Правительство Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Департамент программной инженерии
Образовательная программа 09.03.04 «Программная инженерия»

ОТЧЕТ
по технологической практике

в (на) ООО "Вебгеймс"
(название организации, предприятия)

Выполнил(а) студент(ка)
Группы БПИ141__

_____ Хузин Т.В _____
(подпись) (инициалы, фамилия)

Руководитель практики от предприятия

_____

(должность, ФИО)

Дата _____ _____ _____

(оценка) (подпись)

Москва – 2016

ВВЕДЕНИЕ

Целью практики было ознакомление с методами и приёмами анализа данных, а также получения опыта применения их на реальных данных.

Исследовательский анализ данных - это процесс получения скрытой информации из имеющихся данных. Выводы, полученные в результате такого исследования, могут применяться для оптимизации бизнес-процессов и принятия стратегических решений.

Отчёт содержит описание исследования, проведённого в компании Webgames - разработчике игр для социальных и мобильных платформ. Оно было направлено на выявление различных групп пользователей по их поведению и определение изменений в их поведении с течением времени.

В рамках кластеризации пользователей были выполнены подбор параметров для кластеризации, разделение на периоды, очистка от выбросов, подбор количества кластеров, визуализация и интерпретация результатов.

ОПИСАНИЕ ИГРЫ

Проект "Привидения" ("Ghost Tales") представляет собой игру, ориентированную на западных женщин от 25 до 40 лет. Проект представлен в популярных социальных сетях (Facebook, ВКонтакте, Одноклассники) и на iOS.

Игровой процесс состоит в прохождении игроком квестов главной сюжетной линии и побочных квестов. Порядок прохождения главных сюжетов определён, побочные могут проходиться игроком независимо.

В процессе игры пользователь может покупать ("донатить") и добывать в процессе игры ("гриндить") кристаллы и монеты, которые могут применяться в игровом магазине.

ИСПОЛЬЗОВАННЫЕ ИНСТРУМЕНТЫ

Программные инструменты:

1. Python 2.7
2. IPython Notebook
3. библиотека Pandas
4. библиотека sciPy
5. библиотека numpy
6. библиотека matplotlib
7. библиотека networkx
8. система контроля версий git

ПОДГОТОВКА ДАННЫХ

Для анализа были выгружены два файла pinfo.csv и pact.csv. Первый файл содержит описание аккаунта игрока, в том числе id и время регистрации в формате POSIX. Второй файл содержит логи активности игроков, в том числе id игрока; дату; прогресс игрока в главном квесте(ml_num); количество пройденных квестов к концу игры(quest_end); оценку "гринда" (grindnum); количество купленных кристаллов (buynum). Все данные были анонимизированы для защиты персональных данных.

Выбор параметров для кластеризации

Для подготовки данных использовался скрипт preprocessing.ipynb. Данный скрипт сначала разделяет логи по времени с момента регистрации. Всего периодов 5:

1. момент регистрации - 7-й день;
2. 8-й день - 14-й день;
3. 15-й день - 28-й день:
4. 29-й день - 56-й день;
5. 57-й день - 200-й день.

После разделения логов по периодам он составляет для каждого периода таблицу, состоящего из следующих полей:

id	ID игрока
activity_quantity	количество строк лога, соответствующих этому игроку
med_ml_num	скорость прохождения основного квеста
buynum	среднее количество покупок игрока
grind	среднее grindnum
quest_speed	скорость прохождения побочных квестов

Эти поля являются параметрами кластеризации, так как являются значимыми и легко интерпретируемыми. При этом, учитывались лишь игроки, которые проявляли какую-нибудь активность.

Результаты работы сохраняются в разных файлах вида preprocessed_*время*.csv, где *время* - верхняя граница временного периода.

Проверка параметров на независимость

Для этого был написан скрипт correlation_test.ipynb. Данный скрипт проверяет гипотезу о наличие статистической гипотезы о наличии связи между разными парами параметров кластеризации с помощью теста Стьюдента. Гипотезы были отвергнуты.

КЛАСТЕРИЗАЦИЯ ДАННЫХ

Выбор оптимального числа кластеров

Перед проведением кластеризации надо было выбрать оптимальное число кластеров. Для этого было использовано правило локтя (elbow rule). Для этого был написан скрипт elbow_rule_for_times.ipynb.

