Беларусь, 2015).