Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра математической
кибернетики и компьютерных наук

РАЗРАБОТКА СЕРВИСА УВЕДОМЛЕНИЙ И РЕКОМЕНДАЦИЙ ТОВАРОВ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИИ DATA MINING

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 411 группы
направления 02.03.02 — Фундаментальная информатика и информационные технологии
факультета КНиИТ
Шевченко Антона Сергеевича

Научный руководитель
старший преподаватель _____ 10.06.2017 М. И. Сафрончик

Заведующий кафедрой
к.ф.-м.н. _____ 10.06.2017 С. В. Миронов

Саратов 2017

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

При реализации товаров и услуг важно учитывать предпочтения целевой аудитории для того чтобы рекомендовать им наиболее полезные для них товары и делать индивидуальные предложения о скидках, тем самым увеличивая объем продаж. Таким образом, одним из ключевых моментов маркетинга является создание списка клиентов, заинтересованных в конкретном товаре или услуге. Использование технологии Data Mining позволит повысить эффективность маркетинговых мероприятий в несколько раз за счет анализа поведения пользователей и выявления их предпочтений.

Целью данной бакалаврской работы является реализация сервиса автоматических Push–уведомлений и рекомендации товаров посетителям крупных интернет магазинов с применением технологии Data Mining. Ставится задача разработать функционал для анализа получаемых данных о просмотре страниц товаров пользователями с использованием нескольких алгоритмов Data Mining и создать на основе результатов этого анализа блок рекомендаций для каждого товара на сайте. Данная система позволяет, как привлекать внимание покупателей, за счёт ненавязчивых уведомлений о просматриваемых товарах, так и почти полностью исключает необходимость ручного добавления рекомендуемых товаров для каждой страницы товара в интернет магазине.

Для разработки используются следующие программные стредства: Python, PHP, JavaScript, jQuery, HTML 5, CSS 3, Bootstrap 3, дистрибутивы развёрнуты на Virtual Machine, СУБД Postgre SQL, программа для написания кода NetBeans IDEA.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Термин Data Mining в сфере компьютерных технологий означает «извлечение информации» или «добыча данных». Цель Data Mining — поиск скрытых правил и закономерностей в очень больших объемах данных. Ведь дело в том, что разум человека не приспособлен к восприятию большого количества разнообразной информации. Человек, в большинстве своем, не способен осмысливать более трех взаимосвязей даже в небольших объемах информации. В то же время, традиционная статистика, которая являлась основным инструментом анализа данных, потеряла свою мощь и популярность при решении, допустим, задач из реальной жизни. Дело в том, что в данном подходе используются усредненные характеристики выборки и методы математической статистики являются эффективными в основном для проверки сформулированных гипотез, но определение еще не открытой гипотезы часто бывает непосильной задачей [1].

Data Mining имеет в своём составе несколько методов анализа информации, описывающих тип анализа и операцию восстановления данных.

1. **Ассоциация** является наиболее известным и простым методом интеллектуального анализа данных. Для распознавания моделей используется простое сопоставление нескольких элементов одного типа. Например, анализирую покупки клиентов магазинов, можно заметить тот факт, что в дополнении клубники покупают сливки.
2. **Классификация** позволяет описывать несколько атрибутов класса и идентифицировать каждый класс на основе выделенных атрибутов. Данный метод можно использовать, например, для получения информации о типе объекта. Допустим транспорт можно легко классифицировать по типу кузова (легковые, грузовые и т.д.) используя информацию об имеющихся атрибутах (размер, форма кузова и т.д.). Классификацию можно использовать для других методов в качестве входных данных.
3. **Кластеризация** позволяет объединить отдельные элементы в общую группу элементов, после чего получить подробное и структурированное заключение. При кластеризации на простом уровне используются атрибуты в качестве основных для определения кластера похожих результатов.
4. **Прогнозирование** в сочетании с остальными методами Data Mining поз-

воляет прогнозировать и анализировать тенденции, классификацию и сопоставлять модели с отношениями. То есть проанализировав прошлое, можно попытаться предсказать будущее.

5. **Последовательный метод** полезен для выявления тенденций.
6. **Дерево решений** представляет собой дерево, состоящее из листьев и веток и каждый узел которого имеет только два исхода. Прохождение каждого из узлов позволяет отнести объект к одному из заранее заданных классов.
7. **Метод классификация-кластеризации** использует кластеризацию для нахождения ближайших соседей и позволяет уточнить классификацию [2].

Каждый из методов Data Mining имеет определенный набор свойств, каждое из которых играет большую роль при выборе метода для анализа данных. Методы можно сравнивать друг с другом путем оценки характеристики их свойств.

Основными свойствами и характеристиками методов Data Mining являются:

— гибкость;
— популярность;
— быстрота;
— проверяемость;
— трудоемкость;
— точность;
— интерпретируемость;
— масштабируемость.

