Оглавление

1.1 Введение

В настоящий момент довольно остро стоит вопрос о сохранении тайны связи при использовании электронной почты, чата, социальных сетей и иных электронных средств коммуникаций. Закон о сохранении тайны связи не охватывает публичные сервисы.[1] Помимо этого, опубликованные Эдвардом Сноуденом данные наглядно демонстрируют, что межправительственные системы слежения (созданные для борьбы с терроризмом) используются для достижения экономических и политических целей, нарушая права граждан на тайну частной жизни и тайну переписки.[2] [3]

Люди часто недооценивают значение метаданных и комплексного анализа. Под комплексным анализом далее будет подразумеваться сочетание методов, подходов, инструментов по интеллектуальной добыче данных (Data Mining), обработки больших объемов неструктурированных и слабосвязанных данных (Big Data).
Зачастую остается неосознанным тот факт, что вступая в электронные сообщество человек переносит элементы реального общения в виртуальное. На первый взгляд может показаться, что изменяется лишь форма взаимодействия, но на самом деле меняется его суть. К примеру, при реальной встрече двух людей, если это встреча не была кем либо еще подслушана, то она остается известной только участникам и тем, кого они информировали об этом. В социальной же сети сервер выступает безоговорочным трентом (доверенным центром), который хранит информацию обо всех происходящих событиях.[4]. Таким образом, личная жизнь человека становится син-

[1] Тайна связи, электронная почта и российские суды (`http://www.securitylab.ru/blog/personal/emeliyannikov/37733.php`)

[2] Правительство США предало интернет. Нам надо вернуть его в свои руки (`http://habrahabr.ru/post/192852/`)

[3] Эдвард Сноуден (`http://ru.wikipedia.org/wiki/\T2A\CYRS\T2A\cyrn\T2A\cyro\T2A\cyru\T2A\cyrd\T2A\cyre\T2A\cyrn,_\T2A\CYREREV\T2A\cyrd\T2A\cyrv\T2A\cyra\T2A\cyrr\T2A\cyrd`)

[4] далее по тексту я предполагаю под событием любое взаимодействие с сервером социальной сети, например факт авторизации, просмотр страницы, написание сообщения, создание группы, комментарий и т.д. - в общем все то, что можно делать с помощью социальной сети

хронизированной с электронным сервером и человек, обладающий личностью, превращается в очередного пользователя системы. Так как одновременно происходят миллионы событий на первый план выходят метаданные: кто вступал во взаимодействия, когда, как долго, характер взаимодействия, метаинформация касательно события (количество символов в сообщение, количество участников встречи и прочее). Значительную часть метаинформации об отдельном человеке возможно проанализировать даже не являясь трентом, потому что сама природа социальных сетей заставляет пользователей делиться информацией (конечно если пользователь специально не установил некоторый уровень приватности).

Последние достижения в области обработки естественных языков позволяют автоматически извлекать факты из произвольного текста, написанного на естественном языке. Так в научной работе сотрудников IBM Research во главе с Джалал Махмудом (Jalal Mahmud), в которой демонстрируют возможность определить местонахождение человека по его постам в Twitter с точностью до 70% (определяется обычно город или округ).[5] Основная идея алгоритма, придуманного сотрудиками IBM, заключается в том, что само содержание сообщений, которые публикуются в Twitter, несет в себе информацию о местонахождении пользователя. Современные инструменты позволяют ее извлечь. Так, например, в посте может быть ссылка на фото или пост в другой социальной сети, в которой отмечена гео-информация. Кроме того, анализируется семантика текста для извлечения фактов. К примеру,из текста "Сергей не забудь самовар, встречаемся в Туле"можно извлечь следующие факты: Место - Тула, Объект - Сергей. Всю необходимую информацию исследователи извлекают напрямую из Twitter с помощью Streaming API, в основном используя GET statuses/firehose.[6])

[5]Who will retweet this?: Automatically Identifying and Engaging Strangers on Twitter to Spread Information (`digital.cs.usu.edu/~kyumin/pubs/lee14iui.pdf\begingroup\let\relax\relax\endgroup[Pleaseinsert\PrerenderUnicode{вЃ}intopreamble]`)

