МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра дискретной математики и информационных технологий

ПРИМЕНЕНИЕ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ ДЛЯ КЛАССИФИКАЦИИ КРАУДФАНДИНГОВЫХ ПРОЕКТОВ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Поповой Екатерины Игоревны

Научный руководитель
доцент, к.э.н. _____ Г. Ю. Чернышова

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2019

ВВЕДЕНИЕ

Краудфандинговые площадки становятся успешными бизнес-моделями финансирования, являясь альтернативой классическому банковскому кредитованию. Прежде чем открывать сборы средств для проектов на крудфандинговых платформах, следует проработать все детали плана проекта, представление и сделать оценку вероятности того, что проект станет успешным. Для оценки нового проекта, можно использовать результаты предыдущих проектов, найдя сходства и различия посредствам Data Mining. Однако возникает проблема сбора информации о проектах с краудфандинговых площадок, сведения о проектах хранятся в гипертекстовом виде. Чтобы получить данные о проектах для последующей обработки методами Data Mining, применяется технология web scraping. Для удобства оценки краудфандинговых проектов, анализ проектов оформляется в виде готового инструмента, которым является web-приложение.

Цель бакалаврской работы – разработка web-приложения с использованием методов Data Mining для оценки проектов на краудфандинговой платформе.

Поставленная цель определила следующие задачи:

1. сбор данных о проектах по технологии web scraping с краудфандинговой платформы;
2. разработка классификационной модели для оценки краудфандинговых проектов с помощью алгоритма Data Mining;
3. создание web-приложения для оценки краудфандинговых проектов на платформе;
4. использование построенной модели для оценки успешности проектов на краудфандинговой площадке.

Методологические основы «Применение методов интеллектуального анализа данных для классификации краудфандинговых проектов» представлены в работах A. Agrawal, C. Catalini, A. Goldfarb, А. А. Барсегяна, М. С. Куприянова, В. В. Степаненко, И. И. Холод, Т. В. Алексеевой, Ю. В. Амириди, В. В. Дика, Г. Шилдт, Дж .Ховарда.

Теоретическая значимость бакалаврской работы заключается в описании инструментов, использующихся для классификации проектов на краудфандинговой платформе.

Практическая значимость бакалаврской работы заключается в сборе данных с краудфандинговой платформы по технологии web scraping, создании классификационной модели на основе собранных данных о проектах и проектировании web-приложения для прогноза успешности проекта на платформе Boomstarter.

Бакалаврская работа состоит из введения, 3 разделов, заключения, списка использованных источников и 1 приложения. Общий объем работы – 53 страницы, из них 44 страницы – основное содержание, включая 4 рисунка и 4 таблицы, список использованных источников информации – 20 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе «Сбор данных о проектах с краудфандинговой платформы» выбирается и применяется библиотека для осуществления сбора данных с краудфандинговой платформы. Затем полученные данные подвергаются описательной статистике для выявления явных зависимостей.

Для сбора данных о проектах с краудфандинговой платформы воспользуемся технологией web scraping(скрейпинг). Она позволяет извлекать данные из страниц web-сайта с помощью специально обученного алгоритма. Этот алгоритм использует HTML-теги и CSS-селекторы на страницах и переход по ссылкам, извлекая необходимую информацию в хранилища разного рода. Хранилищами могут быть базы данных или текстовые документы. Скрейпинг помогает автоматически собрать важную часть данных. Для создания обучающего алгоритма необходимо знать разметку страниц и структуру сайта, для выделения пути по которому будет производится переход по страницам. Данные, полученные в результате работы алгоритма такого рода, используются для структурирования, отбрасывания не важных для поставленной задачи сведений и анализа полученных данных.

Существует много библиотек для различных языков программирования, облегчающих извлечение данных из HTML-страниц. В настоящий момент создано не мало готовых приложений, через пользовательский интерфейс которых можно настроить работу алгоритма. Это программное обеспечение может пытаться автоматически распознать структуру данных страницы или предоставить интерфейс записи, который устраняет необходимость вручную писать код или некоторые функции сценариев, которые можно использовать для извлечения и преобразования контента. Некоторое программное обеспечение для скрейпинга web-страниц также можно использовать для непосредственного извлечения данных из Application Programming Interface (API). Большинство полнофункциональных приложений, распространяющихся в сети Интернет, платные (по истечении срока пользования demo-версией), но в процессе их использования выявляются ошибки при настройке алгоритма. В связи с этим эффективнее использовать библиотеки языков программирования и непосредственно через код производить настройку алгоритма скрейпинга.

