Санкт-Петербургский государственный университет
математико-механический факультет

Кафедра системного программирования

Поиск спамеров на основе анализа пользователей сервиса Youtube

Курсовая работа студента 445 группы
Филатова Владимира Константиновича

Научный руководитель …… Владимир Суворов
EMC Санкт-Петербург

2013

Оглавление

1. Введение

Современное общество постоянно использует интернет не только для того, чтобы получать некоторую информацию, но и для того, чтобы делиться своей. На данный момент есть много социальных сетей, видео хостингов, файловых хостингов и других различных ресурсов сети с большими объемами информации и базами данных. Количество таких ресурсов растет постоянно, так как меняется мир и меняются информационные технологии. Не понятно, что нас ждет через десять, пятнадцать или двадцать лет. Люди каждый год придумывают что-то новое. Где-то экономят на памяти, где-то на пространстве хранения, на скорости вычислений и прочих других алгоритмах.

Но самое главное, большинство людей оставляют подробную информацию о себе и выкладывают много различного контента на различных ресурсах сети. Эта информация может быть об увлечениях человека, его образовании, его местах путешествий и других сферах его деятельности. Поэтому многие люди и компании решили выделять информацию из таких ресурсов.

В связи с большим количеством таких компаний и ресурсов, появилось новое направление в компьютерной индустрии, такое как data mining. Этот термин ввели еще в 1989 году, и он предполагает выделение некоторых «скрытых» данных из уже имеющихся.

Основу методов Data Mining составляют всевозможные методы классификации, моделирования и прогнозирования, основанные на применении деревьев решений, искусственных нейронных сетей, генетических алгоритмов, эволюционного программирования, ассоциативной памяти, нечёткой логики. К методам Data Mining нередко относят статистические методы (дескриптивный анализ, корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ, компонентный анализ, дискриминантный анализ, анализ временных рядов). Такие методы, однако, предполагают некоторые априорные представления об анализируемых данных, что несколько расходится с целями Data Mining (обнаружение ранее неизвестных нетривиальных и практически полезных знаний).

Подразделом этого направления является направление community mining, которое предполагает выделение сообществ людей из базы информации о них. Например, из какой-нибудь социальной сети можно выделить людей работающих в одной компании, или людей болеющих за одну футбольную команду и т.п. Причем выделять сообщества можно не только из социальных сетей.

Большим развивающимся компаниям необходимо знать кучу информации о своих сотрудниках, так как все работники должны работать в комфортных для них условиях, чтобы эффективность труда была высокая. Например, если в компании все люди увлекаются футболом, то можно организовать турнир между отделами, или сделать

сборную команду для игр против других компаний. Если часть сотрудников любит выезжать на природу, то можно сделать корпоративные выезды.

Направление community mining занимается сбором такой статистики. Все сотрудники могут оставлять ссылки на свои странички в социальных сетях, и компания может искать необходимые группы людей по какому-нибудь направлению.

2. Постановка задачи

Для начала определим несколько понятий, которые нам пригодятся.

Спамер – пользователь ресурса, который неоднократно нарушает правила поведения. Например, на видеохостингах он может выкладывать видео порнографического характера, на различных форумах писать однообразные сообщения не по тематике и др.

Таких пользователей в интернете на различных ресурсах стало очень много. Они есть практически на всех ресурсах, и очень часто их ищут вручную. Например, пользователи указывают, что какой-то пользователь является спамером, и когда наберется довольно много жалоб, администраторы ресурса его заблокируют, пока не прояснят, является он спамером или нет. Но пока люди не скажут, что он спамер, его информация и сообщения спокойно лежат на ресурсе, и её может увидеть кто угодно.

Такие люди не всегда выкладывают запрещенный контент или вредят ресурсу просто так. В основном они преследуют личные мотивы для продвижения своего сайта, чтобы поисковая машина при запросах выдавала его сайт в числе первых строчек, что иногда мешает, так как в основном эти сайты не есть хорошие.

Главная проблема, порождаемая поисковым спамом, заключается в том, что он генерирует множество мусорного контента, затрудняя эффективную работу поисковых серверов, искажает объективное ранжирование интернет-ресурсов и релевантность поисковых результатов. В итоге это во многом обесценивает Интернет как источник получения объективной информации.

Нежелательно, чтобы на видеохостинге по поиску «смешарики» выдавалась бы видео, содержащее сцены порнографического характера, насилия и другого запрещенного контента. Необходимо как-то быстро кластеризовать пользователей, которые нарушают правила ресурсов и понять вообще, можно ли это сделать: можно ли найти у спамеров ресурса одинаковые характеристики. То есть необходимо определить, обладают ли спамеры одного какого-нибудь ресурса одинаковым поведением.

Введем еще одно определение – «спам фанат». Спам фанат – пользователь, который следит за контентом, который выкладывают спамеры. Например, он может смотреть видео, которые выложил спамер, комментировать, подписываться и другие вещи разрешенные видеохостингом.

Возникает гипотеза, а не соединены ли спамеры и спам фанаты в одну социальную сеть? Можно ли по группе спамеров, через связи со спам-фанатами найти других спамеров?

Рассмотрим известный сервис в интернете с названием Youtube.

Youtube — сервис, предоставляющий услуги видеохостинга. Пользователи могут добавлять, просматривать и комментировать те или иные видеозаписи. Благодаря простоте и удобству использования YouTube стал популярнейшим видеохостингом и третьим сайтом в мире по количеству посетителей. В январе 2012 ежедневное количество просмотров видео на сайте достигло 4 млрд. На сайте представлены как профессионально снятые фильмы и клипы, так и любительские видеозаписи, включая видеоблоги.