Графики зависимости функции стоимости от количества кластеров для каждого правила приведены в приложении 1.

Согласно правилу локтя, было решено использовать по 4 кластера на каждый период.

Процесс кластеризации

Сама кластеризация проводилась с помощью скрипта clustering_by_time.ipynb.

Порядок работы скрипта:

1. частичное удаление выбросов (все элементы, выходившие за границы 98% квантили были к ней приравнены);
2. приведение всех измерений к одинаковому масштабу;
3. кластеризация алгоритмом KMeans;
4. восстановление исходного масштаба для центроидов, вывод их в файл, сохранение элементов кластеров.

РЕЗУЛЬТАТЫ И ИХ ИНТЕРПРЕТАЦИЯ

Центроиды кластеров, полученных в результате работы алгоритма K-Means приведены в приложении 2.

Каждый кластер имеет индекс.

Кластеры 2 за первый, второй, пятый и 1 за третий и четвёртый считаются выбросами, так как имеют крайне низкую численность пользователей.

Остаются три кластера игроков, которые проявляли активность, и один кластер неактивных.

Визуализация графа переходов пользователей по разным кластерам с течением времени и число пользователей в кластерах представлены на рисунках 6 и 7 в приложении 3.

Описание рёбер графа на рис. 6:

- in - какой процент от итогового кластера представляют те, кто перешли по ребру.
- out - какой процент от исходного кластера представляют те, кто перешли по ребру.

На рёбрах на рисунке 7 записано количество перешедших.

Описание кластеров

Кластеры с индексом -1 - кластер неактивных игроков, которые никак себя не проявляли весь период ("отколовшиеся").

Остальные кластеры описываются в виде цепочек кластеров, так как у разных кластеров в разные периоды есть сильная схожесть. Были выделены цепочки "гриндеры", "донатеры" и "пескари". Также есть две когорты "вернувшихся", то есть тех, кто был ранее в кластере неактивных игроков.

1) цепочка кластеров "**гриндеры**":

3 -> 1 -> 2- >3 ->1

Наименьшая по числу игроков группа(число игроков колеблется около отметки в 60 человек). В начале игры весьма походят на группу "донатеры", но в дальнейшем расходятся с ними.

Имеют самую низкую частоту заходов за период, высокую скорость прохождения побочных квестов, вторую скорость прохождения основного квеста (причём, довольно быстро).

В первые две недели имеют меньший средний платёж, чем "донатеры", но общая сумма с пользователя у "донатеров" больше за счёт большей активности.

Значимые источники из других групп: переходы от "донатеров" в первые две недели и одна из двух когорт "вернувшихся" из кластера -1.

2) цепочка кластеров "**донатеры**":

1-> 3 ->3 -> 2 -> 3

Больше всех покупают и быстрее всех продвигаются в основном квесте. Быстро проходят побочные квесты. Очень часто играют. Во время игры гриндят больше "пескарей", но в разы меньше, чем гриндеры.

Численность колеблется в диапазоне 600- 800 человек.

Основные притоки цепочки: "пескари" - каждый период кроме последнего от 14% до 32% от итогового числа; обленившиеся или разбогатевшие "гриндеры" - от 17 до 35% "гриндеров" в каждый период (в пределах 5% от итогового числа).

Максимизируют количество на второй неделе, потом уходят понемногу к "пескарям".

3)"пескари":

Все кластеры с индексом 0.

Заходят чаще "гриндеров", но реже "донатеров"; платят и гриндят гораздо меньше, чем донатеры и гриндеры; в итоге получают очень малый прогресс.

Очень многочисленны, сравнимы лишь с "отколовшимися".

Одна из двух когорт "вернувшихся" идёт в "пескари"; также очень популярный переход от донатеров.

Эти цепочки есть стандартные пути развития игрока: игрок скорее всего будет вести себя так же, как и в первую неделю.

Нестандартные переходы:

- не заходить в игру в первую неделю после регистрации, потом стать гриндером или пескарём;
- перестать донатить и стать пескарём;
- быть пескарём, а потом начать активно гриндить или донатить.

