Применение Data Mining в индустрии видеоигр

Суворов С.В., к.э.н., профессор кафедры «Прикладная информатика»,
ФГБОУ ВО «Московский политехнический университет», Москва, Россия
Жиляева И.А., к.э.н., доцент кафедры «Прикладная информатика»,
ФГБОУ ВО «Московский политехнический университет», Москва, Россия
Евтихов В.Г., к.т.н, доцент, кафедра «Прикладной информатики»,
ФГБОУ ВО «Московский политехнический университет», Москва, Россия
Баранова А.В., магистрант кафедры «Прикладная информатика»,
ФГБОУ ВО «Московский политехнический университет», Москва, Россия

Аннотация. Данная статья посвящена вопросам интеллектуального анализа больших данных игровой платформы Steam, которые включают в себя 609 549 наблюдений, с целью определения игрового поведения российских геймеров. К набору данных была применена технология Data Mining посредством инструментов: Python 3.7, Jupyter Notebook 6.0.1.

Ключевые слова. Интеллектуальный анализ данных, большие данные, подготовка данных, система управления базами данных, дамп базы данных

Application of data mining in the video game industry

Suvorov S.V., candidate of economic sciences, Professor of the «Applied Informatics» Department at the Moscow polytechnic University, Moscow, Russia
Zhilyaeva I.A., candidate of economic sciences, Associate Professor of the «Applied Informatics» Department at the Moscow polytechnic University, Moscow, Russia
Evtihov V.G., candidate of technical Sciences, Associate Professor of the «Applied Informatics» Department at the Moscow polytechnic University, Moscow, Russia
Baranova A.V., magistrate of the «Applied Informatics» Department at the Moscow polytechnic University, Moscow, Russia

Annotation. This article is devoted to the issues of big data mining of the Steam gaming platform, which includes 609,549 observations, in order to determine the gaming behavior of Russian gamers. Data Mining technology was applied to the data set using the following tools: Python 3.7, Jupyter Notebook 6.0.1.

Keywords: data mining, big data, data preparation, database management system, database dump

Сегодня рынок видеоигр является одним из самых востребованных сегментов. В 2019 году выручка индустрии видеоигр составила $152,1 млрд, что на 9,6% больше по сравнению с предыдущим годом [1]. Сегмент насчитывает более 2 млрд игроков. С ростом числа пользователей, увеличивается и объем пользовательских данных. Претерпевают изменения и технологии их обработки, старые методы просто не справляется с объемами, нужен качественно новый метод, которым стал интеллектуальный анализ данных (Data mining – «добыча данных, извлечение данных»). Применение методов Data mining к данным видеоигр - новый подход к извлечению знаний в игровой индустрии.

Применение технологии Data Mining к наборам игровых данных помогает компаниям совершенствовать свой продукт, выявлять состояние рынка, прогнозировать тенденции его развития, и, как следствие, получать больше прибыли. Данная технология позволяет выявлять скрытые закономерности в больших наборах данных. Она состоит из изучения предметной области; постановки задачи; подготовки данных, необходимых для проведения анализа; построения модели, а также ее проверки, оценки, применения и корректировки [4].

Для исследований был найден набор данных в формате .sql [2]. Данный набор данных был собран с игровой платформы Steam и проанализирован доцентом кафедры информатики Даниэлем Заппала из университета Brigham Young для конференции ACM Internet Measurement Conference 2016 года. Данные были собраны в 3 этапа: 1 этап – с 28 февраля 2013 года по 18 марта 2013 года – сбор информации об учетных записях пользователей; 2 этап – с 5 мая 2013

года по 5 ноября 2013 года – сбор информации о списках друзей пользователей, о играх пользователей, о членствах в группах, об игровом времени пользователей, о приложениях; 3 этап был завершен 9 апреля 2014 года, когда был собран «полный» список идентификаторов приложений каждого продукта [3].

Сервис Steam, созданный корпорацией Valve, является крупнейшей игровой платформой в мире. Сервис Steam был запущен в 2003 году. Каталог Steam, по состоянию на май 2014 года (момент сбора данных), насчитывал 108,7 млн пользователей [3]. Игровая платформа включает в себя не только магазин игр, но и социальную сеть, стриминговый сервис, набор инструментов для издателей и разработчиков игр, позволяющий эффективно распространять игры. Посредством платформы Valve занимается сбором, хранением, и обработкой игровых данных, которые в последствии используются компанией при создании игр.