Масштабируемость — одно из свойств вычислительной системы, обеспечивающее предсказуемый рост характеристик системы (при добавлении к ней вычислительных мощностей), таких как:

— производительность;
— скорость;
— надежность.

В таблице 1 показано сравнение характеристик некоторых методов [3].

Таблица 1 – Сравнительная характеристика методов.

Алгоритм	Точность	Масштаб	Интерпретируемость
Линейная регрессия	нейтральная	высокая	высокая
Нейронные сети	высокая	низкая	низкая
Методы визуализации	очень низкая	высокая	высокая
Деревья решений	низкая	высокая	нейтральная
K-ближайшего соседа	высокая	нейтральная	нейтральная
Алгоритм	Быстрота	Трудоемкость	Популярность
Линейная регрессия	нейтральная	нейтральная	нейтральная
Нейронные сети	низкая	нейтральная	низкая
Методы визуализации	высокая	высокая	высокая
Деревья решений	высокая	высокая	нейтральная
K-ближайшего соседа	нейтральная	нейтральная	нейтральная

Каждая из характеристик оценена по следующей шкале:

— очень низкая;
— низкая;
— нейтральная;
— средняя;
— высокая.

Из таблицы 1 можно заметить, что каждый метод обладает своими сильными и слабыми сторонами. Но ни один из методов не может обеспечить решение всех задач Data Mining. Большинство предлагаемых в сети интернет инструментов для анализа Data Mining реализуют сразу несколько методов. В таких статистических пакетах, как SPSS или SAS существует широкий спектр разнообразных методов (статистических и кибернетических). Но для возможности их использования и интерпретации результатов работы статистических методов требуются необходимые знания в области статистики и анализа [4].

Универсальность каждого из инструментов довольно часто приводит к выявлению некоторых ограничений в его возможностях. Однако положительной стороной использования таких универсальных пакетов является то, что можно достаточно легко сравнивать все результаты построения разным способом нескольких моделей. Данная возможность существует для примера в пакете Statistica. Данный пакет позволяет сделать оценку, при помощи при-

менения различных моделей к одинаковому набору данных и последующем сравнении его характеристик.

Следует отметить, что на сегодняшний день наибольшее распространение технология Data Mining получила при решении бизнес–задач. На данный момент методы Data Mining используется почти во всех сферах деятельности человека, где есть большие объемы данных [5].

Основные из них:

- банковское дело;
- финансы;
- CRM;
- телекоммуникации;
- маркетинг;
- страхование;
- фондовые рынки;
- другие.

Перед началом работы над сервисом были проведены работы по анализу возможных нагрузок на сервер, расчет необходимого дискового пространства для хранения действий пользователей и подбор необходимых средств для дальнейшей разработки.

Для реализации функционала Web–сервиса был выбран язык программирования Python и в качестве фреймворка был выбран Django. Для создания красивого и удобного дизайна были использованы HTML (язык гипертекстовой разметки), CSS (каскадная таблица стилей) и Bootstrap 3 (HTML, CSS и Javascript фреймворк). За основную СУБД была взята Postgre SQL.

Главная страница сервиса содержит всю необходимую информацию для того, чтобы начать пользоваться им (Кнопки «Регистрация» и «Вход», ссылка на документацию и логотип сервиса) [6].

Для того, чтобы зарегистрироваться в системе, необходимо будет перейти на страницу регистрации и заполнить форму. Форма содержит поля «Фамилия» и «Имя», «Адрес электронный почты» и «Пароль».

После регистрации в системе на введенный адрес электронной почты будет отправлено письмо с подтверждением. После этого можно будет зайти панель управления магазином/магазинами воспользовавшись формой входа.

После входа в панель управления можно добавить новый магазин в систему. Для этого разработана страница «Добавить сайт», на которой необходимо указать Название, Адрес сайта, Часовой пояс и основной цвет сайта.

После добавления магазина в систему, будет сформирован уникальный JavaScript код.

Перед тем, как установить его на все страницы сайта, необходимо отредактировать код страницы товара в соответствии со стандартом микроразметки — Schema.org [7].

После установки JavaScript кода на все страницы сайта и настройки страницы товара, начнется сбор всей информации о действиях пользователя на странице. Скрипт будет анализировать только страницы товаров и пропускать посещение других страниц, т.к. основные данные с которыми работает система это — товары. На главной странице сервиса можно будет увидеть график посещаемости (существует возможность настроить промежуток времени, за который показывать данные, распечатать график или изменить его вид) и детально рассмотреть каждый из визитов при помощи таблицы, расположенной под графиком.

Каждый визит можно рассмотреть более детально. Для этого необходимо нажать на необходимую запись в таблице визитов. После этого откроется страница с подробной информацией по визиту.

На основе собранных данных о действиях пользователей, им индивидуально будут предлагаться наиболее подходящие товары и услуги в формате Push–уведомлений.

