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

УДК 004.4

Ковалёв
Станислав Павлович

Анализ поведения пользователей сайтов и приложений при помощи машинного обучения

АВТОРЕФЕРАТ
на соискание степени
магистра технических наук

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

Научный руководитель
Калугина М.А.
к.ф.-м.н., доцент

Минск 2019

КРАТКОЕ ВВЕДЕНИЕ

В настоящее время существует огромное количество сайтов, сервисов, мобильных и десктопных приложений. Каждый год появляются новые приложения, развиваются и изменяются старые, некоторые приложения исчезают. Удобный процесс обмена идеями через интернет позволяет постоянно придумывать идеи для новых приложений, в которых заинтересованы пользователи.

Каждое приложение имеет уникальный функционал, постоянных пользователей, заинтересованных в этом функционале и использовании приложения. Приложение помогает им либо решать повседневные задачи, либо оптимизировать выполнение работы, либо развлекать.

Мир быстро меняется, люди начинают интересоваться новыми приложениями, им может не доставать функционала в текущем приложении, на рынке появляются новые приложения-конкуренты и т.д. Таким образом, с одной стороны, перед владельцами приложения стоит задача удержать пользователей в своем приложении, предложить им необходимые возможности. С другой стороны, у владельцев приложения может возникать желание увеличить прибыль от приложения в случае, если нет нужды заботиться удержанием пользователей.

Эти и множество других проблем помогают решать сервисы веб-аналитики. Они позволяют следить за посещаемостью ресурса, находить самые заинтересованные в продукте группы пользователей, статистику посещения страниц. Большое количество сайтов и приложений создается небольшими компаниями, которые не могут позволить себе команду бизнес аналитиков. Таким компаниям выгодно использовать сервис веб-аналитики.

Сервисы веб-аналитики предоставляют возможность построения разнообразных графиков о поведении пользователей. Чтобы выявить закономерности и сделать некоторые выводы о поведении пользователей в приложении необходимо провести большое количество времени за изучением этих графиков.

Одним из таких графиков является сегментация пользователей по некоторым признакам, таким, например, как возраст, национальность, пол. С помощью этого графика выявляется поведение в приложении людей из определенных слоев населения. Часто используемыми являются графики удержания пользователей и воронки поведения пользователей, которые помогают определить удачные и неудачные изменения в приложении, а также места в приложении, которые не удовлетворяют пользователей.

Проведение анализа, таким образом, требует много времени и сил, и нет гарантии, что в итоге будут достигнуты какие-либо результаты. В таком случае можно обратиться к машинному обучению. ЭВМ способна обрабатывать сотни тысяч записей за несколько секунд, анализируя поведение пользователей, находя закономерности в их поведении.

Интересным графиком в смысле анализа поведения пользователей является сегментация. Но обычная сегментация по стандартным признакам

(возраст, пол, социальное положение, доход) не позволяет правильно разделять пользователей на группы, а лишь позволяет понять, как в среднем ведет себя данная группа. В то время как часть группы может вести себя в приложении совсем в другом ключе, чем остальная часть.

Необходимо делить пользователей на группы не по национальности/возрасту/полу, а по их конкретному поведению в приложении. Простые результаты, которые каждую неделю будут давать информацию о пользователях, об изменении их поведения, о группах пользователей, которые начинают переставать пользоваться сервисом.

Поэтому необходимо делать анализ с помощью машины. При этом не будет траты времени на изучение графиков. Результаты анализа будут присылаться владельцам приложения для дальнейшей работы с ними.

Все это позволяет сделать вывод, что задача разработки алгоритма анализа поведения пользователей сайтов и приложений является актуальной задачей с научной и практической точки зрения.

Исследования в диссертации проведены на базе сервиса веб-аналитики, ранее реализованного мной, который позволяет сохранять действия пользователей сайтов и приложений и анализировать их поведение с помощью построения различных типов графиков.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является разработка алгоритма для анализа поведения пользователей приложений при помощи машинного обучения (МО).