ЗАКЛЮЧЕНИЕ

В результате исследования выявлено, что пользователей можно разделить на две группы: активные и неактивные. Активные пользователи, в свою очередь, делятся на три группы: "пескари", которые мало играют и мало платят; "донатеры", которые много платят и много играют; "гриндеры", которые много играют, но мало платят. По количеству пользователей "пескарей" в примерно 100 раз больше, чем "донатеров" и в 1000 раз больше, чем "гриндеров".

Также обнаружено, что поведение игрока можно очень точно предсказать по его поведению в первую неделю после регистрации: игроки, по большей части, относятся к тому же кластеру, к которому они были отнесены вначале. Однако из этого правила есть исключения:

- те, кто не заходит в игру в первую неделю после регистрации, потом начинает играть как гриндер или пескарь;
- те, кто из "донатеров" переходят в "пескари";
- те, кто из "пескарей" перешли в разряд "донатеров" и "гриндеров".

Пути применения и улучшения исследования:

- более детальный анализ признаков пользователей, которые начинают играть не сразу, чтобы продумать стратегию привлечения большего числа "отколовшихся";
- анализ "пескарей" с целью определения потенциальных "донатеров ";
- таргетированные акции;
- анализ "донатеров", которые уходят в "пескари" с целью их дополнительной стимуляции.

С исходным кодом скриптов можно ознакомиться через Github: https://github.com/AngelicosPhosphoros/ghost_clustering

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1) Официальный сайт Jupyter Notebook [Электронный ресурс] URL: http://jupyter.org/
2) Обзор алгоритмов кластеризации данных. Habrahabr.ru[Электронный ресурс] URL: https://habrahabr.ru/post/101338/
3) Pandas Documentation [Электронный ресурс] URL: http://pandas.pydata.org/pandas-docs/version/0.18.0/
4) Документация пакетов scipy и numpy [Электронный ресурс] URL: http://docs.scipy.org/doc/

ПРИЛОЖЕНИЕ 1

ГРАФИКИ ЗАВИСИМОСТИ ФУНКЦИ СТОИМОСТИ ОТ ЧИСЛА КЛАСТЕРОВ

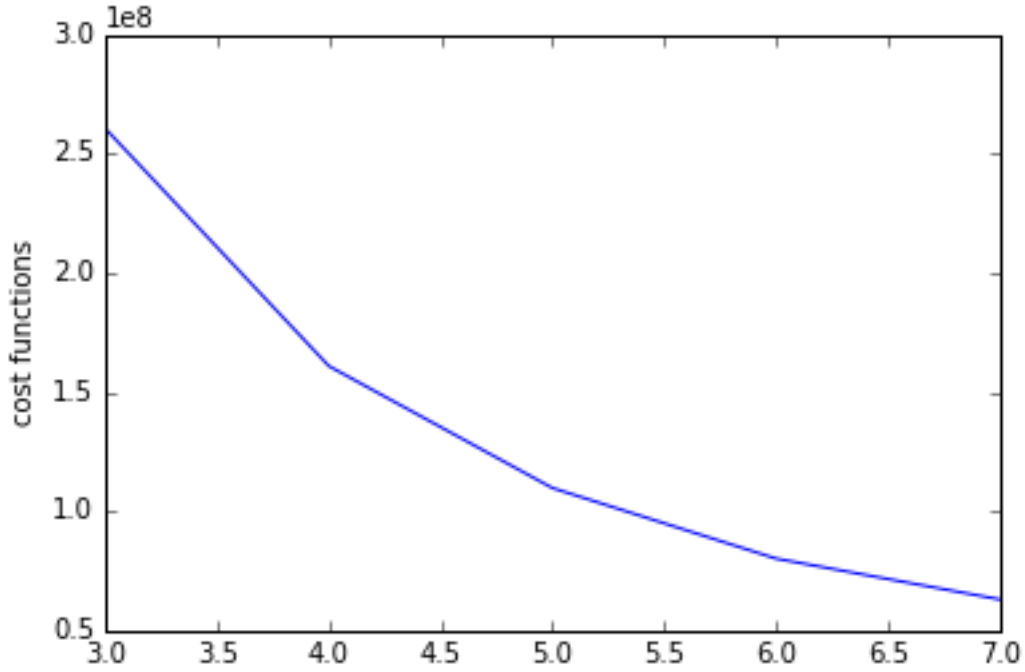

Рисунок 1. График для первого периода.