На Youtube есть пользователи, которые являются спамерами. Раньше это были пользователи, которые выкладывали порно видео, сейчас же это в основном пользователи, которые выкладывают ссылки на другие ресурсы.

Наверняка на Youtube уже есть готовые решения, чтобы фильтровать спамеров на сервисе, но спамеры там есть, а значит можно попробовать сделать фильтрацию лучше, и попробовать дофильтровать тех спамеров, которые там есть сейчас.

Моей задачей является построение классификатора, который мог бы по определенной группе пользователей сервиса Youtube определять, является он спамером или нет. А также проверить соединены ли спамеры и спам фанаты в одну некую социальную подсеть. То есть определить, можно ли по одним спамерам найти других.

Для основы я буду брать данные, которые мне сможет предоставить сервис Youtube (http://youtube.com) через свой API. Он может предоставить информацию о выложенном видео, информацию о пользователях: кто комментирует видео, список подписчиков и прочее другое.

Результатом будет являться приложение, которое по заданной группе спамеров выдаст список связанных с ними спамеров (для подтверждения гипотезы про «социальную сеть» спамеров), а так же классификатор, который по группе пользователей будет делить их на спамеров и не спамеров по определенным критериям.

3. Обзор существующих решений.

Проблема поиска спамеров не нова. Спамеры появились уже давно, и после того, как они появились, многие компании начали искать решения. В основном для каждого ресурса необходимо писать свои методы определения спамеров, так как у каждого сайта свои методы хранения, свои пользователи, своди данные. Например, спамеры на видеохостинге могут отличаться от спамеров на twitter.

Существует два основных подхода выделения спамеров: Это изучение информации (например сообщений) и фильтрация IP адрессов. Но эти оба подхода имеют свои

недостатки. Например спамеры могут изменять сообщения и свои IP адреса динамически, для того, чтобы их было трудно отследить.

Есть похожая работа, написанная четырьмя людми: **Sarita Yardi, Daniel M. Romero, Grant Schoenebeck и Danah Boyd,** где они попытали кластеризовать пользователей на Twitter, а также посмотрели, как связаны «плохие» и «хорошие» пользователи. Им удалось показать, что спамеры на Twitter связанны друг с другом через других полоьзователей, а также построили классификатор, который на 300 пользователях пропустил 27 спамеров и сделал 12 хороших пользователей спамерами.(http://firstmonday.org/ojs/index.php/fm/article/view/2793/2431)

На сайте Youtube скорее всего есть какие-то методы решения задачи поиска спамеров. После того как они изменили дизайн, они избавились от многих пользователей, которые распространяли порно в чистом виде. Но спамеры все равно остались, и захотелось как-то улучшить то, что они сделали.

В основном на всех крупных ресурсах интернета, пользователи сети сами могут сообщать о проблемах. Например, на сервисах Youtube и Вконтакте есть кнопка «отметить как спам». При достаточном наборе жалоб, администраторы сайта временно заблокируют пользователя, и при подтверждении того, что пользователь является спамером, удалят его.

Некоторые ресурсы ищут спамеров вручную, сажая специально обученных людей, которые смотрят за всем, что выкладывают пользователи на их сайт. Им приходится просматривать много страниц, что-то постоянно обновляется и необходимо как-то помочь им быстрее находить некоторый спам, чтобы на ресурсе его стало меньше.

4. Этапы решения задачи

В качестве основного языка программирования был выбран язык Java, так как он является платформенно независимым, объектно-ориентированным языком программирования. Под этот язык Youtube предоставляет API для того, чтобы разработчики могли использовать сервис через свои приложения, например, добавлять видео, комментировать и др. вещи. В ходе данной работы мы будем использовать API сервиса для получения информации о пользователях и видео, которое они загрузили.

Для начала необходимо было накачать данные с сервиса, чтобы можно было составить статистику и пытаться кластеризовать их.

Сбор данных из сервиса Youtube.

Посмотрим, что нам может предоставить сервис Youtube через свое API.

а) Информация о пользователе

Youtube API предоставляет обширную информацию о пользователе, которую тот заполнил: имя, никнэйм, возраст, хобби и пр. Эта информация была актуально до того как в 2013 году сервис поменял свой дизайн и эту информацию заполняли. Но спустя время Youtube понял, что на их сервисе это мало кому надо. Поэтому отсюда нам пригодится только информация о никнейме пользователя, чтобы иметь уникальное поле пользователей сервиса. Раньше была полезна информация о дате последнего захода, из которой мы могли бы отсеивать тех людей, которые заходили, например, больше года назад. Сейчас у всех стоит одна дата.

б) Информация о видео

Простой запрос:

```
        String feedUrl =
"http://gdata.youtube.com/feeds/api/users/GoogleDevelopers/uploads";

VideoFeed videoFeed = service.getFeed(new URL(feedUrl), VideoFeed.class);
```

выдаст список видео, которое выложил определенный пользователь username на свой канал. Так можно просмотреть все видео, которое у него есть и информацию о них.