Для достижения поставленной цели необходимо решить следующие *задачи*:

1. Изучить предметную область, связанную с целью работы, провести анализ подходов и алгоритмов МО для анализа поведения пользователей приложений;
2. Создать архитектуру системы, которая позволит реализовать необходимый алгоритм;
3. Разработать алгоритм для анализа поведения пользователей приложений;
4. Реализовать разработанный алгоритм с использованием выбранных подходов и методов МО.

Объектом исследования является разработка и реализация алгоритма с использованием методов и алгоритмов МО для анализа поведения пользователей приложений.

Предметом исследования является сервис веб-аналитики с функционалом, позволяющим анализировать поведение пользователей приложений.

Личный вклад соискателя

Все изложенные в диссертации результаты исследования получены соискателем лично с учетом рекомендаций и замечаний научного руководителя.

Опубликованность результатов диссертации

По теме диссертации опубликовано 2 работы, из них 1 статья в научном журнале «Студенческий» №18(62), 1 тезис на 55-ой юбилейной научной конференции аспирантов, магистрантов и студентов БГУИР.

Структура и объем диссертации

Структура диссертационной работы обусловлена целью, задачами и логикой исследования. Состоит из введения, общей характеристики работы, трех глав, заключения, списка использованных источников, списка публикаций автора и приложений.

Общий объем работы составляет 61 страницу, из которых основного текста – 42 страницы, включая 14 иллюстраций и 4 формулы, список использованных источников из 38 наименований на 3 страницах и 4 приложения на 12 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ

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 проведен анализ предметной области, поиск и изучение подходов машинного обучения, которые могут быть использованы для анализа поведения пользователей сайтов и приложений. Выполнен анализ применяемых методов и алгоритмов обработки данных, выявлены их достоинства и недостатки. Итогом является выбор и описание выбранных подходов и алгоритмов.

Машинное обучение включает в себя множество различных семейств методов, они отличаются своими достоинствами при применении к некоторым задачам и невозможностью их применения к другим типам задач.

По способам решения задачи разделяют на обучение «с учителем» и «без учителя». Обучение «с учителем» требует существования ответов, с помощью которых алгоритм будет обучаться давать ответ. Обучение «без учителя» объединяет задачи, выявляющие описательные модели, например, закономерности в покупках, совершаемых клиентами большого магазина.

Для решения поставленных задач решено использовать методы обучения «без учителя», такие как кластеризация и анализ главных компонент.

Вторая глава посвящена разработке архитектуры программного обеспечения, подготовке тестовых данных для последующего анализа.

Для демонстрации работоспособности алгоритма анализа использован сгенерированный мной набор метрик поведения пользователей в системе управления проектами небольших групп. Использование подразумевает регистрацию, вход в аккаунт, открытие главной страницы, досок с карточками, создание карточек и т.д. При генерации этого набора были заданы два типа пользователей: корпоративные и персональные.

В процессе разработки системы были использованы следующие инструменты:

– Python, как язык разработки;

– Django, широко распространенный веб-фреймворк для разработки сайтов, веб-приложений, API на языке python;

– PostgreSQL в качестве сервера базы данных;

– Библиотека машинного обучения Scikit-learn;

– Библиотека Pandas;

– Библиотека Matplotlib.

В третьей главе описывается алгоритм решения поставленной задачи. Проведено исследование различных алгоритмов кластеризации, описаны подходы для улучшения качества результатов. Алгоритм действий представлен на рисунке 1.

Рисунок 1 – Алгоритм анализа поведения пользователей

Во время выполнения разработанного алгоритма происходит выборка информации из базы данных и её подготовка к дальнейшему анализу, применение кластерного анализа для разбиения пользователей на группы, описание всех полученных групп, поиск изменений в группах в сравнении с предыдущей неделей, генерация отчета по результатам проведенного анализа.