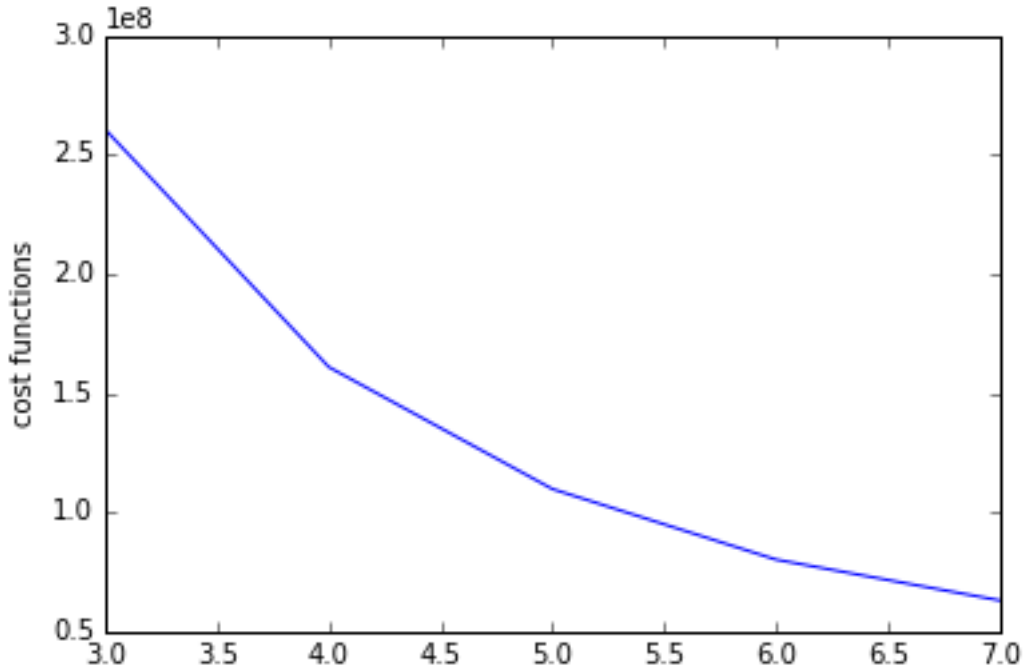

Рисунок 2. График для второго периода.