в) Комментарии

Комментарии являтются одними из данных для подтверждения гипотезы про сеть спамеров и спам-фанатов. Наверняка тот, кто смотрит запрещенное видео, смотрит его не у одного пользователя. Благодаря комментариям мы можем построить связи между некоторыми пользователями. Так же по комментариям раньше можно было найти поисковык спамеров (те кто выкладывает ссылки своих сайтов для продвижения), но Youtube сделал политику запрещения выкладывания ссылок в комментарии.

```
String commentUrl = videoEntry.getComments().getFeedLink().getHref();

CommentFeed commentFeed = service.getFeed(new URL(commentUrl),
CommentFeed.class);
```

Из результатов запроса мы можем узнать автора оставленного и комментария и текст самого комментария.

г) Подписчики

Благодаря библиотеке можно было узнать, на кого подписан пользователь. Человек, который распространяет запрещенный контент, наверняка подписан на других, которые тоже его распространяют. После того, как Youtube поменял дизайн, эта функция не поддерживается. Так как наверно они думают, что это никому не нужно.

```
String feedUrl =
"http://gdata.youtube.com/feeds/api/users/GoogleDevelopers/subscriptions";


SubscriptionFeed feed = service.getFeed(new URL(feedUrl),
SubscriptionFeed.class)
```

д) Новостная лента пользователя.

Благодаря этой ленте мы можем узнать активность пользователя на сервисе, что он делал и когда: комментировал, выкладывал, подписывался и многое другое.

```
String feed =
"http://gdata.youtube.com/feeds/api/events?author=users";UserEventFeed
activityFeed = service.getFeed(new URL(feedUrl), UserEventFeed.class);
String title = activityFeed.getTitle().getPlainText();


....
```

На самом деле Youtube обладает не очень удобным API для этой работы. Все запросы ограничены и дают ограниченный результат набором не более чем 25 строк, в связи с этим постоянно приходится писать различные итераторы для того, чтобы полностью получить более точную и подробную картину данных пользователей сервиса Youtube

После того как мы определи, какие данные есть данные на сервисе, попробуем проверить гипотезу: «соединены ли спамеры и спам-фанаты в одну сеть».

Гипотеза про «социальную сеть»

До того как Youtube не поменял свой дизайн и не почистил весь контент, там были выложены видео порнографического характера, которое люди комментировали, а значит интересовались им. Те кто выкладывал это видео, мы назовем спамерами, а те, кто подписывался на них, комментировал, добавлял в избранное мы назовем «спам-фанатами». Эти множества могут пересекаться. Любой спамер может быть спам-фанатом и наоборот. Здесь нет четкой границы. Попробуем определить связаны ли как-то одни спамеры с другими через спам-фанатов или других спамеров. Можно ли по одной группе спамеров отыскать других.

Для подтверждения или опровержения гипотезы необходимо было написать приложение, которое могло бы соединять пользователей в один граф, где вершинами были бы сами пользователи, а ребрами были бы комментарии, избранное, подписчики и пр.

Приложение должно использовать базу данных. В качестве СУБД было решено выбрать Mysql, так как она является наиболее приспособленной для web приложений.

Также преимуществами MYSQL является многопоточность, быстрая работа, масштабируемость, бесплатность, интерфейс с языком JAVA и другими языками.

База данных состоит из двух основных таблиц:

1) people – информация о пользователях на сервисе YouTube
2) videos – информация о видео на сервисе

И трех соединительных таблиц

3) subscribers – таблица, соединяющая ID пользователей и ID их подписчиков
4) uploads – таблица, соединяющая ID пользователей и ID видео, которое они загрузили на сервис
5) comments – таблица, соединяющая ID пользователей и ID видео, которое они прокомментировали

Схема базы данных представлена на Рис.1

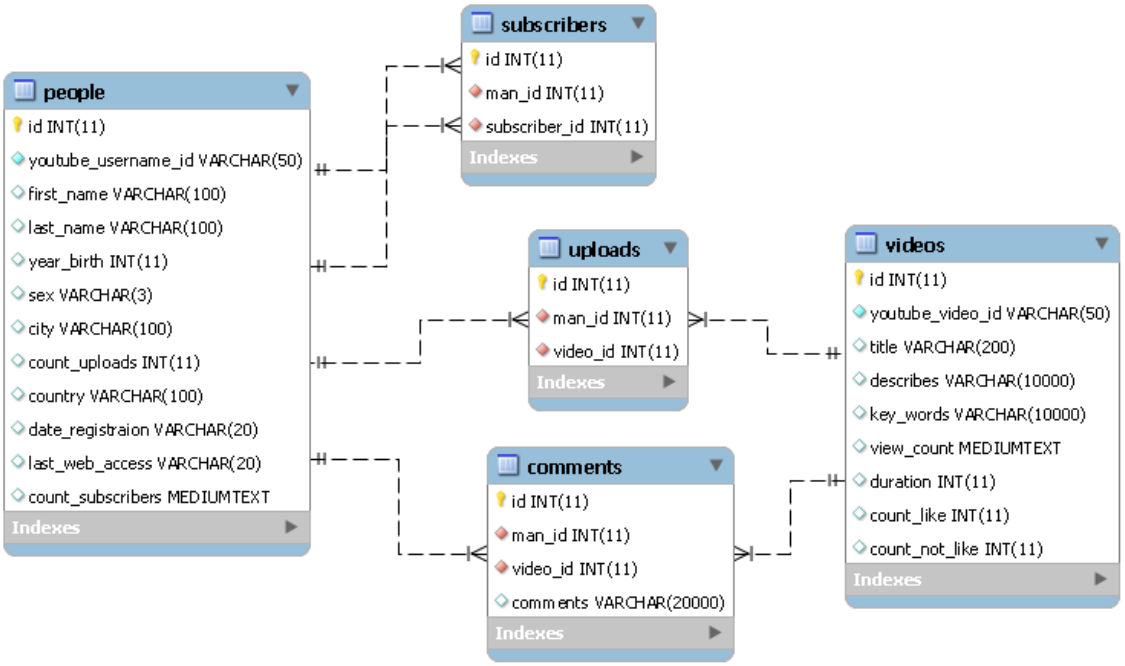

Рис 1.