[6]подробнее см. Twitter Rest API (`https://dev.twitter.com/docs/api/1.1`

1.1.1 Цели и задачи дипломного проекта

Целью данного дипломного проекта является создание веб-приложения, демонстрирующего возможности по сбору сведений об отдельном человеке с использованием косвенных данных, полученных только из открытых источников. Особенно интересным представляется создать приложение для автоматизированного анализа страницы в социальной сети с выделением дополнительных сведений о человеке, на основе сведений о его друзьях. Веб-сервис создается как инструмент, демонстрирующий значимость косвенных данных в процессе сбора сведений о человеке или в случае, если пользователь хочет остаться анонимным.

Данный сервис задуман с целью проверки оценки уровня защищенности персональной информации, которую пользователь оставляет конфиденциальной, становясь участником виртуального сообщества, но которая может быть получена в результате анализа косвенных данных.

Данный сервис является попыткой защитить конечного пользователя, демонстрируя ему часть той информации, которую о нем могут собрать специальные службы.

Для реализации данных целей следовало решить следующие задачи:

- анализ легитимности функционала приложения;
- анализ существующих научных подходов для реализация данной задачи;
- анализ существующих web-сервисов, которые предоставляют дополнительную информацию о пользователе с помощью анализа косвенных признаков. Составление описания для каждого решения;
- реализация обязательного и дополнительного функционалов;

- тестирование и доработка приложения.

Веб-приложение не является социально опасным по следующим причинам:

- пользователь сервиса имеет возможность анализа только той страницы, для которой известны данные авторизации;
- сервис безопасен для пользователя т. к. авторизация происходит по средствам API социальной сети и данные авторизации не передаются на сервер приложения;
- мировой опыт показывает, что уже созданы куда более мощные средства для анализа данных. Однако, все они являются достоянием специальных служб.[7]

1.1.2 Основной функционал приложения

Обязательный функционал позволит определить пол, возраст, ВУЗ некоторого человека в социальный сети Вконтакте, на основе данных получаемых в автоматическом режиме. Дополнительным функционалом является:

- извлечение id пользователя из страницы пользователя (id не всегда указан явно);
- проектирование и реализация модуля для сбора статистических данных, подтверждающих предположение автора о существовании корреляции между персональными данными пользователя и его друзей (по таким категориям как пол, возраст, место учебы);
- проектирование и реализация модуля, отвечающего за обработку данных, собранных модулем статистики.

[7] Эдвард Сноуден на TED: Как нам вернуть Интернет (озвучка) (`http://habrahabr.ru/post/217079/`)

Оценка уровня конфиденциальности закрытых персональных данных пользователя на основе активности в социальной сети

1.2 Анализ существующих решений

Социальные сети занимают значительное место в жизни современных людей. И как следствие их огромной популярности, давно появились проекты, дополняющие их функционал.

Продукты, которые взаимодействуют с социальными сетями, можно разделить на следующие категории:

- продукты для SMM - автоматизируют работы по привлечению внимания к брендам через социальные сети;
- социальные агрегаторы - упрощают управление несколькими аккаунтами в социальных сетях. Как правило, позволяют настроить крос-постинг. Иногда присутствует функция собирания ленты, сообщений и комментариев с различных аккаунтов;
- сервисы анализа сообществ и трендов в социальных сетях - позволяют вести SMM на более высоком уровне, проверять эффективность собственных pr-компаний и отслеживать активность конкурентов;
- нативные приложения для социальной сети - дополняют функционал социальной сети. Значительную долю приложений этого класса занимают игры.