Стоит отметить тот факт, что сайты нередко обновляются, изменяя

структуру кода разметки, так что при работе со скрейпингом нужно быть готовым адаптироваться к изменениям. HTML-скрейпинг далеко не всегда может помочь извлечь необходимые данные. Проблема в том, что данные могут генерироваться налету внутри страницы, и тогда анализ кода страницы при подобном подходе бессилен. Существуют различные методы, позволяющие извлечь данные в этом случае, все они базируются на исполнении кода и использовании данных, полученных в результате исполнения, но способ исполнения и способ доступа к полученным данным может отличаться.

В рассматриваемой технологии скрейпинга сайтов отсутствует универсальное решение. Для каждого сайта создаётся уникальный алгоритм, в зависимости от набора данных, который нужно собрать. Если возникает необходимость сбора данных с нескольких ресурсов, то для каждого из них необходимо писать отдельный алгоритм, изучив структуру сайта, что является очень время- и ресурсозатратным.

Многие web-сайты стремятся защитить свой контент: запрещают доступ при частых запросах, используют динамическую смену верстки и другие способы. Все эти преграды возможно преодолеть, но потребуется большие затраты сил, что может перечеркнуть основные ценности скрейпинга: простота и высокая скорость получения структурированного контента.

По результатам сравнения библиотек лидирует библиотека jsoup-1.12.1. Она предоставляет очень удобный API для извлечения и манипулирования данными, используя Document Object Model (DOM), CSS и jquery методы. jsoup-1.12.1 реализует спецификацию WHATWG HTML5 и анализирует HTML в том же DOM, что и современные браузеры.

Среди всего многообразия краудфандинговых платформ для сбора данных о проектах была выбрана платформа https://boomstarter.ru. Выбранный инструмент скрейпинга с лёгкостью позволяет обойти архитектуру этой платформы для сбора данных о проектах, не встретив преград в виде мер, защищающих от сбора контента с сайта.

Со страницы каждого проекта можно собрать информацию о:

— месте нахождения проекта (url);
— названии проекта (title);
— категории, в которой он представлен (category);
— цели проекта в виде денежной суммы (goal);

— реальном итоге сбора средств на текущий момент времени (pledged);

— количестве вкладчиков, поддержавших проект (backers);

— денежном сборе в процентах на настоящий момент(percentage);

— дате запуска проекта (start);

— дате окончания сбора средств (end).

Выделены значения следующих показателей, которые способны повлиять на успешность проекта:

— цель проекта, руб;

— собранные средства, руб.;

— количество вкладчиков, чел.;

— успешность проекта, %.

Проект считается успешным, если сбор денежных средств составляет не менее 100% своей цели (в случае избыточного финансирования возможны значения выше 100%). Выбор этой меры определяется принципом работы «все или ничего» данной платформы.

В 2013-2016 г.г. средняя величина целевой суммы проекта и объем собранных средств увеличивалась. В 2017г. наметилась тенденция к уменьшению размера целевой суммы, при этом объем собранных средств в среднем увеличился. В 2018 г. в проектах указывались большие целевые суммы, но объем собранных средств уменьшился. Проекты, отнесенные к категориям, связанным с настольными играми и изданиями, традиционно собирают большие суммы на платформе Boomstarter.

В результате анализа библиотек скрейпинга для языка Java было принято решение осуществлять скрейпинг посредствам jsoup-1.12.1. Итогом произведённого скрейпинга данных стала таблица в виде набора всех проектов с краудфандинговой платформы boomstarter.ru. В процессе комплексной оценки проектов, которые были успешны с точки зрения объема собранных средств в 2018 г., следует отметить, что наибольшие суммы были получены проектами с относительно небольшими целевыми запросами.

Во втором разделе «Построение классификационной модели с использованием алгоритмов Data Mining» выбирается оптимальная задача интеллектуального анализа данных и способ отражения функциональной зависимости. Для выбранного способа отражения функциональной зависимости побирается наиболее точный метод построения. Далее строится классификаци-

онная модель для данных с крудфандинговой платформы.

Методы Data Mining решают следующие аналитические задачи:

— классификация,
— кластеризация,
— регрессия,
— поиск ассоциативных правил.