Первым шагом необходимо подсчитать поведенческие метрики для каждого пользователя. При этом активно используется библиотека pandas. Для этого нужно запросить все события пользователей из базы данных за последнюю

неделю. Полученные записи о действиях подвергаются обработке таким образом, чтобы итоговая выборка содержала информацию о различных поведенческих метриках, таких как:

1. сумма всех событий по дням недели;
2. сумма всех событий в будние/выходные дни;
3. время первого события по дням недели утром, и среднего за будние/выходные;
4. время последнего события по дням недели вечером, и среднего за будние/выходные;
5. число каждого отдельного события за неделю.

Второй шаг представляет собой процесс разбиения пользователей на группы. Но для того чтобы выбрать наиболее подходящий алгоритм действий при кластеризации, было проведено исследование для выбора наиболее подходящего алгоритма кластерного анализа и методов по улучшению результатов кластеризации.

В процессе этого исследования было проведено:

1. сравнение трех алгоритмов кластеризации (*k*-средних, агломеративная кластеризация с методом одиночной связи, агломеративная кластеризация с методом Ward) на качество результатов и быстродействие работы;
2. исследование возможности улучшения результатов путем уменьшения размерности данных при помощи анализа главных компонент PCA;
3. исследование возможности применения алгоритма DBSCAN и фильтрации аномальных объектов для улучшение результатов кластеризации.

Результаты исследования текущего пункта опубликованы в работах [1-А, 2-А].

В результате таких исследований был предложен следующий алгоритм действий для поиска необходимого разбиения. Наглядно алгоритм представлен на рисунке 2, шаги описаны далее:

1. Выделить 3 главные компоненты в исходном наборе с помощью PCA;
2. Передать эти данные на вход методу DBSCAN для фильтрации аномальных значений;
3. После того как значения отфильтрованы, провести агломеративную кластеризацию методом Ward для различного числа кластеров – от 2 до 8;
4. Выбрать разбиение с наилучшим значением метрики «силуэт»;
5. Добавить отфильтрованные ранее объекты в ближайшие кластеры.

Следующим этапом является создание описания для каждой из групп – необходимо выделить наиболее значимые признаки группы. Это параметры, значения которых наиболее существенно отличаются от соответствующих значений других групп.

Далее происходит сравнение результатов анализа текущей недели с результатами прошлой. Этот шаг позволяет узнать изменения в

пользовательских характеристиках, позволит заметить тенденции в изменении характеристик или числа пользователей различных групп.