В качестве приложения были написаны два сервлета, которые взаимодействуют между собой. В принципе это мог бы быть и один сервлет, но так как библиотека для JAVA сервиса Youtube, конфликтовала с библиотекой для работы с графами, пришлось разделить приложение на два.

Первый сервлет получает на вход список спамеров и начинает парсить их. Сервлет скачивает с сервиса информацию о видео пользователей, кто комментировал это видео, на

кого подписаны пользователи и др. Всю информацию он скачивает в базу данных, описанную выше.

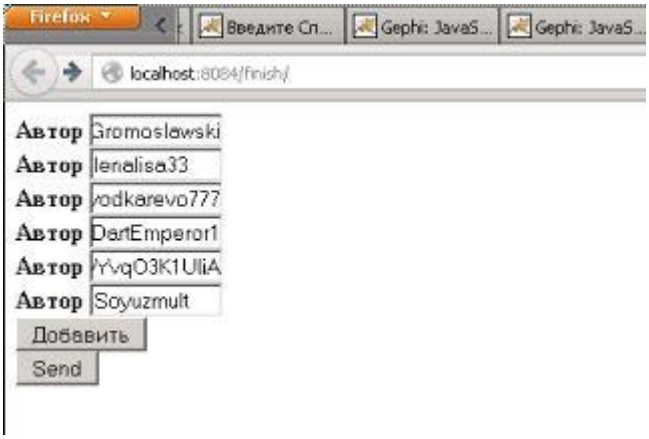

Рис 2.

Второй сервлет берет данные из базы и строит граф пользователей. После того как граф построен, сервлет визуализирует его в браузере для наглядности отображения графа, чтобы можно было посмотреть, какой пользователь с каким связан. Все данные графа сводятся в общую таблицу, по которой можно посмотреть информацию о пользователях.

Поиск спамеров на сервере YouTube

<< < Страница 1 из 23 > >> На странице 20 строк Установить

User	CountUploads	Cluters1	Clusters2	Count Subscribers	SameDescribes	degree	closeness	SameCommt
theyoungturks	14511	1	1	793425	0.0	19	1.0	1
moviemaniacsde	4276	1	1	97103	0.16	1	1.0	2
collegehumor	1934	1	1	4156550	0.0	29	1.0	5
psychetruth	1697	1	1	238864	0.0	5	1.0	37
satanrulezzzz	1436	1	1	33716	0.08	1	1.0	14
tytuniversity	1379	1	1	82439	0.0	2	1.0	86
xtraonline	717	1	1	6329	0.0	1	1.0	1
natgeowild	709	1	1	152245	0.0	13	1.0	0
jpizzle1122	666	0	1	604031	0.0	1	1.0	1351
hiphopnewzcentral	530	1	1	1223	0.0	1	1.0	0
killuminaton	486	1	2	92	0.0	1	0.0	2
gamebombcentral	435	1	1	203391	0.64	8	1.0	28
mrchelios47	324	1	1	594	0.0	1	1.0	0
thebreakingnews0	315	1	1	1055	0.04	1	1.0	0
townsquare	255	1	1	34325	0.0	2	1.0	2
microprikol	201	1	1	165724	0.0	3	1.0	7
thisishorosho	200	1	1	1576224	0.0	3	1.0	0
ikrazeyyhd	178	1	2	488	0.0	1	0.0	1
azatgw	165	1	1	4884	0.24	1	1.0	54
miltonciyo	158	1	1	1139	0.0	1	1.0	1

Рис 3.

Для работы с графами был использован фрейморк Gephi, который обладает своим API для разработчиков на Java. Он позволяет за несколько секунд генерировать граф из базы данных, где может быть более 200000 вершин и 500000 ребер. Если добавлять к составлению графа дополнительные условия, то он будет генерировать дольше. Но его основным плюсом является то, что он может создавать графы из базы mysql. Надо просто

запросом указать ему, что является вершинами в графе, и какие вершины между собой соединены. Также под этот фрейморк есть хороший html визуализатор. Для отображения графа достаточно подставить файл gexf, в который можно легко импортировать из JAVA API Gephi toolkit. Основной минус фреймворка – это то, что библиотека конфликтует с библиотекой Youtube API.

В ходе тестирования, до того как Youtube не поменял дизайн, действительно подтвердилась теория о том, что спамеры как-то соединены в сеть. По набору одних спамеров удалось найти других через связи с ними.

Теперь на Youtube нельзя оставлять ссылки в комментариях, и на сервисе введено много других ограничений. Сервис ввел какую-то проверку, и на сайте нет видео, которое содержит, например, порнографию. Но это не значит, что там нет спамеров, появились другие ухищрения.

Так как же можно искать спамеров на сервисе? Было решено попробовать написать классификатор, который выделял бы из группы пользователей спамеров по различным параметрам.

Для начала определим, какие спамеры есть на сервисе сейчас.