Основной целью данного анализа являлся сбор сведений о существующих решениях в отрасли(какой?) и об общих тенденциях в развитии(чего?). Автор не стремился составить список лучших или всех приложений определенного рода. В анализ существующих решений вошли только те продукты, которые показали интересные технические достижения, новаторство в от-

расли (какой?)или были интересны по другим причинам. Стоит понимать, что на данный момент количество сервисов и приложений, использующих интеграцию с социальными сетями,исчисляется уже сотнями тысяч и описывать их все не имеет смысла.[8]

1.2.1 smm-продукты

Такие проекты автоматизируют задачи с использованием инструментария, предоставляемого непосредственно социальными сетями. К примеру, публикация постов в определенное время или статистика популярности сообщений. Также продукты этого класса могут автоматизировать любые другие действия, упрощающие социальный медиа-маркетинг. (smm)[9]

1.2.2 социальные агригаторы

Кроме того, существуют сервисы, которые собирают информацию из разных социальных сетей, блогов и других ресурсов в один источник.[10] Стоит отметить, что не все сервисы четко вписываются в тот или иной класс приложений, потому как многие из них достаточно самобытны, быстро изменяются и иногда перестают существовать. Так, только за время подготовки дипломной работы перестали функционировать ряд сервисов:

- **twinfluence**

  был простым инструментом для измерения совокупного влияния твитов и их фолловеров. В качестве бонуса предо-

[8] имеется ввиду все приложения, которые имеют функцию интеграции с какой-либо социальной сетью

[9] Social media marketing (`http://ru.wikipedia.org/wiki/Social_media_marketing`)

[10] 20 Ways To Aggregate Your Social Networking Profiles (`http://mashable.com/2007/07/17/social-network-aggregators/`)

ставлял статистику некоторых социальных сетей. В данный момент недоступен по домену, на котором находился проект, стоит переадресация на компанию, в которой работают бывшие владельцы Twinfluence;

– **TweetEffect**

– отражал изменение количества фолловеров после каждого сообщения. Сервис перестал работать после изменения в twitter API;

– **Tweetoclock.com**

- помогал отследить время использования пользователями своего твиттер-аккаунта. В данный момент недоступен.

К самым интересным социальным агрегаторам, которые продолжают функционировать можно отнести:

– **Hootsuite**

- один из самых надежных и доступных инструментов, который постоянно совершенствует свой интерфейс и возможности. В настоящее время есть поддержка Twitter, Facebook Pages, Facebook, LinkedIn, Ping.fm, Wordpress.com, MySpace и Foursquare. HootSuite обладает функционалом, который позволяет настроить отправку поста на множество источников в несколько кликов.[11] Ключевыми характеристиками являются:

  * Планирование. Выбор между обновлением постов онлайн или по заранее заготовленному расписанию;

[11] 7 Social Media Aggregation Tools To Simplify Your Streams `http://socialmediatoday.com/SMC/192312`

* Гибкая работа с url. Добавление ссылок-счетчиков для отслеживания кликов и получение детальной информации об аудитории;
* RSS канал. Возможность добавить отправку постов в блоги и социальные медиа по RSS каналу;
* Закладки и аплет для браузера. Возможно использование фирменного аплета для браузера, чтобы быстро поделиться необходимой информацией.

Для данного исследования представляется наиболее важным выделить существующие методы получения информации и поиска в социальных сетях, в то время как остальные особенности сервисов отходят на второй план. Был проведен анализ существующих решений, выделен ряд приложений, которые с помощью косвенных данных и методов автоматического анализа позволяют:

– «вычислить» дополнительную информацию о человеке, которую он не указывал в явном виде;
– найти на web-ресурсах информацию недоступную обычным поисковым системам;
– получить релевантную информацию, которая обычно слишком низко ранжируется.

1.2.3 web-приложения для поиска людей

В сети Интернет представлен ряд приложений для поиска аккаунтов людей сразу во множестве социальных сетей. Стоит отметить, что в данный момент количество социальных сетей уже исчисляется десятками и это только те, которые имеют значительное (более нескольких миллионов) и живое сообщество.[12]

[12]Top 15 Most Popular Social Networking Sites (`http://www.ebizmba.com/articles/social-networking-websites`)

Существует большое количество CMS,конструкторов сайтов, которые позволяют достаточно быстро создать свою собственную социальную сеть или отдельный блог с интеграцией с другими блогами, построенными на той же технологии.[13] Все сети имеют свои особенности, поэтому агрегация этого многообразия - задача непростая, и ее можно решить несколькими способами. К основным проблемам, которые необходимо решить такому приложению, относятся:

- написание адаптеров для каждого источника информации;[14]
- решение вопросов разряженности данных (социальные сети обладают различным функционалом и данными о своих пользователях;
- скорость работы - агрегатор собирает информацию с других сервисов, а,значит, впадает в зависимость от скорости работы 3-их лиц, что не всегда может быть надежно.