Главная цель данной работы, это проведение анализа больших данных игровой платформы Steam для определения игрового поведения российских геймеров. Поставленные задачи работы: Развертывание дампа базы данных, Импорт базы данных в СУБД, Формирование выборки, Построение модели для проведения анализа, Анализ полученных результатов и их визуализация.

В формате .sql представлен дамп базы данных. Дамп базы данных состоит из описания структуры базы и/или содержащихся в ней данных, обычно в виде команд SQL. Дамп сам по себе базой не является, он лишь позволяет ее воссоздать. Версия СУБД дампа – Mariadb 5.5.52 for Linux (x86_64), вес дампа - 160 ГБ. Была произведена попытка развернуть данный дамп с помощью Open Server 5.2.2 в HeidiSQL путем импорта дампа. К сожалению, не хватило программных мощностей и удалось развернуть только 20,3 ГБ из 160 ГБ за 3 дня. Таким образом, удалось загрузить порядка 188 млн строк данных. Затем была предпринята попытка развернуть дамп на сервере. После успешного развертывания дампа, собранная база данных была скачана и импортирована в dbFordge Studio for MySQL.

База данных содержит следующие таблицы:

– achievement_percentages – содержит идентификатор рассматриваемой игры; название достижения в игре; процент игроков, которые закончили это достижение из всех игроков, которые владеют этой игрой;

– app_id_info – содержит идентификатор рассматриваемого приложения (игр большинство); название; тип; стоимость приложения; дату выпуска; рейтинг; возрастное ограничение; режим (многопользовательский или нет);

– friends - содержит идентификатор пользователя Steam, у которого был запрошен список друзей; идентификатор пользователя Steam, являющегося другом рассматриваемого пользователя; дату и время, когда пользователи стали друзьями;

– games_1 – содержит идентификатор пользователя Steam; идентификатор данного приложения в библиотеке пользователя; общее время, в течение которого пользователь запустил это приложение в течение двух недель; общее время, в течение которого пользователь запускал это приложение с момента добавления его в свою библиотеку; отметка времени, когда эти игровые данные были запрошены;

– games_2 – аналогично games_1(отличие - разное временя запроса);

– games_daily – аналогично games_1 (отличие – данные запрашивались каждые две недели);

– games_developers – содержит идентификатор приложения; название разработчика приложения;

– games_genres – содержит идентификатор приложения; название жанра приложения;

– games_publishers – содержит идентификатор приложения; название издательства;

– groups - содержит идентификатор разработчика приложения

– player_summaries – содержит идентификатор пользователя Steam, логин пользователя; отметку времени, когда эти игровые данные были

запрошены из API; группу пользователя; время создания учетной записи; название игры, которую запускал пользователь в момент времени, когда игровые данные были запрошены из API; код страны, в которой проживает пользователь; код региона, в котором проживает пользователь; статус в котором находится пользователь; отметку времени, когда эти игровые данные были запрошены из API.

С помощью SQL запросов был проведен отбор данных по двум критериям: по стране (Россия), по количеству общего игрового времени (>10 минут).

Для формирования выборки из таблиц запрашивались следующие данные: логин пользователя; код страны, в которой проживает пользователь; код региона, в котором проживает пользователь; идентификатор игры; название игры; рейтинг игры; жанр игры; общее время, в течение которого пользователь запускал игру с момента добавления его в свою библиотеку. Таким образом, было получено 1 047 347 наблюдений.

После очистки данных от дубликатов и выбросов, была получена выборка, состоящая из 609 549 наблюдений. Данная выборка была экспортирована в файл формата .csv для последующего анализа.

Набор данных исследовался с помощью языка программирования Python 3.7 и применения библиотек NumPy, Pandas, Plotly. Используемая среда для проведения анализа и визуализации полученных результатов – Jupyter Notebook 6.0.1.

Всего было проанализированно 101 006 геймеров и 1 713 игр. Так, общее игровое время геймеров составляет 5 195 лет, среднее время, проведенное в игре - 450 часов, в среднем пользователь владеет 6 – 7 играми.

Проанализировав по играм общее время, в течение которого пользователи запускали игру с момента добавления в свою библиотеку, был составлен топ 10-игр (Рис. 1). Самой популярной игрой в России является – Dota 2, в нее сыграли порядка 13161882 часов.