Виды спамеров на сервисе

Спамеров на Youtube можно разделить на несколько категорий:

а)Новые порно спамеры – те которые выкладывают видео с надписью смотри ниже, и в описании кидают ссылку на порно сайт.. Они выкладывают видео за 1-2 дня, поэтому дата выкладки видео у них отличается на 1-2 дня. Описание к видео – http://url.

http://www.youtube.com/user/kroe0825/feed

http://www.youtube.com/user/1308771000?feature=watch

http://www.youtube.com/user/manjykantanransovan?feature=watch

б) Старые порно спамеры, возможно просто кто-то их взломал… Имеют нормальное видео, нормальную активность, может в последнее время выложено нормальное видео без спама… Имеют видео выложенное давно с описанием как в пункте а). Время выкладывания видео может быть разным…

http://www.youtube.com/user/defiancenl

http://www.youtube.com/user/pizza38921111

в) Пользователи не порно спамеры, которые имеют разное описание, но на одну тематику, не обязательно начинается со ссылки, но имеют её. Также проявляют нормальную активность, комментируя видео, добавляют избранное, подписываются и многое другое.

http://www.youtube.com/user/MihailAveryanovRu

http://www.youtube.com/user/alekc323/videos?view=0&flow=list&sort=dd

http://www.youtube.com/user/CigarettesCheap/videos?flow=list&view=0

http://www.youtube.com/user/ShaimovDinislam/

г) Пользьзователи, которые не имеют ссылок в описании, начало описания нормальное – а концовка повторяется у всех. Дата выкладывания видео одна и та же. (В основном отличается на 1-2 дня)

http://www.youtube.com/channel/UCfXtn_ldSxFMxIl2m0WAtsg

д) Пользователи, которое имеют разное описание, ссылку выкладывают прямо в видео, время выкладывания видео как в пункте а) 1-2 дня

е) Пользователи порно спамеры, которые не имеют ссылку вначале описания, но она есть, активность только порно, и это старые спамеры…

http://www.youtube.com/channel/UC_My_zk2PUYyOmDZvKLUleg?feature=watch

з) люди имеют нормальное видео, картинка видео похожа порно, название видео порно, но описание может быть нормальным, само видео тоже может быть нормальным

http://www.youtube.com/user/bilnik8/videos?flow=list&view=0&sort=dd

и) Люди, которые выкладывают видео, ко многим видео одинаковое описание, без ссылок, просто одинаковое…

http://www.youtube.com/user/supermorgan66/videos?flow=list&view=0&sort=dd

Некоторые типы спамеров могут пересекаться, но это основные типы спамеров, которые удалось найти на сервисе.

Параметры кластеризации

После того, как определили типы спамеров, необходимо было определить, какими похожими характеристиками обладают спамеры.

а) Ссылки на другие сайты в описаниях

Первое, что сразу бросилось в глаза – это то, что практически у всех спамеров в описаниях есть ссылки на сторонние ресурсы, не относящиеся к сервису Youtube или какому-нибудь другому известному сервису. Некоторые просто так распространяли порно сайты, некоторые продвигали свои сайты. Спамеров второго типа обычно называют

поисковыми спамерами. Они оставляют ссылки для повышения релевантности ресурса в различных поисковых системах.

Попробуем посчитать отношение количества видео со ссылками ко всему количеству видео, выложенного каждым пользователем. Составим таблицу распределения для выборочной таблицы пользователей сервиса (около 400 пользователей)

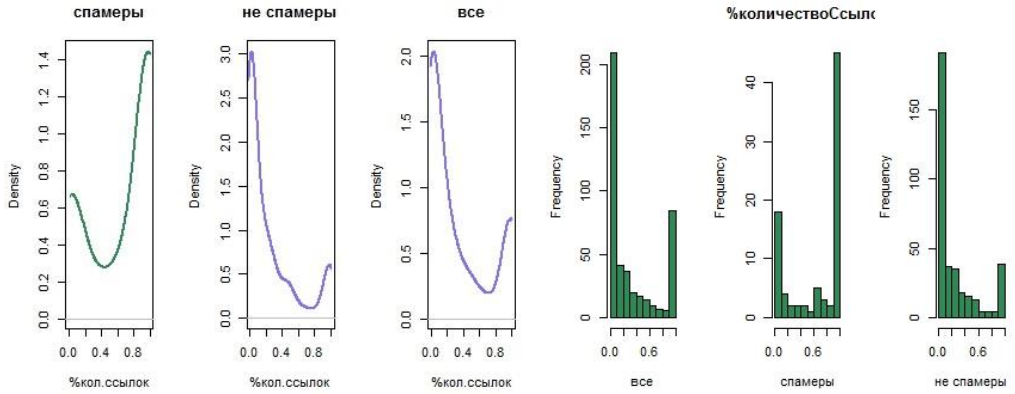

Рис 4.

Из гистограмм и графиков видно, что этот параметр может помочь при кластеризации пользователей, так как распределение вероятностей для спамеров и не спамеров противоположны друг другу.

б) Одинаковое описание ко многому видео.

Вторым параметром, который сразу бросился в глаза, является отношение количества видео, с одинаковым описанием к количеству видео, выложенном пользователем. При поиске спамеров было замечено, что у многих из них описание к видео совпадает, из чего следует, что, скорее всего они являются спам-ботами. Графики и гистограммы приведены на рис 5.

