0.1 Введение

В настоящий момент довольно остро стоит вопрос о сохранении тайны связи при использовании электронной почты, чата, социальных сетей и иных электронных средств коммуникаций. В настоящий момент закон о сохранении тайны связи не охватывает публичные сервисы.[1] Кроме того, опубликованные Эдвардом Сноуденом данные наглядно демонстрируют, что межправительственные системы слежения (созданные для борьбы с терроризмом) используются для достижения экономических и политических целей, и нарушают права граждан на тайну частной жизни и тайну переписки1.

Целью проекта является создание веб-приложения, демонстрирующего различные возможности по сбору сведений об отдельном человеке с использованием только открытых источников.[2] Особенно интересным представляется создать веб-сервис для автоматизированного анализа страницы в социальной сети с выделением дополнительных сведений о человеке, на основе сведений о его друзьях. Веб-сервис планируется создать в демонстрационный целях т.к. решение о публикации своих данных осуществляется непосредственно человеком.

Люди часто недооценивают значение метаданных и комплексного анализа. Под комплексным анализом далее будет подразумеваться сочетание методов, подходов, инструментов по интеллектуальной добыче данных (Data Mining), использованию больших объемов данных (Big Data), Люди не задумываются о том, что вступая в определенные электронные сообщества их личные данные может сообщить не только владелец аккаунта, но и другие участники сообщества. Причем, чем ближе они знакомы, тем больше данных они могут передать, иногда даже сами того не подозревая. Так же большой интерес вызывает возможность автоматического извлечение фактов из текста на естественном языке. Не так давно исследователь сотрудники IBM Research во главе с Джалал Махмудом (Jalal Mahmud) опублиоквали научную работу в которой демонстрируют возможность опередить местонахождение человека по его постам в Twitter с точностью до 70% (определяется обычно город или округ).[4] Основная идея алгоритма придуманного сотрудиками IBM заключается в том, что само содержание твитов несет в себе информацию о местонахождение и 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 анализируются симантика текста для извлечение фактов, например из текста "Сергей не забудь самовар, встрчаемся 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 [5]

[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]

[4]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]`)

[5]подробнее см. Twitter Rest API (`https://dev.twitter.com/docs/api/1.1`) и публика-

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 источников.

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

0.2 Основной функционал приложения

Обязательный функционал позволит определить пол, возраст, ВУЗ некоторого человека в социальный сети Вконтакте, на основе данных получаемых в автоматическом режиме. Состав дополнительного функционала, сообщающий значимую дополнительную информацию о человеке, будет определен в процессе разработки, т.к. на начальном этапе не представляется возможным определить его из-за большого размера проекта социальной сети Вконтакте.

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

0.2.1 Цели и задачи дипломного проекта

Задачи:

- анализ лигитимности функционала приложения;
- анализ существующих web-сервисов, которые предоставляют дополнительную информацию о пользователе с помощью анализа косвенных признаков;
- анализ существующих научных подходов для реализация данной задачи;
- составление описания для каждого решения;
- анализ законности существования приложений данного типа;

цию Jalal Mahmud, Jilin Chen, Michelle Zhou, Jeffrey Nichols Who will...ad Information

- анализ существующих научных подходов для реализация данной задачи;
- реализация обязательного функционала. Уточнение и реализация дополнительного функционала;
- тестирование и доработка приложения.

0.3 Анализ существующих решений

Вследствие огромной популярности социальных сетей, в интернете уже давно стали появляться проекты, дополняющие их функционал.

0.3.1 smm-продукты

Такие проекты автоматизируют задачи с использованием инструментария, предоставляемого непосредственно социальными сетями, например, публикация постов в определенное время, статистика популярности сообщений. Так же продукты этого класса могут автоматизировать любые другие действие упрощающие социальный медиа маркетинг (smm)[6]

0.3.2 социальные агригаторы

Так же существуют проекты, программные продукты или сервисы, которые собирают информацию из разных социальных сетей, блогов и других ресурсов в один источник.[7]

Вышеперечисленные классы программ и сервисов являются самыми распространенными в силу того что их возможно монетизировать и данная вид служб востребован пользователями. Стоит отметить что не все сервисы четко вписываются в тот или иной класс приложений, потому как многие из них достаточно самобытны и быстро изменяются и даже зачастую перестают существовать. Так за время подготовки дипломной работы перестали функционировать 4 сервиса. Для данного иследования представляется наиболее важным выделить существующие методы получен 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 не доступную обычным поисковым системам, получить релевантную информацию которая обычно слишком низко ранжируется.