Уведомление показывается через каждые N просмотров страниц сайта. Параметр N задается в панели управления сайтом [8].

Товар для уведомления выбирается при помощи алгоритма бинарного решающего дерева. Деревья решений прекрасно справляются с такого рода задачами, которые требуют отнесения объектов к одному из заранее известных классов. При их построении большая часть внимания уделяется выбору атрибута, по которому будет происходить разбиение, остановка или отсечение ветвей. Дерево состоит из нескольких уровней и позволяет наиболее точно выбрать интересующий товар из массива данных на основе заданных критериев. На каждом внутреннем узле находится такое условие, которое разбивает множество на подмножества. В качестве такого условия выбирается один из атри-

бутов товара. Выбранный атрибут разбивает множество таким образом, чтобы образованные подмножества содержали только те объекты, которые относятся к одному и тому же классу. Для процесса остановки построения деревьев используется, например, ограничение глубины дерева до определенного заданного значения. Дерево классифицирующее товары для push–уведомлений имеет глубину равную пяти. Это позволяет с нужной точностью подобрать товар и одновременно сделать это быстро. В то же время алгоритмы построения деревьев решений очень часто строят довольно сложные деревья, которые содержат большое количество ненужных данных и ветвей. Такие деревья обучающее множество разбивает на большое количество подмножеств, которые содержат все меньше объектов, тем самым увеличивая объем дерева и его сложность. Для решения такой проблемы часто применяется процесс отсечения ветвей. Дерево, построенное для решения задачи анализа товаров в интернет магазинах с большим количеством посещений и товарных позиций, не является слишком «ветвистым», что позволяет не переполнять дерево большим количеством ненужных данных и условий [9].

Также, на каждой странице товара будет сформирован блок «Рекомендуемых» товаров на основе действий всех пользователей. Для того, чтобы в блоке рекомендаций выводить действительно полезные товары, которые могут пригодиться пользователю при покупке рассматриваемого товара, был использован алгоритм k–ближайших соседей. Число k — это количество соседних товаров в массиве, которые сравнивается с классифицируемым товаром. В данном случае было выбрано k равное 50, т.е. каждый товар сравнивается с 50–ю соседями. В процессе анализа случайным образом выбирается 50 визитов посетителей, в которых есть исходный товар. Определяются товары, со страницы которых был переход на рассматриваемый товар и складываются в массив данных. Для каждого найденного товара вычисляется количество его повторений в массиве, категория товара и время проведенное на странице товара. Далее массив сортируется по количеству повторений и времени в порядке убывания. При спорных ситуациях алгоритм задействует значение категории товара. На последнем шаге алгоритма выбираются 3 верхних элемента и передаются во View для вывода в блок «Рекомендации». Данный функционал был внедрен на сайт hector.ru, что позволило увеличить конверсию сайта и количество просматриваемых страниц [10].

ЗАКЛЮЧЕНИЕ

В настоящей работе были изучены понятие и основные аспекты Data Mining, основные классы задач решаемых при помощи Data Mining, разработан сервис для автоматического показа Push–уведомлений посетителям крупных интернет магазинов, разработан функционал для анализа получаемых данных о просмотре страниц товаров пользователями с использованием алгоритмов бинарного решающего дерева и k–ближайших соседей и на основе полученных результатов создан блок рекомендаций для каждого товара на сайте. Функционал был внедрен на несколько сайтов в тестовом режиме (hector.ru, rus-hunters.ru, alltape.ru). После внедрения система позволила как привлекать внимание покупателей, за счет ненавязчивых уведомлений о просматриваемых товарах, так и почти полностью исключила необходимость ручного добавления рекомендуемых товаров для каждой страницы товара в интернет магазине.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 *Myatt, J.* Making Sense of Data: A Practical Guide / J. Myatt. — Wiley-Interscience, 2006.

2 *Margaret, H.* Data mining introductory and advanced topics / H. Margaret. — Pearson, 2002.

3 *Selen, D.* Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools / D. Selen. — Manning, 2016.

4 *Barlou, M.* Real-Time Big Data Analytics / M. Barlou. — Oreilly, 2013.

5 *Zafarani, R.* Social Media Mining: An Introduction / R. Zafarani. — Cambridge University Press;, 2014.

6 *Maklenan, D.* Data Mining with Microsoft SQL Server / D. Maklenan. — Wiley, 2008.

7 *Lor, C.* Data-ism: The Revolution Transforming / C. Lor. — HarperCollins, 2015.

8 *Hasti, T.* The Elements of Statistical Learning: Data Mining, Inference, and Prediction. / T. Hasti. — Springer, 2001.

9 *Pilon, K.* Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference / K. Pilon. — Addison-Wesley, 2015.

10 *Berri, M.* Data Mining Techniques for Marketing, Sales, and Customer Relationship / M. Berri. — Wiley, 1997.

10.06.2017