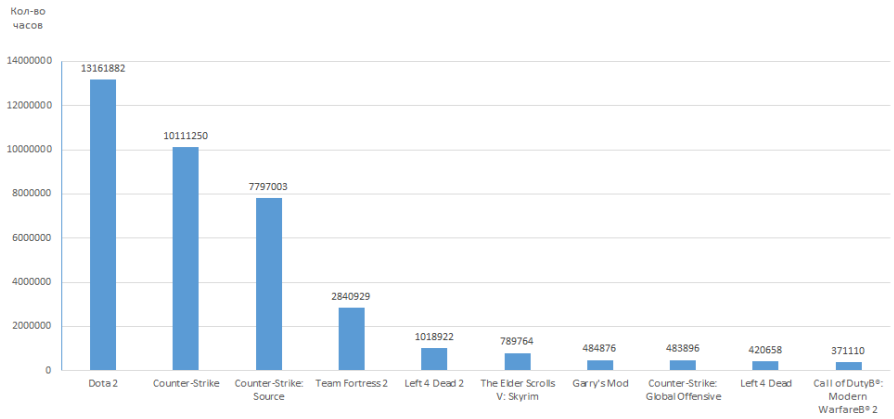

Рис. 1 – Топ -10 игр

По среднему времени проведенному в игре лидирует – Counter Strike (Рис. 2).

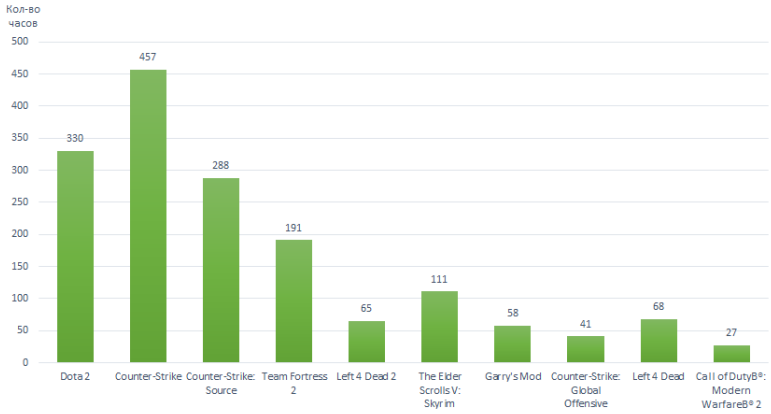

Рис. 2. – Среднее время проведенное в игре на 1 пользователя

Проведя исследование распределения общего игрового времени по регионам, был получен вывод о том, что большего всего в России играют в Липецкой области (Рис. 3).

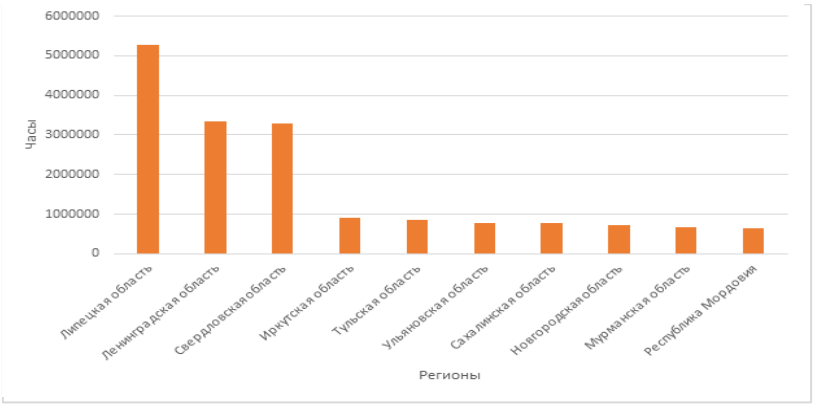

Рис. 3 – Топ - 10 регионов по общему игровому времени

По среднему времени, проведенному в игре лидирует Красноярский край (Рис. 4). Данное явление связано с тем, что в красноярском крае пользователей

меньше, но играют они больше по сравнению с пользователями из других регионов.