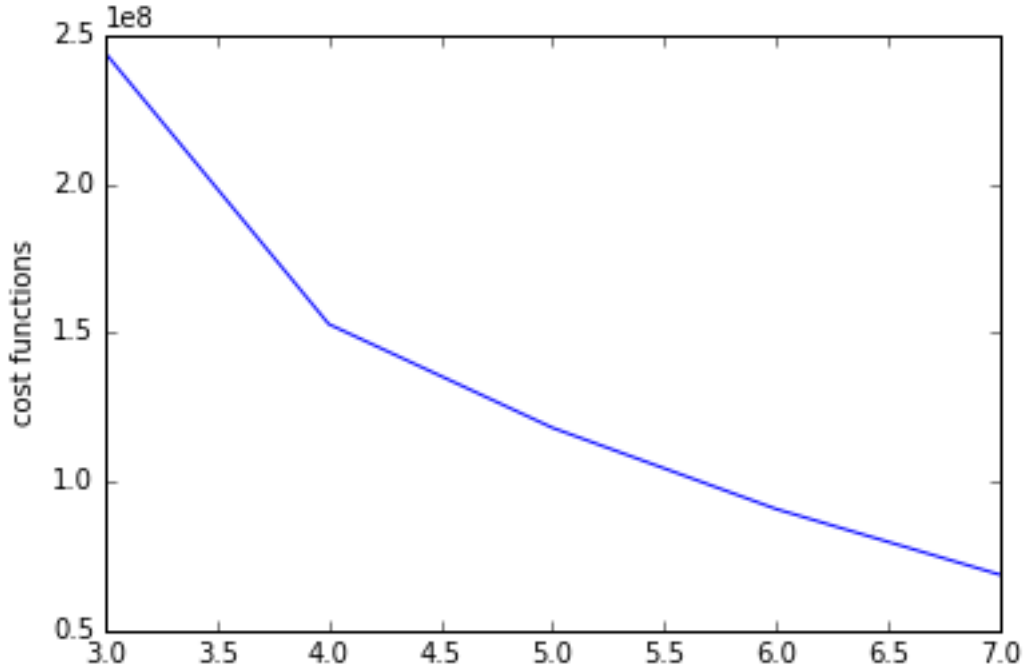

Рисунок 3. График для третьего периода.

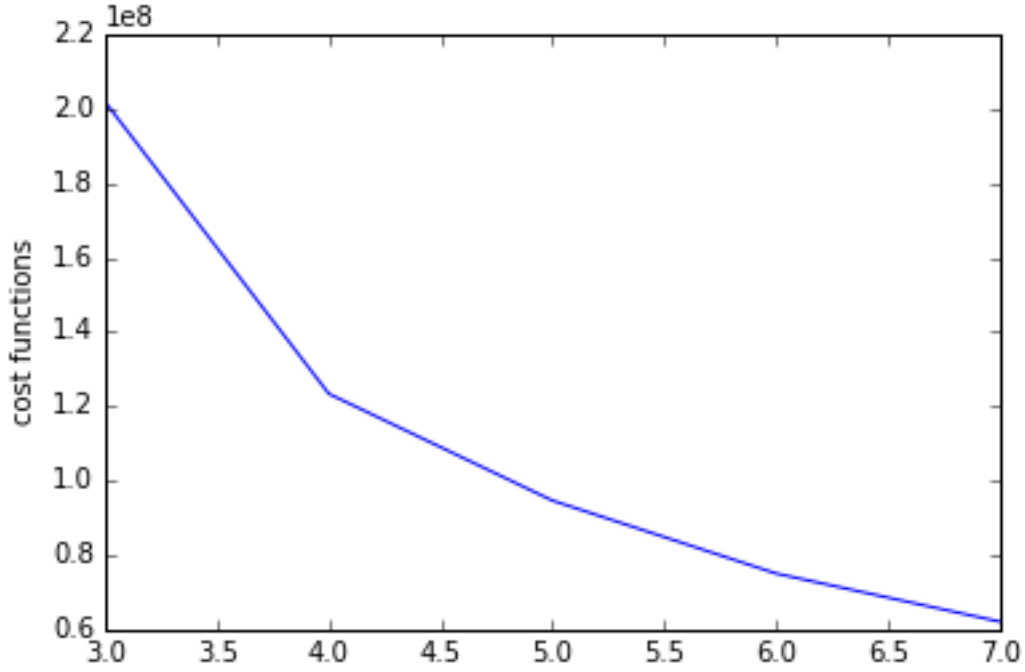

Рисунок 4. График для четвёртого периода.

