[colorlinks,filecolor=black,citecolor=black,colorlinks=true]hyperref

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 переписки.

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 ин-

[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]

терес 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, 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 [5]

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 происхо-

[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{вГЋ}intopreamble]`)

[5] подробнее см. Twitter Rest API (`https://dev.twitter.com/docs/api/1.1`) и публикацию Jalal Mahmud, Jilin Chen, Michelle Zhou, Jeffrey Nichols Who will...ad Information

дит 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;
- 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 отме-

[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/`)

тить 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 ряд сервисов:

- **twinfluence**

  был простым инструментом для измерения совокупного влияния твитов 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**

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

К самым интересным социальным агрегаторам можно отнести:

- **Hootsuite**

  - Один из самых надежных и доступных инструментов, HootSuite постоянно совершенствует свой интерфейс и возможности. Онлайн-доступ позволяет войти в свой аккаунт с любого места, чтобы контролировать свои аккаунты. В настоящее время, есть поддержка Twitter, Facebook Pages, Facebook, LinkedIn, Ping.fm, Wordpress.com, MySpace и Foursquare. HootSuite обладает функционалом, которые позволяют настроить, отправку поста во

множество источников в несколько кликов.[8] Ключевыми характеристиками являются:

- Планирование. Выбор между обновлением постов он-лайн или по зарание загатовленному расписанию.
- 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 использовать фирменный аплет для браузера, что бы быстро поделиться информацией

- **Tweetdeck**

  - Оригинальный и популярный инструмент для twitter, Originally a popular tool for tweeters, Tweetdeck has evolved into a comprehensive platform that services Facebook, LinkedIn and MySpace. Built using Adobe Air, it has a blend of rich-technology and customizable features that end users will enjoy using. The Tweetdeck platform is available for desktop, iPhone and iPad. It also plays well with others so you can use it on your Mac, PC or Linux system. Like other social media aggregators, Tweetdeck has a column-style format that silos your information

Для данного исследование 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

[8] 7 Social Media Aggregation Tools To Simplify Your Streams `http://socialmediatoday.com/SMC/192312`

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 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.[9] 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.[10] 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[11]
- 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:

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

[10] 8 Great Social Networking CMS (`http://www.cmscritic.com/8-great-social-networking-cms`)

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

- **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).

http://qwant.com

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

- **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. Сервис является прекрасным примером того, насколько эффективным может быть автоматизированное использование различных источников данных.

0.3.4 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 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:

- **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. С 2012 года партнер Яндекс (используется в формирование 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 или географии. Пользователь может узнать, наиболее релевантные твиты, ссылки, фотографии и видео для любой терма из индекса Topsy в сотни миллиардов твитов. Пользователи могут групповые термы в сохраненных тем и настройки индивидуальных оповещений и ежедневных дайджестов деятельности.

  Подводя итог, можно сказать что topsy - является одним из лидеров на рынке извлечения данных из социальных сетей, но в силу того что рынок чрезвычайно разнообразен и имеет множество особенностей в разных странах мира, то topsy не является едиственным 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"[12]

    – Широкий охват поиска. Порядка 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, что особенно важно для

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

формирования эффективной информационной политики, выбора подходящих площадок взаимодействия с аудиторией и выявления лидеров мнений. Возможность реагирования

- Разнообразие отчетов, экспорт данных
- Автоматическая генерация отчетов по шаблону и рассылка по электронной почте по заданной схеме.
- Возможность заказать аналитический отчет у экспертов в области мониторинга социальных медиа.
- Ролевой доступ, система назначения заданий, журналирование действий операторов в системе[13]

- **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.[14]. Проект особенно не развивается, масштаб проекта не большой, 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 не стабильно;

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

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

- **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.

- **sleepingtime.org**

  - Простой сервис с одно единственной функцией - определение времени сна по твиттам. Принцеип работы достаточно прост: сервис анализирует последние 1000 твитов и по ним строит приближенное расписание сна человека. Сервис обладает красивым интерфейсом и набором людей и областей из которых можно проанализировать людей, например шоу-бизнес, it-специалисты, политики, спортсмены.

0.3.5 Выводы

По итогам анализа этих проектов были сделаны следующие выводы:

- большие объемы данных позволяют построить более детальную аналитику, чем локальный анализ
- так как в разных социальных сетях сидят одни и и те же людей, то при отслеживании каких либо общественных изменений, как правило, достаточно глубоко анализа одной из платформ, поэтому большинство сервисов заточены на Twitter, как наиболее удобную и открытую социальную сеть из всех.

Несогласованное предложение?

- достаточно интересным оказался функицонал сайта sleepingtime.org, который анализирует время публикации постов. В дальнейшем возможно развить отсюда следующие направления:
  - вычислить время, в которое пользователь активено пишет посты в социальной сети

– вычислить время сна

– примерно вычислить сколько часов в день пользователь проводит в социальной сети

я точно не реализую эту функциональность, стоит ли тогда писать об этом? может как то изменить предложение?)

достаточно интересным оказался функицонал сайта sleepingtime.org, который анализирует время публикации постов. В дальнейшем возможно развить отсюда три направление: а)вычислить время, в которое пользователь активено пишет посты в социальной сети б)вычислить время сна в)примерно вычислить сколько часов в день пользователь проводит в социальной сети интересный функционал klout.com, который реадизует механизм махожий на arhive.org, но для отдельных аккаунтов Twitter‘а. Конечно интересно повторить данный функционал для других платформ, но это представляется маловероятным из-за особенностей социальных сетей и сильного отличия Twitter от других сервисов.