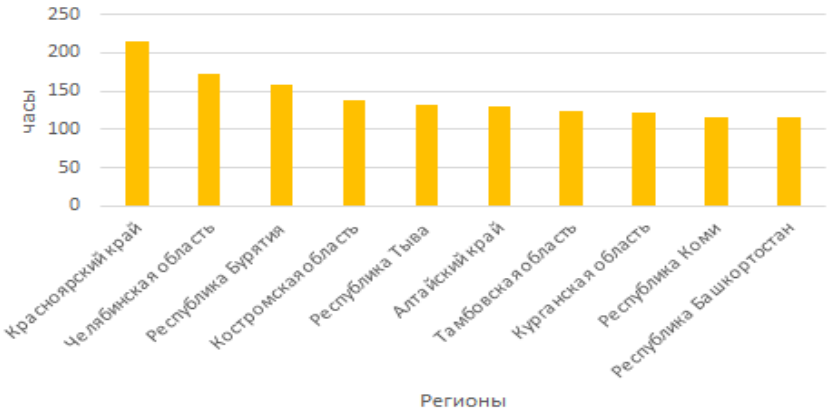

Рис. 4 – Среднее время проведенное в игре на 1 пользователя

Проанализировав рейтинги игр в зависимости от жанра, был сделан вывод, что наибольшие рейтинги имеют игры жанров: действие (Action) - максимальный рейтинг 96, ролевые (RPG) - максимальный рейтинг 96, приключения (Adventure) - максимальный рейтинг 95, стратегия (Strategy) - максимальный рейтинг 94.

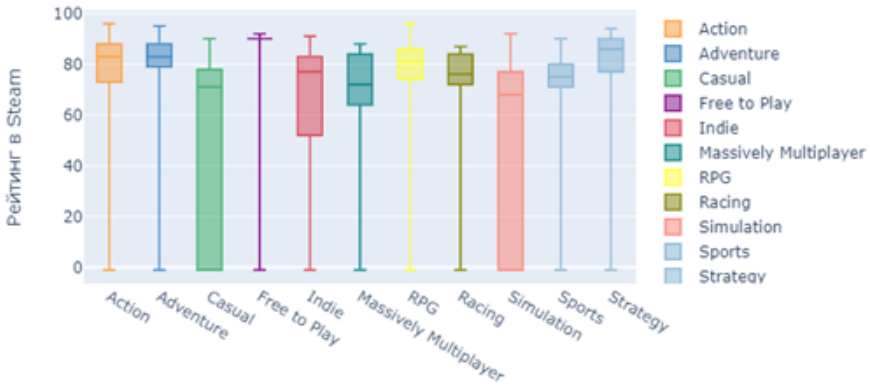

Рис. 4 – Зависимость рейтинга от жанра

В данной статье произведен интеллектуальный анализ больших данных игровой платформы Steam, с помощью языка программирования Python 3.7 в среде Jupyter Notebook 6.0.1. Было установлено, что самой популярной игрой у российской аудитории является Dota 2, а больше всего времени в игре проводят жители Липецкой области. Также россияне предпочитают игры жанров – действие, ролевые и приключения.

Видеоигры – это мощный генератор больших объемов пользовательских данных телеметрии, а также данных о производстве и производительности игр,

что потенциально является ценным источником бизнес-аналитики. Эти данные помогут в разработке игр, их улучшении и продвижении в соответствии с требованиями игрока и его областью интересов, и как следствие, помогут увеличить прибыль компании.

Библиографический список:

1. Global game market report 2019 [Электронный ресурс]. Режим доступа: https://resources.newzoo.com/hubfs/2019_Free_Global_Game_Market_Report.pdf

2. Steam Dataset [Электронный ресурс]. Режим доступа: https://steam.internet.byu.edu/

3. Zappala D. Condensing Steam: distilling the diversity of gamer behavior [Электронный ресурс]. Режим доступа: https://internet.byu.edu/static/papers/steam-imc-2016.pdf

4. Луньков А.Д., Харламов А.В. Интеллектуальный анализ данных: учеб. пособие [Электронный ресурс]. Режим доступа: http://elibrary.sgu.ru/uch_lit/1141.pdf

References:

1. Global game market report 2019 [Electronic resource]. Access mode: https://resources.newzoo.com/hubfs/2019_Free_Global_Game_Market_Report.pdf

2. Steam Dataset [Electronic resource]. Access mode: https://steam.internet.byu.edu/

3. Zappala D. Condensing Steam: distilling the diversity of gamer behavior [Electronic resource]. Access mode: https://internet.byu.edu/static/papers/steam-imc-2016.pdf

4. Lunkov D.A., Kharlamov A.V. Data mining: a tutorial [Electronic resource]. Access mode: http://elibrary.sgu.ru/uch_lit/1141.pdf