Задача классификации сводится к определению класса объекта по его характеристикам. Важно, что в условии этой задачи множество классов, к которым можно отнести объект, заранее известно. Для решения поставленной цели в данной работе, наиболее подходящая модель – задача классификации. Данный вид задачи является предсказывающей задачей обучения с учителем. Собрав данные, нужно построить модель, обучить её и инициировать запросы на прогнозирование успешности или не успешности проекта.

В задаче классификации обнаруженная функциональная зависимость между переменными может быть представлена в виде:

— классификационных правил,
— деревьев решений,
— математической функции.

Для поставленной цели, наиболее выгодной моделью представления является дерево решений. Реализация на языке программирования Java данной модели достаточно проста, и существует большое количество документации, в которой построение дерева решений является базовым примером.

Для обработки данных средствами Data Mining были выбраны программное обеспечение Weka и библиотека weka для языка Java. Выбор был сделан на основе того, что данные инструменты занимают ведущие позиции, и зачастую используются другими средствами интеллектуального анализа данных в качестве ядра. Программное обеспечение Weka предоставляет несколько методов построения деревьев решений: DecisionStump, HoeffdingTree, J48, LMT, Random- Forest, RandomTree, REPTree. Для того, чтобы выбрать наиболее оптимальный метод, были вычислены Correctly Classified Instances (корректно классифицированные объекты), Relative absolute error (относительная абсолютная ошибка) и Time (время, затраченное на построение модели) по каждому из методов. Анализ полученных вычислений указывает на то, что метод J48 является наиболее подходящим. Этот метод реализует алгоритм

C4.5, рассмотрим его более подробно.

Была проведена предварительная подготовка: все данные, собранные по средствам web scraping считаны и оформлены в виде файла с расширение «.arff».

Затем была создана классификационная модель методом J48 с использованием программного обеспечения Weka и «.arff»-файла. Модель была помещена в «jar»-файл.

Для решения поставленной цели, из задач Data Mining была выбрана задача классификации с отражением функциональной зависимости в виде дерева решений. Была проведена оценка методов построения деревьев решения для данных о проектах с краудфандинговой платформы. Самым оптимальным из них является метод J48, который реализует алгоритм C4.5. В результате использование программного обеспечения Weka, проектирования классов и методов с использованием библиотеки weka, была получена классификационная модель, способная прогнозировать успешность или неуспешность проекта на краудфандинговой платформе Boomstarter.ru.

В третьем разделе «Создание web-приложения для оценки успешности краудфандинговых проектов» на основе выбранного инструмента создаётся web-приложение для оценки успешности краудфандинговых проектов. Затем производится оценка точности модели, которая используется в web-приложении.

Web-приложение имеет много преимуществ. К web-приложению можно получить доступ с помощью любого устройства. Оно адаптивно, его компоненты способны учитывать особенности операционных систем и платформ. Для использования не нужно специального клиентского программного обеспечения помимо браузера. Подобное приложение может обеспечивать некоторую степень сетевой безопасности. Существует множество web-фреймворков, которые облегчают написание крупных web-приложений, например, Spring MVC, и позволяют абстрагироваться от многих технических моментов в написании кода web-приложения. Но для того, чтобы детально просмотреть все этапы создания web-приложения, была выбрана технология Servlet (сервлеты) в Java Enterprise Edition. Сервлеты – особый тип Java-программ, выполняемый в пределах web-контейнера (также называемый контейнером сервлетов, например, Tomcat), которые позволяют обрабатывать запросы клиентов

и ответы сервера. Сервлеты создаются и уничтожаются их контейнерами, а не разработчиком, и действуют как промежуточный уровень между клиентами и другими приложениями, запущенными на сервере. Сервлет способен принимать данные от клиента через GET- и POST-запросы, работать с cookie и параметрами сеанса. Также он обрабатывает данные через дополнительные уровни приложений и отправляет выходные данные клиенту как в текстовом, так и в бинарном форматах (HTML, XML, PDF, JPG, GIF и т.д.), во многих случаях используя Java Server Pages (JSP) файлы.

Была создана оболочка web-приложения с использованием технологии Servlet и библиотеки Apache Tomcat.

С помощью программного обеспечения Weka было выполнено построение модели на основе полученной информации о проектах. Модель строилась для решения задачи классификации. Для построения модели был загружен «arff»-файл с набором выявленных данных по проектам с тремя независимыми переменными: projectTarget, investors, projectDuration и зависимой переменной crowdfunding, по которой будет происходить классификация. Для проверки модели была выбрана стратифицированная перекрестная проверка. При подобном подходе имеющиеся в наличии данные разбиваются на n частей. Далее на (n - 1) частях данных производится обучение модели. Оставшаяся часть данных используется для тестирования. Цикл повторяется n раз. Таким образом, каждая из n частей данных используется для тестирования.