Основными представителями web-приложений для поиска людей являются:

- **http://people.yandex.ru**

  `people.yandex.ru` – это специализированная поисковая вертикаль, с помощью которой возможно быстро находить размещенные в открытом доступе профили людей в социальных сетях. Для поиска не требуется регистрация в социальных сетях. Характерной чертой является то, что сервис очень бережно относится к персональным данных пользователей:

[13] 8 Great Social Networking CMS (`http://www.cmscritic.com/8-great-social-networking-cms`)

[14] конечно, существует Open API, но многие социальные сети имеют особенности, поэтому все-таки необходим индивидуальный подход

* не собирает и не хранит у себя никаких дополнительных данных о пользователе, лишь ищет и индексирует уже существующую информацию;
* индексирует только те профили, индексация которых не запрещена самим пользователем;
* индексирует только публично доступные данные, которые видны любому незалогиненному в социальной сети пользователю;
* склеивает только те профили, которые явно и публично ссылаются друг на друга (или в двух профилях проставлены взаимные ссылки друг на друга, или в одном из них есть провалидированная, т.е. требующая авторизации, ссылка на другой).

http://qwant.com

qwant.com — поисковая система с особым методами ранжирования и поиском по англоязычным социальным сетям (в этом она напоминает people.yandex.ru);

– **http://spokeo.com**

spokeo.com — сайт для поиска людей, агрегирующий информацию из более чем 60 он-лайн и офф-лайн источников, таких как: телефонные справочники, социальные сети, фотоальбомы, маркетинговые исследования, списки рассылки, государственные переписи, бизнес-сайты и другие. Основные базы для поиска на английском языке и, как следствие, позволяет довольно точно отследить людей, пользующихся иностранными сайтами в повседневной жизни. Сервис является прекрасным примером того, насколько эффективным может быть автоматизированное использование различных источников данных.

1.2.4 Сервисы анализа сообществ и трендов в социальных сетях

В интернете содержится огромное количество книг, инструкций и примеров психологических анализов страницы из социальной сети, но сервисы для автоматизации этого процесса практически отсутствуют. Это можно объяснить тем, что на такого рода сервисы сложно монетизировать. Естественно, что у самих владельцев есть подобные и даже куда более мощные средства. Так, например, система Матрикснет от Яндекс умеет классифицировать следующим образом пользователей (вставить инфу о матркснете с хабра + яка) .

Данный класс приложений похож на мой проект тем, что с помощью автоматических алгоритмов он анализирует состояние и изменения в сообществах и социумах, в то время как я анализирую отдельного человека. Некоторые из этих приложений уникальны и весьма интересны, и на основании этого включены в анализ. Интересно что многие сервисы для анализа twitter'а являются некоммерческими и вследствии этого быстро теряли поддержку.Так, например, в 2011 году эти сервисы еще существовали,были популярны и хорошо работали:

- **http://topsy.com**

  topsy.com - realtime поисковая система, специализирующаяся на поиске и аналитике по социальным медиа, таким как блоги, twitter, google+ и другим социальным сетям. Компания является сертифицированным партнером twitter и поддерживает индекс всех сообщений, начиная с момента создания twitter в 2006 году. Запуску предшествовали три года разработки. С 2012 года партнер Яндекс (используется в формировании новостной ленты), в 2013 куплена Apple за $200 мл. Ключевые характеристики:

* анализ миллиардов разговоров в реальном времени;
* мгновенное получение новостей и информации об изменении в цитируемости;
* поиск наиболее влиятельных пользователей Twitter по любой тематике;
* просмотр продвижения любого хештега в Twitter. Возможность отследить искуственное раскручивание;
* интерактивный анализ по ключевым словам и авторам, каталогизация по темам, влиянию, эмоциональной окраске, языку и географии. Пользователь может узнать наиболее релевантные твиты, ссылки, фотографии и видео для любого поискового терма из индекса Topsy в сотни миллиардов твитов. Пользователи могут группировать термы, настраивать индивидуальные оповещения и ежедневные дайджесты по группам поисковых запросов

Подводя итог, можно сказать что topsy - является одним из лидеров на рынке извлечения данных из социальных сетей. Однако, в силу того, что рынок чрезвычайно разнообразен и имеет множество особенностей в разных странах мира - topsy не является единственным представителем этого класса сервисов.

– **http://www.kribrum.ru/**

- система мониторинга и анализа социальных медиа для управления репутацией в Интернете. Позволяет отслеживать и анализировать упоминания бренда, продуктов, услуг и ключевых персон компании. Система в автоматическом режиме находит отзывы, обрабатывает их, определяет эмоциональную окраску высказываний и выгружает информацию в виде наглядных графиков и интерактивных отчетов. Интересно, что это одна из немногих отечественных

разработок на этом рынке. Продукт принадлежит компании "Ашманов и партнеры"[15]

* широкий охват поиска. Около 700 000 отслеживаемых площадок. Постоянно добавляются новые источники, в т.ч. по запросу пользователя;
* фильтрация спама, точность выборки. Система учитывает только те отзывы, которые относятся к объекту мониторинга, отсеивает спам и сообщения, в которых бренд упомянут вскользь;
* автоматическое определение тональности и тематики сообщений;
* собственная лингвистическая технология, которая позволяет системе «понимать» правила построения предложений, анализировать связи между словами и автоматически определять тональность высказывания (хорошо, плохо, нейтрально) относительно объекта мониторинга с точностью более 80
* оперативность обновления данных. Данные попадают в систему в период от 15 минут до 2-4 часов после публикации;
* система позволяет определить общий охват, а также «вес» каждого упоминания и его автора. Это особенно важно для формирования эффективной информационной политики, выбора подходящих площадок взаимодействия с аудиторией и выявления лидеров мнений. Предусмотрена возможность реагирования;
* разнообразие отчетов, экспорт данных;
* автоматическая генерация отчетов по шаблону и рассылка по электронной почте согласно заданной схеме;
* возможность заказать аналитический отчет у экспертов в области мониторинга социальных медиа;

[15]Крибрум | Ашманов и партнеры `http://www.ashmanov.com/services/kribrum`

* ролевой доступ, система назначения заданий, журналирование действий операторов в системе.[16]

– **TweetStats**

TweetStats - создает инфографику на основе постов человека в twitter по следующим направлениям:

* количество твиттов в час;
* количество твиттов в месяц;
* количество твиттов в зависимости от времени (день, ночь, день недели). Есть функция сохранения результатов анализа.[17]. Проект особенно не развивается, масштаб проекта небольшой, сервис просто хорошо справляется с заявленной функциональностью. Tweets per month Tweet timeline Reply statistics

показывает количество сообщений по месяцам, частоту сообщений в зависимости от времени дня и дня недели. Проект некоммерческий, не развивается, некоторые функции работают нестабильно;

– **Twitteranalyzer**

Twitteranalyzer - статистика по направлениям: Пользователи, Друзья, Упоминания, Группы, - и более мелким подуровням, что позволяет получить довольно много информации для анализа; Также перестал работать;

– **sleepingtime.org**

- простой сервис с одной единственной функцией - определение времени сна по твиттам. Принцип работы достаточно прост: сервис анализирует последние 1000 твитов

[16] Что такое Крибрум `http://www.kribrum.ru/about/`

[17] TweetStats - Graph your Twitter Stats `http://www.tweetstats.com/`

и по ним строит приближенное расписание сна человека. Сервис обладает красивым интерфейсом, набором людей и областей из которых можно проанализировать людей, например шоу-бизнес, it-специалисты, политики, спортсмены;