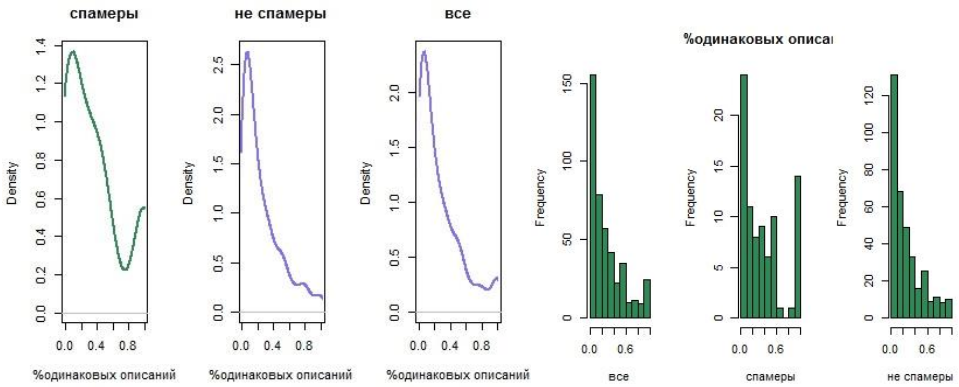

Рис 5.

В) Количество видео выложенное за 1-2 дня.

Еще одним основным параметром можно выделить отношение количества видео, выложенное за 2 дня подряд ко всему количеству видео, выложенному пользователем. Скорее всего, это не пользователи, а спам-боты, которые «засоряют» Youtube.

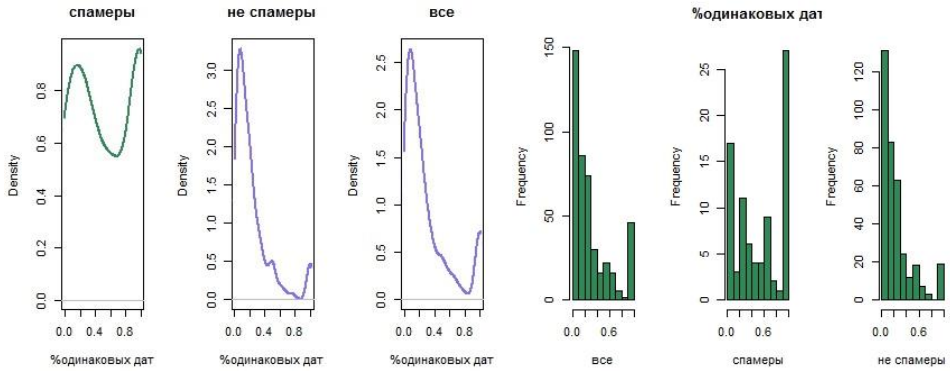

Рис 6.

г) Фид загрузки

И последний не бинарный классификатор, который удалось выделить это процент фида загрузки, то есть это отношение количества фида с названием «загружено видео» ко всему количеству видео. Если это число близко к нулю, то можно определить, что пользователь имеет некоторую активность на сервере: комментирует видео, добавляет избранное и др. У пользователей-спамеров, которых удалось найти вручную было выявлено, что это отношение стремится к единицы. Это значит что спамеры практически не имеют никакой активности, кроме того, что выкладывают видео.

К сожалению Youtube не дает весь список Фида пользователя, а только последние 150 записей, но как показала практика, этого достаточно, чтобы определить активность пользователя на сервере.

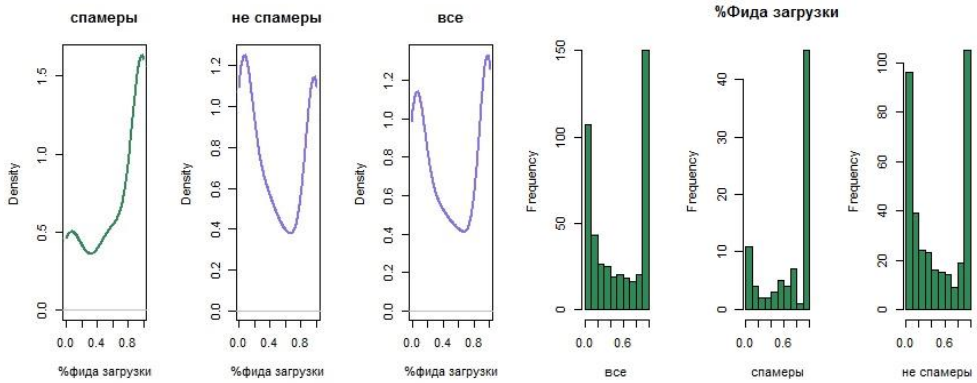

Рис 7

В добавок к 4 основным метрикам, описаных выше, были еще выделены дополнительные бинарные метрики

д) Количество ссылок больше 3

е) Количество одинаковых описаний больше 3

Также дополнительно был создан словарик часто встречающих спам-слов. Были пропарсены описания к видео и потом составился словарик из наиболее встречающихся слов, длина которых больше трех (чтобы не было предлогов как в русском, так и в английском языке). Таким образом добавилось еще три метрики:

ж) Отношение спам слов в названии и в описании к суммарному количеству слов в них.

з) Для спамеров был создан отдельный словарик не совместимых слов, построенных на словарике из пункта ж), в котором содержатся слова не совместимы по описани. (например Порно ДТП, ржач и т.п), и они есть в списке часто встречающихся слов. Если количество слов было больше ставили 1. (Например порно-дтп).

е) для порно спамеров был создан еще один словарик наиболее употребляемых слов по терме порно. Если количество слов было больше 3 ставилась 1.