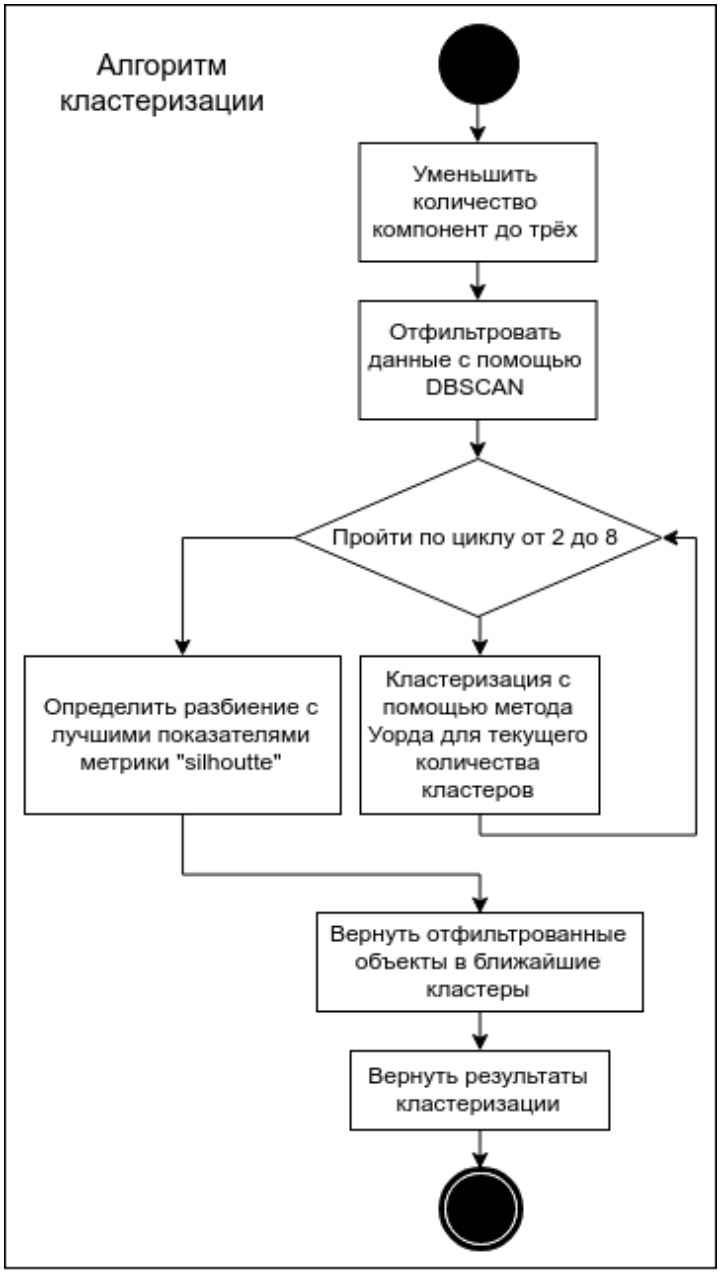

Рисунок 2 – Алгоритм кластеризации

Полученные результаты анализа текущей недели, полученные группы и их характеристики сохраняются в базу данных для последующего анализа.

Последним шагом является создание отчёта о результатах проведенного анализа. По итогам предыдущих действий система располагает данными, которые необходимо показать в отчёте, нужно лишь подготовить их для отображения и передать на вход шаблонизатору Django, который, используя существующий в системе шаблон, сгенерирует итоговый отчет.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Разработан алгоритм проведения анализа поведения пользователей с помощью МО. В ходе работы алгоритма происходит обработка действий пользователей за последнюю неделю, разделение пользователей на различные группы по поведенческим характеристикам, описание признаков полученных групп, поиск изменений в группах по сравнению с предыдущей неделей.
2. Разработан подход для улучшения эффективности кластеризации данных путем использованием анализа главных компонент (PCA). Предложен алгоритм сравнения работоспособности данного подхода в зависимости от количества выделяемых компонент.
3. Разработан подход для улучшения результатов кластерного анализа путем предварительного использования алгоритма кластеризации DBSCAN для фильтрации аномальных объектов в исходной выборке [1-А, 2-А].

Рекомендации по практическому использованию результатов

1. Полученные результаты формируют теоретическую и практическую базу для разработки ПО для решения задачи анализа поведения пользователей, а точнее, поиска групп пользователей с помощью кластерного анализа, описания найденных групп, нахождения изменения поведения различных групп пользователей. Это позволяет использовать разработанные алгоритмы в автоматическом режиме в системах веб-аналитики, что помогает уменьшить время проведения анализа человеком и отказаться от взаимодействия с графиками различных типов.
2. Разработанные методы и алгоритмы анализа поведения пользователей могут применяться в автоматизированных системах веб-аналитики.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1-А. Ковалёв С.П., Калугина М.А. Комбинация алгоритмов K-средних и DBSCAN при анализе поведения пользователей // Студенческий: электрон. научн. журн. 2019. № 18(62). URL: https://sibac.info/journal/student/62/140295 (дата обращения: 15.05.2019).

2-А. Ковалёв С. П. Использование алгоритма кластеризации DBSCAN для фильтрации выбросов в данных / Ковалёв С. П. // 55-я юбилейная научная конференция аспирантов, магистрантов и студентов. - 2019. - № 4(20). - с. 198-200.