– **klout.com**

- веб-сайт и мобильное приложение, которое использует аналитику по социальным медиа для выставления ранга от 0 до 100 под названием "Klout Score"по направлениям:

  * True Reach - количество пользователей, на которых вы оказываете влияние;
  * Amplification - охват зоны влияния. Когда вы публикуете что-то, как много людей отвечает на ваш пост или перепечатывает его. Чем больше люди реагируют на ваши посты и сообщения, тем выше зона влияния;
  * Network Score - как ваша аудитория реагирует на ваше влияние. Как часто пользователи (друзья, подписчики или их друзья) делятся вашим контентом со своими читателями и как далеко он расходится по сети? Чем больше вас упоминают, тем выше этот показатель.[18]

  По заявлению разработчиков ранг является корреляцией между вкладом человека в контент социальных сетей и

[18]Работаем с сервисом Klout - а как вы влияете на вашу аудиторию в социальных сетях? https://www.facebook.com/notes/mike-ponomarenko/\T2A\cyrr\T2A\cyra\T2A\cyrb\T2A\cyro\T2A\cyrt\T2A\cyra\T2A\cyre\T2A\cyrm-\T2A\cyrs-\T2A\cyrs\T2A\cyre\T2A\cyrr\T2A\cyrv\T2A\cyri\T2A\cyrs\T2A\cyro\T2A\cyrm-klout-\T2A\cyra-\T2A\cyrk\T2A\cyra\T2A\cyrk-\T2A\cyrv\T2A\cyrery-\T2A\cyrv\T2A\cyrl\T2A\cyri\T2A\cyrya\T2A\cyre\T2A\cyrt\T2A\cyre-\T2A\cyrn\T2A\cyra-\T2A\cyrv\T2A\cyra\T2A\cyrsh\T2A\cyru-\T2A\cyra\T2A\cyru\T2A\cyrd\T2A\cyri\T2A\cyrt\T2A\cyro\T2A\cyrr\T2A\cyri\T2A\cyryu-\T2A\cyrv-\T2A\cyrs\T2A\cyro\T2A\cyrc\T2A\cyri\T2A\cyra\T2A\cyrl\T2A\cyrsftsn\T2A\cyrn\T2A\cyrery\T2A\cyrh-\T2A\cyrs\T2A\cyre\T2A\cyrt\T2A\cyrya\T2A\cyrh/233439080026653

тем, насколько контент, создаваемый пользователем, востребован другими пользователями социальных сетей. Аналитика производится на основе данных сайтов Twitter, Facebook, Google+, LinkedIn, Foursquare, YouTube, Instagram, Tumblr, Blogger, WordPress, Last.fm и Flickr.[19] [20] Klout оценивает степень влияния, используя такие показатели как:

* сколько авторов отслеживает пользователь;
* сколько авторов отслеживают пользователя;
* количество ретвитов;
* упоминания в списках авторов;
* за сколькими спам/мертвыми авторами следит пользователь;
* какова степень влияния тех, кого ретвитит пользователь;
* количество приватных сообщений.

Полученная информация объединяется с информацией из Facebook, комментариями, отметками о понравившейся публикации, количеством друзей. Все эти данные отображаются в «Klout Score», который показывает степень влияния пользователя в социальных сетях. Данный сервис подвергается постоянной критике [21] [22] [23] из-за того, во что он фактически властвует над человеческими судьбами. Так, в 2012 году в США одного вместо квалифицированного специалиста с низким рейтингом klout взяли неопытного

[19] How can you measure influence? (http://www.simplyzesty.com/Blog/Article/July-2010/How-can-you-measure-influence)

[20] http://klout.com/corp/about http://klout.com/corp/about

[21] Why Klout scores are possibly evil (http://money.cnn.com/2011/11/15/technology/klout_scores/index.htm)