Число 3 в этих метриках сначала было выбрано случайно. Но в результате тестирования в дальнейшем оказалось, что оно дало лучшую классификацию группы пользователей на спамеров и не спамеров.

Еще было решено не выделять пользователей, у которых количество выложенного видео меньше трех, из-за боязни ухудшения результатов. В дальнейшем есть возможность рассмотреть еще и их.

Для того чтобы разделить пользователей на две части, необходимо было написать программу, которая бы использовала алгоритм кластеризации. Необходимо дать этой программе набор пользователей, а она сказала бы, кто из них является, а кто нет. Рассмотрим три алгоритма кластеризации, которые возможно помогут нам при определении спамеров. Один из них (Kmeans) является алгоритмом обучения без учителя. Вдобавок к нему, попробуем построить два алгоритма обучения с учителем (Наивный Байесовский классификатор и SVM), и проверить, какой из этих алгоритмов поможет нам определить спамеров наилучшим образом.

Алгоритмы кластеризации

1. Kmeans

Kmeans является наиболее популярный методом кластеризации. Он был изобретён в 1950-х годах математиком Гуго Штейнгаузом и почти одновременно Стюартом Ллойдом. Особую популярность приобрёл после работы Маккуина. Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров.

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

Он разбивает множество элементов векторного пространства на заранее известное число кластеров k.

Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике.

Алгоритм завершается, когда на какой-то итерации не происходит изменения кластеров. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом шаге суммарное квадратичное уклонение V уменьшается, поэтому зацикливание невозможно.

В ходе тестирования классификатору давался различный набор параметров. Для начала ему дали все параметры, описанные в части про метрики, кроме словарика порно спамеров, так как решили прокластеризовать всех, а не только порно спамеров. В результате получили число кластеризации true-negative (отношение не найденных спамеров ко всем спамерам) равным 0.3, а число false-positive(FP) (отношение не спамеров в найденных спамерах получилось) 0.25 Эти данные были получены на выборке из 400 пользователей. Захотелось как-то улучшить результаты. Путем различного выбора параметров было получено, что оптимальными являются параметры а)-ж), и удалось улучшить кластеризацию в два раза. Это пока максимальный результат, который удалось получить.

В результатах kmeans дал лучшие результаты в сравнении с другими классификаторами, которые использовались.

2. Наивный байесовский классификатор

Наивный байесовский классификатор — простой вероятностный классификатор, основанный на применении Теоремы Байеса со строгими (наивными) предположениями о независимости.

Теорема Байеса:

$$p(C|F_1, \ldots, F_n) = \frac{p(C)\ p(F_1, \ldots, F_n|C)}{p(F_1, \ldots, F_n)}.$$

Её смысл на обывательском уровне можно выразить следующим образом. Теорема Байеса позволяет переставить местами причину и следствие. Зная с какой вероятностью

причина приводит к некоему событию, эта теорема позволяет расчитать вероятность того что именно эта причина привела к наблюдаемому событию.

Цель классификации состоит в том чтобы понять к какому классу принадлежит документ, поэтому нам нужна не сама вероятность, а наиболее вероятный класс. Байесовский классификатор использует оценку апостериорного максимума (Maximum a posteriori estimation) для определения наиболее вероятного класса. Грубо говоря, это класс с максимальной вероятностью.

$$c_{map} = \text{argmax}_{c \in C} P(d|c)P(c)P(d)$$

То есть нам надо рассчитать вероятность для всех классов и выбрать тот класс, который обладает максимальной вероятностью.

Необходимо попробовать «скормить» классификатору 90% пользователей, и посмотреть, как он определит оставшиеся. В качестве обучающего множества классификатору давались различные 320 пользователей, где было указано, кто является спамером, а кто нет. В результате получили значения False-Positive где-то около 0.6, а True-Negative около 0.8, что говорит об ужасной кластеризации.

3. SVM

Метод опорных векторов - это метод машинного обучения, целью которого является попытка классифицировать входные наборы данных в один из двух классов. Для эффективной работы метода сначала необходимо использовать обучающую выборку, состоящую из входных и выходных данных, которая необходима для построения модели метода опорных векторов, и которую в дальнейшем можно будет использовать для классификации новых данных.

Для построения модели метода опорных векторов нужно взять обучающие входные данные, отобразить их в многомерное пространство, а затем использовать регрессию, чтобы найти гиперплоскость (гиперплоскость - это поверхность в n-мерном пространстве, которая разделяет его на два подпространства), которая лучше всего разделяла бы два класса входных данных. После обучения модели она способна классифицировать новые входные данные в один из классов при помощи разделяющей гиперплоскости.

По существу, метод опорных векторов является методом входов/выходов. Пользователь вводит входные данные, и на основе разработанной (при помощи обучения) модели получает выходные результаты. Теоретически, число входов для метода опорных векторов лежит в диапазоне от одного до бесконечности. Однако, в практическом применении, есть определенные ограничения на размер входной выборки, которые зависят от вычислительной мощности. Например, пусть для конкретного применения метода опорных векторов используются N входов (N - целое число, в диапазоне от 1 до бесконечности). Тогда задача метода опорных векторов заключается в том, чтобы

сопоставить все входные данные размерности N и найти такую гиперплоскость размерности N-1, которая наилучшим образом разделяла бы обучающую выборку.