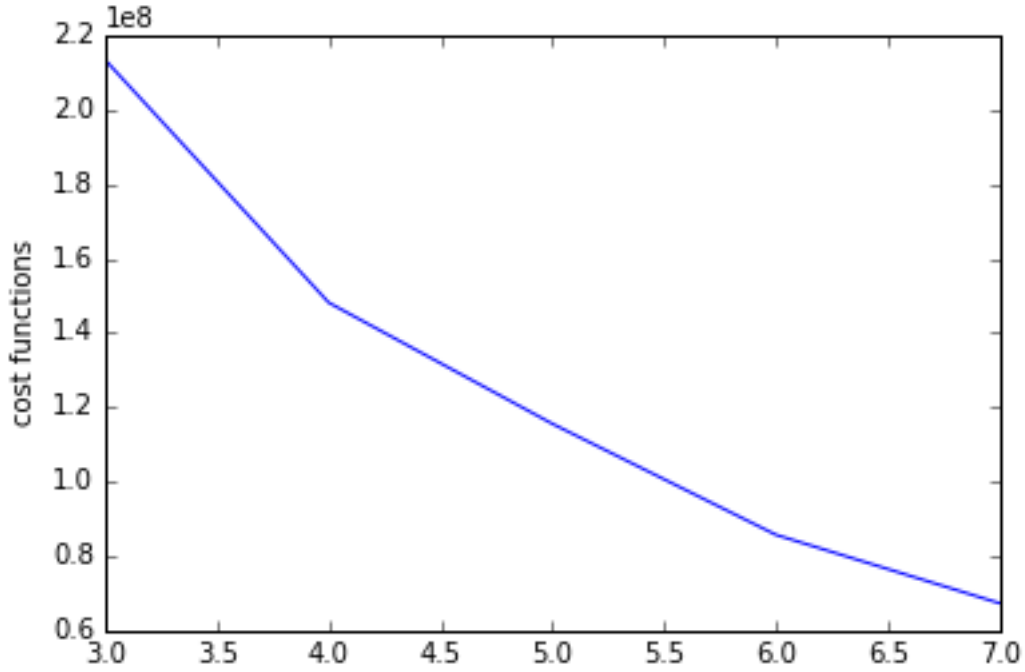

Рисунок 5. График для пятого периода.

ПРИЛОЖЕНИЕ 2

ТАБЛИЦА ЦЕНТРОИДОВ, ПОЛУЧЕННЫХ В РЕЗУЛЬТАТЕ РАБОТЫ АЛГОРИТМА K-MEANS

0 - 7						
cluster_id	user_count	activity_quantity	med_ml_num	buynum	grind	quests_speed
0	45399	8,900	0,008	102,452	886,252	9,106
1	603	10,378	0,013	373,944	76611,432	17,289
2	1	4,885	0,000	0,000	1642401,439	1,000
3	60	8,812	0,011	205,512	300466,119	16,895
8 - 14						
cluster_id	user_count	activity_quantity	med_ml_num	buynum	grind	quests_speed
0	52057	24,879	0,006	75,944	773,261	6,381
1	84	24,697	0,007	156,192	238896,661	9,517
2	2	8,120	0,000	0,000	1517648,119	1,000
3	780	32,215	0,010	237,630	55587,778	12,169
15 - 28						
cluster_id	user_count	activity_quantity	med_ml_num	buynum	grind	quests_speed
0	57431	66,271	0,005	62,083	796,682	4,932
1	3	76,641	0,017	5,361	1591612,749	6,523
2	78	50,254	0,006	303,839	219534,376	9,132
3	714	94,287	0,008	191,609	55576,060	9,555
29 - 56						
cluster_id	user_count	activity_quantity	med_ml_num	buynum	grind	quests_speed
0	61521	157,531	0,004	55,296	608,002	4,308
1	2	217,220	0,002	6,352	1346751,785	1,165
2	704	257,845	0,007	216,403	42350,337	8,825
3	74	73,537	0,006	283,127	173482,110	8,437
57 - 200						
cluster_id	user_count	activity_quantity	med_ml_num	buynum	grind	quests_speed
0	65784	368,564	0,003	48,288	452,941	3,970
1	51	167,397	0,004	308,320	152584,016	8,350
2	1	29,406	0,000	0,000	965654,842	0,500
3	630	653,995	0,006	217,926	33526,476	8,661