0.3.3 Отдельные 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 мо-

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

[7] 20 Ways To Aggregate Your Social Networking Profiles (`http://mashable.com/2007/07/17/social-network-aggregators/`)

мент количество социальных сетей уже исчисляется десятками и это только те, которые имеют значительное (более нескольких миллионов) и живое сообщество.[8] Существует большое количество CMS,конструкторов сайтов позволяющие достаточно быстро создать свою собственную социальную сеть или отдельный блог с интеграцией с другими блогами построенными на той же технологии.[9] Все сети имеют свои особенности, поэтому агрегация этого многообразия - задача не простая, и ее можно решить несколькими способами. К основным проблемам, которые необходимо решить таким приложениям являются:

- написание адаптеров для каждого источника информации[10]
- решение вопросов разряженности данных (социальные сети обладают различным функционалом и данными о своих пользователях)
- скорость работы - агригатор собирает информацию с других сервисов и значит в вподает в зависимость от скорости работы 3-их лиц, что не всегда может быть надежно

основными представителями являются:

• 0.3.4 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:

- 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).

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

[9] 8 Great Social Networking CMS (`http://www.cmscritic.com/8-great-social-networking-cms`)

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

- ### 0.3.5 http://topsy.com

topsy.com - realtime поисковая система, специализирующаяся на поиске и аналитике по социальным медиа, таким как блоги, twitter, google+ и другие социальные сети. Компания является сертифицированным партнером twitter и поддерживает индекс всех сообщений начиная с момента создания twitter в 2006 году. Запуску предшествовали три года разработки. Основные черты:

  - 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
  - множество других инструментов для анализа трендов в социальных сетях

Он делает некоторое количество синтетических пометки , чтобы извлечь тему из твита сделать тему для поиска , а также выполняет классификацию контента, где есть больше текста , чтобы играть с , для ссылок , упомянутых в чириканье. Он понимает, что автор отличается от того, что обсуждается и который имеет в виду которых в почтовых отправлений , которые питаются в его графике влияния , который занимает ссылки в результатах поиска на основе влияния людей, говорящих о этих связей . Это включает в себя глобальную звание пользователем , независимо от темы и сроки , а также ряды на уровне ключевых слов на основе того, что было в чириканье , когда они получили внимание на это С 2012 года партнер Яндекс, в 2013 куплена Apple за $200 мл;

- ### 0.3.6 http://qwant.com

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

- ### 0.3.7 http://spokeo.com

spokeo.com — сайт для поиска людей, аггрегирующий информацию из множества других он-лайн и офф-лайн источников, таких как: телефонные справочники, социальные сети, фотоальбомы, маркетинговые исследования, списки рассылки, государственные переписи, безнесс-сайты, всего — более чем из 60 источников. Основные базы для поиска на английском языке и, как следствие, позволяет довольно точно отследить людей, пользующихся иностранными сайтами в повседневной жизни.

0.3.8 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 манетизировать. Естественно, что у самихвладьцев есть подобные и даже куда мощные средства. Так например система матрикснет от Яндекс умеет классифицировать следующим образом пользователей.

Данный класс приложений похож на мое приложение 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 много сервисов для анализа twitter'а являются некоммерческими и вследствии этого быстро теряли поддержку, так например в 2011 году эти сервисы еще существовали или были популяярны и хорошо работали:

- **TweetStats**

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

- **Twinfluence**

  ыл простым инструментом для измерения совокупного влияния твиплов и их фолловеров, а также в качестве бонуса предоставляет статистику некоторых социальных сетей. В данный момент недоступен, по доменну на котором находился проект стоит переадрисацию на компанию в которой работают бывшие владельцы Twinfluence;

- **TweetEffect**

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

- **Twitteranalyzer**

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

  TwitterCounter - позволяет отслеживать статистику популярности вашего аккаунта, настроить уведомления себе на почту;

Twittergrader - детальная статистика аккаунта с показами всех заходов и так далее;
Klout - статистика аккаунта, динамика роста, его твит-сила; Tweetmetrics - разнообразная статистика аккаунта; Trendrr - мощный сервис для сбора статистики в интернете, включая Твиттер; Trendistic - дает представление о популярности того или иного запроса, представляет его количество повторений в Twitter на графике; TweetBeep - позволяет отслеживать упоминание бренда, имени, сайта, вообще любого слова в Твиттере, присылая уведомления; Analytics.ad.ly - систематизирует информацию о фолловерах; TwitterStreamGraphs - интерактивный инструмент, позволяющий создать график на основе слов упоминающихся в сообщениях пользователей Твиттере; Tweetoclock - поможет отследить время использования пользователями своего твиттер-аккаунта; Radian - мощное средство для анализа социальных сервисов, включающее в себя и инструменты для анализа Твиттера; sleepingtime.org — сайт анализирует время Вашего сна по твиттам. http://www.tweetping.net/