К сожалению этот алгоритм, так же как и Байесовский классификатор не удалось обучить так, чтобы он при тренировочном множестве 90% давал бы на оставшихся 10% FP < 0.5 и TN < 0.5. Возможно, эти два классификатора необходимо обучать на множестве не из 400 человек а из 400000, и проверить на оставшихся 2000 являются они спамерами или нет. Но для этого надо вручную просмотреть 400000 страниц, чтобы определить, кто является спамером, а кто нет,

5. Заключение

В ходе данной курсовой работы было написано приложение, которое проверило гипотезу о том, что порноспамеры и спам-фанаты соединены в одну сеть. По определенной группе порноспамеров можно было найти других спамеров, которое выкладывают порно, через связи со спам-фанатами, которые комментируют видео, добавляют в избранное или подписываются на канал.

Также в ходе работы был написан классификатор, который по группе пользователей пытается разделить выделить из них спамеров, применяя различные метрики, описанные выше в данной работе.

Для тестирования была взята выборка из 450 пользовательных подозрительных на соответсвие спамерам. Затем в этой группе пользователей были определены спамеры для проверки классификатора. Были построены три алгоритма классификации: Kmeans, Байесовский классификатор и SVM. В результате тестирования, было определено, что лучше всего с задачей справился KMeans, У которого получилось 7 ложных срабатываний (то есть он определил хорошего пользователя как спамера) и 11 спамеров он определил как хороших пользователей.

Результат классифицирования методам KMeans представлен на рис. 8

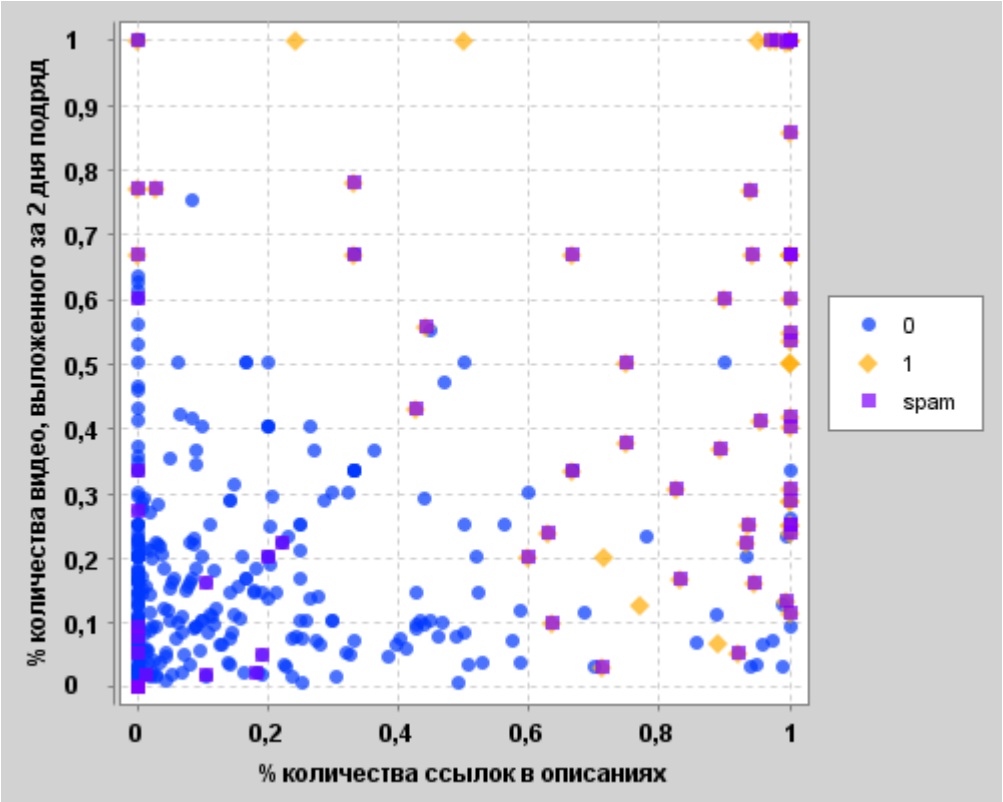

рис. 8

Для оценки классификации были посчитаны числа оценки кластеризации.

1)false-positive(FP) - отношение не спамеров в найденных спамерах получилось 15 %

2)true-negative(TN) - отношение не найденных спамеров ко всем спамерам равно 15 %

Эти два числа говорят о высокой точности качества кластеризации (85% найденных спамеров)

Дальнейше развитие предполагает уменьшение количества ненайденных спамеров (TN) и уменьшение количества пользователей, которых классификатор характеризовал как спамер, а они таковыми не являются (FP). Уменьшение FP является приоритетной задачей, так как лучше не найти спамера, чем назвать спамером, какого-нибудь честного пользователя.

6. Список используемой литературы:

1. http://ru.wikipedia.org/wiki/Data_mining - информация о Data Mining

2. https://gephi.org - информация о фреймворке для работы с графами Gephi

3. http://youtube.com – Сервис Youtube

4. A Tutorial on ν-Support Vector Machines, Pai-Hsuen Chen, Chih-Jen Lin, Bernhard Scholkopf, Department of Computer Science and Information Engineering,National Taiwan University Taipei 106, Taiwan 2 Max Planck Institute for Biological Cybernetics, Tubingen, Germany. – статья об SVM

5. http://en.wikipedia.org/wiki/K-means_clustering - алгоритм Kmeans

6. https://en.wikipedia.org/wiki/Naive_Bayes_classifier - Байесовский классификатор

7. http://www.mysql.ru/ - информация о MySql

8. http://firstmonday.org/ojs/index.php/fm/article/view/2793/2431 - статья по обнаружению спамеров на twitter.