ПРИЛОЖЕНИЕ 3

ГРАФЫ ПЕРЕХОДОВ ПОЛЬЗОВАТЕЛЕЙ ПО ГРУППАМ

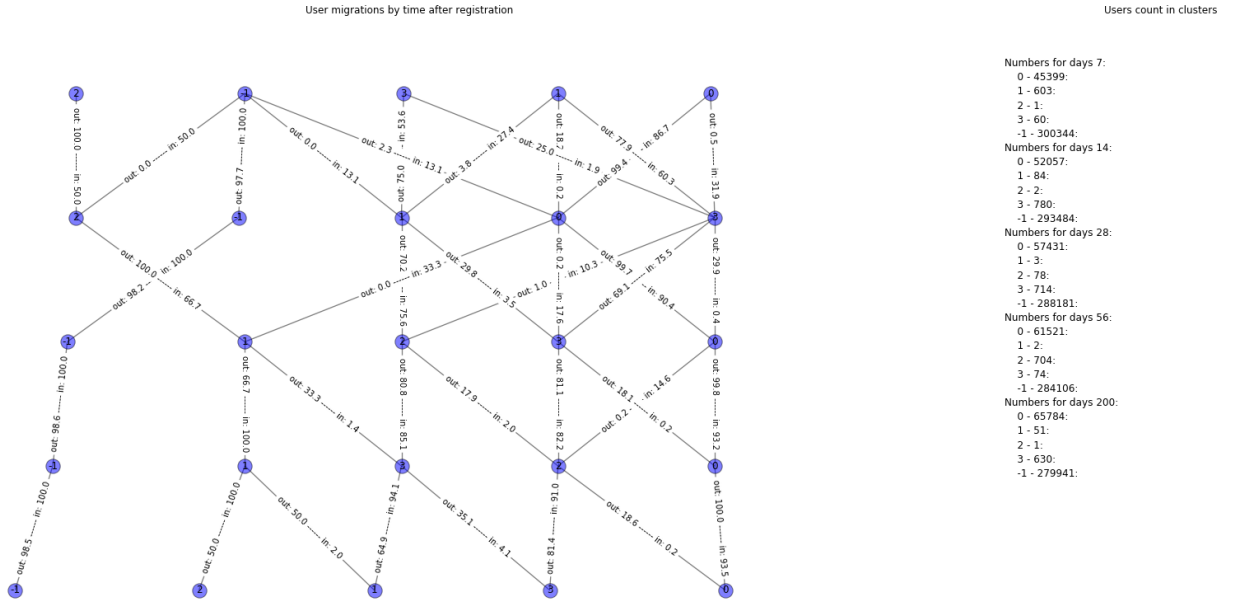

Рисунок 6.

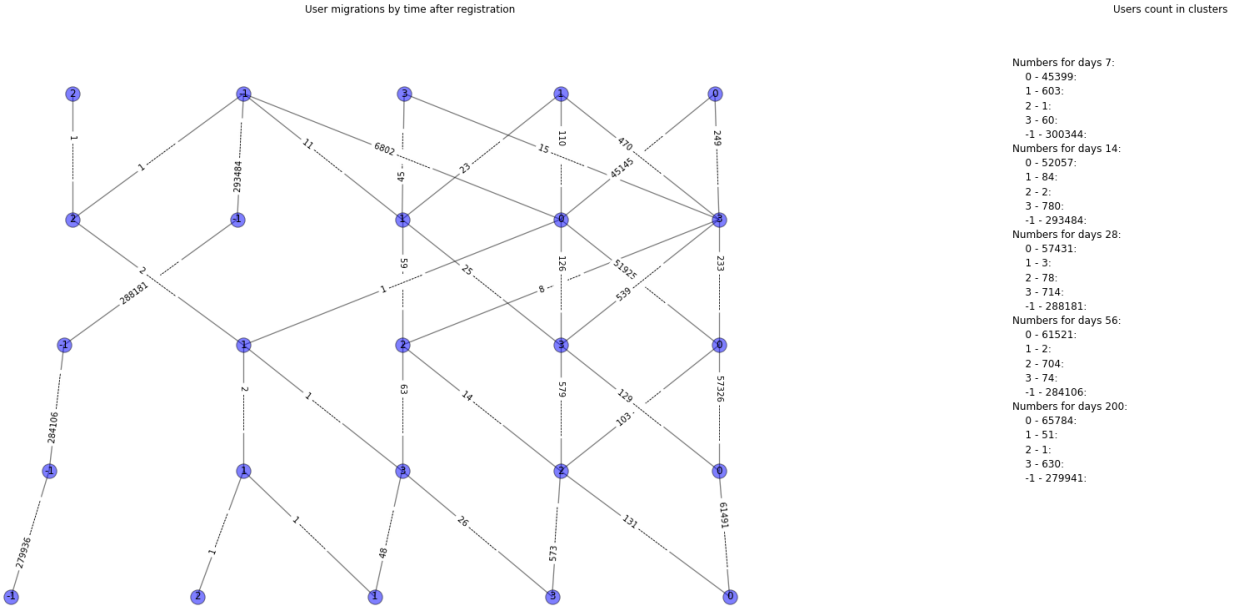

Рисунок 7.