[22] Don't Fall for this Sneaky Klout Trick Designed to Suck You In (http://www.forbes.com/sites/anthonykosner/2012/05/08/klout-uses-this-trick-to-make-you-feel-bad-about-yourself-dont-let-it-ruin-your-life/)

[23] Klout overhauls its business model, but does it answer its critics? http://www.businessesgrow.com/2012/08/14/klout-overhauls-its-business-model-but-does-it-answer-its-critics/

парня с высоким рейтингом.[24]. Джон Скалзи (John Scalzi) из CNN описал принцип, лежащий в основе Klout как «социально зло". В результате использования klout вызывает тревожное состояние у своих пользователей.[25] Британский писатель Чарльз Стросс охарактеризовал klout как "герпес для интернета". Анализ условий использования и лицензионного соглашения klout показывает, что бизнес-модель компании является незаконной в Великобритании, где она противоречит закону Data Protection Act 1998 года; Стросс советует читателям удалить их аккаунты Klout и отказаться от услуг этой компании.[26]

1.2.5 Приложения для платформы vk.com

Отдельно стоит упомянуть приложения, написанные на платформе ВКонтакте — все они реализуются по средствам flash, javascript или как iframe со стороннего сайта. Особенностью приложений под эту платорму является их относительная простота. Как правило, они выполняются на стороне клиента или имеют несложную серверную по сравнению с приложениями для анализа сообществ и трендов. В ходе анализа существующих решений были выявлены следующие приложения:

- Анализатор (`https://vk.com/ianaliz`) — способен проанализировать количество друзей, сколько из них женского пола, сколько мужского, сколько не сообщили такую информацию, примерная дата регистрации в vk.com;
- Радар (`https://vk.com/vkradar`) — сообщают статистику по сообщениям на стене, в группе, по полярности сообще-

[24] см. подробней wired: What Your Klout Score Really Means (`http://www.wired.com/2012/04/ff_klout/`)

[25] Klout Now Measures Your Influence on Facebook `http://mashable.com/2010/10/14/facebook-klout`

[26] Charlie Stross - Evil social networks (`http://www.antipope.org/charlie/blog-static/2011/11/evil-social-networks.html`)

ний и т.д. Относительно несложное приложение с моделью монетизации за дополнительный функционал.

- Модерация пабликов и страниц, анализ популярности; (`http://vk.com/public_tools`) - инструмент для автоматизации повседневных задач модератора. Отличается низкой надёжностью (стало временно недоступно во время написания работы);
- Анализ Аватара (`https://vk.com/avascan`) — выдает результаты близкие к случайным, но на основе аватара пользователя по таким характеристикам как: сексуальность, красота и прочее. Не так давно было заблокировано администрацией платформы.
- Лайк-машина (`http://vk.com/like.machine`) - приложение, позволяющее за дополнительную плату повысить свою популярность и отследить посетителей. Аналогичных приложений для платформы vk, решающих подобные задачи выявлено не было.

По итогам анализа были сделаны следующие выводы:

- представляется интересным показывать результаты работы приложения посредством публикации отчета на стене пользователя;
- возможна монетизация приложения путем бесплатного предоставления пользователю некоторого количество виртуальных денег, которых хватит на то, чтобы попробовать лишь часть функционала, в то время как весь функционал будет стоить дополнительных денег;
- в приложении можно реализовать дополнительный(не основной) функционал, но он должен пониматься пользователем как дополнительный и предоставляться бесплатно.

1.2.6 Выводы

По итогам анализа этих проектов были сделаны следующие выводы:

- большие объемы данных позволяют построить более детальную аналитику, чем локальный анализ;
- часто один и то же пользователь зарегистрирован сразу в нескольких социальных сетях. И для отслеживания каких либо общественных изменений, как правило, достаточно глубоко анализа одной из платформ. Следует отметить, что большинство сервисов заточены на Twitter, т.к. он является наиболее удобной и открытой социальной сетью.
- достаточно интересным оказался функционал сайта sleepingtime.org, который анализирует время публикации постов. В дальнейшем возможно развить отсюда следующие направления:
  * вычислить время, в которое пользователь активно пишет посты в социальной сети;
  * вычислить время сна пользователя;
  * вычислить примерное количество времени, которое пользователь проводит в социальной сети.

