

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.²³ 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 демонстрационных 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) опубликовали 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 сотрудниками 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, 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/firehose⁵

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/Сноуден,_Эдвард)

⁴ Who will retweet this?: Automatically Identifying and Engaging Strangers on Twitter to Spread Information (digital.cs.usu.edu/kyumin/pubs/lee14iui.pdf)

⁵ подробнее см. Twitter Rest API (<https://dev.twitter.com/docs/api/1.1>) и публикацию

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 социальной сети Вконтакте, на основе данных получаемых в автоматическом режиме. Состав дополнительного функционала, сообщаящий 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 легитимности функционала приложения;
- анализ существующих web-сервисов, которые предоставляют дополнительную информацию о пользователе с помощью анализа косвенных признаков;
- анализ существующих научных подходов для реализации данной задачи;
- составление описания для каждого решения;
- анализ законности существования приложений данного типа;

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

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

0.3 Анализ существующих решений

Вследствии огромной популярности социальных сетей, в интернете уже давно стали появляться проекты, дополняющие их функционал. Такие проекты решают иные задачи с использованием инструментария, предоставляемого социальными сетями. Помимо этого, существуют проекты, обобщающие функционал различных социальных сетей в одном месте. Такие проекты получили название агрегаторов.

Был проведен анализ существующих решений, 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.

0.3.1 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 момент количество социальных сетей уже исчисляется десятками и это только те, которые имеют крупное и живое сообщество. Все сети имеют свои особенности, поэтому агрегация этого многообразия - задача не простая, и ее можно решить несколькими способами.

• 0.3.2 <http://people.yandex.ru>

people.yandex.ru – простой сервис (в одну строку) поиска в различных социальных сетях с преобладающим числом русскоязычных пользователей. Также сервис обладает функционалом кластеризации результатов. Данный сервис не содержит функционала, похожего на мое приложение, но является неплохим примером построения интерфейса взаимодействия с пользователем;

• 0.3.3 <http://topsy.com>

topsy.com - realtime поисковая система, специализирующаяся на поиске по социальным медиа, таким как блоги, твиттер, сообщения в социальных сетях;

- **0.3.4** <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);

- **0.3.5** <http://spokeo.com>

spokeo.com — сайт для поиска людей, агрегирующий информацию из множества других он-лайн и офф-лайн источников, таких как: телефонные справочники, социальные сети, фотоальбомы, маркетинговые исследования, списки рассылки, государственные переписи, бизнес-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.

0.3.6 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 самих владельцев 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'a являются некоммерческими и вследствие этого быстро теряли поддержку, так например в 2011 году эти сервисы еще существовали или были популярны 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 твицлов и их фолловеров, а также в качестве бонуса предоставляет статистику некоторых социальных сетей. В данный момент недоступен, по

домену на котором находился проект стоит переадресацию 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/>