Самой важной информацией в характеристики построенного классификатора, являются пункты Correctly Classified Instances и Incorrectly Classified Instances. Они обозначают, что корректно были классифицированы 69%, т.е. 424 проекта. Некорректно были классифицированы 31%, т.е. 191 проект. Время построения классификатора таким методом относительно больше, 0.05 сек, но это компенсируется за счёт большей надёжности.

После построения модели классификатора, она была подключена к web-приложению на языке Java. Далее модель считывается, в неё могут подаваться запросы клиента с данными, полученными из формы.

Для создания web-приложения с оценкой успешности проектов была выбрана технология Servlet на языке Java. web-приложение имеет пользовательский интерфейс, состоящий из формы, в которую необходимо ввести данные о проекте, и страниц - ответов, содержащих результат прогноза об

успешности проекта. Также, к web-приложению была подключена библиотека weka. С помощью этой библиотеки особым образом формировались входные данные в объекты Weka, чтобы затем подать их на вход модели и получить наиболее вероятную классификацию проекта, т.е. прогнозирование успешности проекта.

ЗАКЛЮЧЕНИЕ

При ограниченности доступной информации по краудфандинговым проектам методы получения дополнительных сведений о текущем состоянии процесса сбора средств на конкретной платформе являются востребованными. Результаты применения методов обработки и анализа информативны для сторон, вовлеченных в процесс краудфандинга, предпринимателей и инвесторов. Оценка успешных проектов прошлого позволит принять более объективное решение о поддержке того или иного проекта. Для предпринимателей важно оценить динамику функционирования платформы перед использованием ее для размещения проекта.

Было создано web-приложение для оценки успешности проектов на крудфандинговой платформе Boomstarter на языке Java.

Были собраны данные по всем проектам с 2013 года с использованием библиотеки jsoup-1.12.1. В результате объём набора проектов составил 1876 проектов.

Далее были отобраны данные по проектам за последний год, количество которых составило 615 проектов, и на их основании была построена модель классификатора с использованием программного обеспечения Weka. Эта модель была подключена к web-приложению, созданному с помощью технологии Servlet.

Пользователь через форму может передавать данные своего проекта, они отправляются на сервер, где с помощью библиотеки weka происходит обращение к построенной модели и создаётся прогноз на основе работы классификатора. Построенная модель имеет точность 69 %. В зависимости от результата прогноза пользователь может сделать вывод о потенциальной успешности проекта.

По материалам работы была опубликована статья Чернышова, Г. Ю., Попова, Е. И. «Анализ функционированиякраудфандинговых платформ» / Цифровые технологии в экономике и обраовании : сборник научных трудов.- Саратов, 2019. - С. 43-49.

Основные источники информации:

1. Agrawal, A. Some simple econoics of crowdfunding. Innovation Policy and the Economy / A. Agrawal, C. Catalini, A. Goldfarb [Электронный ресурс]: научно-исследовательская работа - National bureau of economic research,

Cambridge, 2013 – URL: https://www.researchgate.net/publication/272543487_Some_Simple_Economics_of_Crowdfunding (дата обращения: 20.05.2019) – Загл с экрана. – Яз. англ.

2. Барсегян, А. А. Технологии анализа данных Data Mining, Visual Mining, Text Mining, OLAP / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – 2-е изд., перераб. И доп. – СПб.: БХВ-Петербург, 2007. – 384 с.
3. Алексеева, Т. В. Информационные аналитические системы / Т. В. Алексеева, Ю. В. Амириди, В. В. Дик и др.; под ред. В. В. Дика. - М.: МФПУ Синергия, 2013. - 384 с.
4. Шилдт, Г. Java 8. Полное руководство / Г. Шилдт - 9-е изд.: пер. с англ. - М. : ООО "И.Д. Вильяме 2015 - 1376 с.
5. Ховард, Дж. Предиктивный анализ: создание простой модели классификации на Java с использованием Weka [Электронный ресурс]: интерактив. учеб. – URL : https://www.ibm.com/developerworks/ru/library/bd-javaweka/index.html (дата обращения: 20.05.2019) – Загл с экрана. – Яз. рус.