  я точно не реализую эту функциональность, стоит ли тогда писать об этом? может как то изменить предложение?)
- показательным является пример klout, который зарабатывает деньги, меняя наше общество к худшему.

Среди всех приложений я пытался выявить те, которые анализируют метаданные в целях получения дополнительных сведений о пользователе. Таких приложений оказалось немного: sleepingtime, Радар, Анализатор для vkontakte. Приложения Радар и Анализатор не реализуют функционала, который заложен в дипломный проект. Малое количество приложений можно объяснить тем, что вся мощь, заключенная в метаданных,

раскрывается при больших объемах информации. Создание приложений, анализирующих большие объемы, стоит больших денег. Поэтому должна существовать ясная модель монетизации. В гражданской сфере, в основном, востребован мониторинг брендов. Особняком стоит относительно новый сервис klout. Совсем другая ситуация складывается в военном секторе, особенно в области разведки, где созданы огромные системы, такие как Xkeyscore и другие, в которых основной источником анализа являются метаданные.

Можно сказать, что приложений, реализующих заявленный в дипломной работе функционал, выявлено не было.

1.3 Обоснование выбора технических инструментов для реализации приложения

Инструменты которые будут использоваться для построения приложения:

- Python выбран в связи с большой функциональной выразительностью [27] и гибкостью языка[28]. Проект не рассчитан на очень высокие нагрузки, поэтому с одной стороны производительности интерпретатора Python вполне достаточно, а с другой - упрощается процесс написания и сопровождения приложения.

- MySQL выбрана как одна из самых быстрых СУБД при средних и маленьких объемах БД. Также у проекта хорошая документация.

[27] С. Макконнелл — Совершенный код, с.60, Microsoft Press,М.:2012

[28] `http://ru.wikipedia.org/wiki/Python`

– Django выбран за высокую скорость написания приложения и архитектурные преимущества, по сравнению с такими фреймворками как Symfony(PHP) и Dancer (Perl), а также в связи с тем, что написание приложения обработки данных и приложения веб-клиента на одном языке упрощает сопровождение (Django написан на Python).

– Ubuntu server 12.04 выбрана среди прочих аналогов, таких как Fedora, Debian, OpenSuse по следующим причинам:

  * дружелюбное сообщество;
  * безопасность;[29]
  * легкая расширяемость с помощью ppa[30] и центра приложений;
  * активная поддержка. (Ubuntu server 12.04 будет поддерживаться до апреля 2017 года)[31]
    В процессе написания приложения необходимо будет решить следующие проблемы:
  * распознавание информации с сайта;
  * обход страниц друзей пользователя и распознавании информации на их странице;
  * минимизация нагрузки с одного клиента;
  * авторизация пользователя средствами социальной сети (по протоколу OAuth 2.0).

Приложение реализовано по клиент-серверной архитектуре. Сервер обработки данных на Python, база данных на MySQL, клиентская часть (веб-форма) на фреймворке Django. Так как планируется использовать отдельный сервис для обработки дан-

[29] см. подробнее доклад National Technical Authority for Information Assurance (CESG) - End User Devices Security Guidance: Introduction `https://www.gov.uk/government/collections/end-user-devices-security-guidance--2#overview`

[30] `http://help.ubuntu.ru/wiki/ppa`

[31] `http://ru.wikipedia.org/wiki/Ubuntu`

ных, то размещение проекта на виртуальном (shared) хостинге является недостаточным. Проект будет размещен на виртуальном выделенном сервере (VPS), ОС для сервера — Ubuntu 12.04 Все инструменты, которые я применяю, являются продуктами с открытым исходным кодом. Они активно развиваются, поддерживаются и имеют живое сообщество пользователей в т.ч. русскоязычное.

безопасность1 легкую расширяемость с помощью ppa2 и центра приложений активная поддержка (Ubuntu server 12.04 будет поддерживаться до апреля 2017 года)3