Некоторые аспекты задачи исследования распространения информации в социальной сети ВКонтакте

Евгений Рабчевский, Артем Цукерман

ПГНИУ, Пермь, Россия. evgeny@rabchevsky.name, azholy@gmail.com

Аннотация. Статья посвящена некоторым техническим аспектам обширной задачи по исследованию особенностей распространения информации в социальной сети ВКонтакте. В статье описывается работа по сбору информации со страниц пользователей, индексация и поиск по полученным данным, а также анализ топологии Сети, который говорит о безмасштабности исследованного сегмента.

Ключевые слова: поиск в социальных сетях; анализ и мониторинг социальных медиа; аудит сообществ, исследование общественного мнения в социальных сетях.

Введение

На сегодняшний день социальные сети представляют собой коммунникативную площадку, которая активно используется для формирования и манипулирования общественным мнением. Так, например, сегодня не редкость, когда "ВКонтакте" создаются группы для помощи людям, попавшим в беду, или группы любителей определенных брендов, которые продвигают соответствующую продукцию, и так далее. В данном случае социальную сеть можно рассматривать, как общественную среду, на которую пытаются каким-то образом повлиять для дос-

Выходные данные сборника.

тижения определенной цели, например формирования определенного общественного мнения по определенному вопросу. Для решения этой задачи, лицо, заинтересованное в формировании определенного мнения (далее будем называть его автором), должно знать, каким образом максимально эффективно привести сеть в необходимое состояние. А для этого требуется определить максимально эффективные каналы распространения информации по сети, пользователей, потенциально готовых с большей вероятностью принять необходимую позицию (например, тех, кто разделяет некоторые взгляды автора), и так далее.

Помимо формирования общественного мнения актуальны такие прикладные задачи как оценка успешности любого нововведения, будь то сервис, сообщество или рекламная компания. Такая оценка дает возможность разрабатывать сервисы не наугад, а целенаправленно повышать пользовательскую активность. Также актуален анализ аномального изменения пользовательской активности, поиск целевой аудитории бренда и маркетинговые исследования, связанные с получением данных о позиции пользователей по отношению к услугам какой-либо компании, что требует, однако, более глубокого лингвистического анализа.

Все это требует полноценного анализа сегмента социальной сети, с которым придется работать автору. Более того, этот самый сегмент еще нужно каким-то образом выделить. Также следует принять во внимание тот факт, что такая задача уникальна для каждого автора, и при этом очень востребована в целом для сети.

В этой связи, глубокое понимание особенностей распространения информации в социальных сетях в целом, и ВКонтакте в частности, на практике, например, позволит осуществлять более эффективные маркетинговые мероприятия в социальных сетях.

Постановка задачи

Задача исследования особенностей распространения информации в социальных сетях представляется достаточно сложной. Она включает в себя задачи по

- автоматическому сбору информации из Сети
- анализу собранной информации с учетом изменения по времени
- идентификации источников данных, представляющих информацию по определенной теме (тема, распространение которой исследуется)
- непосредственно анализ полученных на первых этапах данных с целью выявления определенных закономерностей и особенностей.

Данная статья представляет собой отчет о промежуточных результатах решения указанной комплексной задачи, которые на данный момент нельзя считать законченным исследованием, а стоит рассматривать лишь как некую подготовительную работу по решению указанной сложной задачи.

Согласование исследований с правилами пользования сайта ВКонтакте

Для сбора информации из Сети вместо стандартного API ВКонтакте использовалась имитация поведения пользователя при просмотре страницы. Согласно пункту 4.2. правил, пользователем должно быть физическое лицо, а в случае сбора информации при помощи программ, согласно пункту 5.3.9. правил, необходимо получить согласие администрации Сайта.

На данный момент, запрос к администрации Сайта на получение разрешения использования программ для сбора информации из Сети не выполнялся. Это связано с тем, что на данный момент сбор информации проводился единоразово, и не представлял собой существенной нагрузки на Сеть. В последующем, такой запрос будет сделан.

Краулер - программа для сбора данных со страниц Сети

Для сканирования страниц пользователей социальной сети Вконтакте, далее Сети, на интерпретируемом высокоуровневом языке программирования Ruby была разработана программа – краулер.

Особенность работы краулера состоит в том, что информация в Сети представляется с помощью динамических технологий. И для сканирования Сети краулеру требуется полностью имитировать работу реального пользователя, это достигается формированием определенной последовательности запросов типа POST (например, полностью имитирующей вертикальную прокрутку страницы), отправляемой на HTTP сервер Сети. Ответы сервера Сети, которые являются сообщениями со страниц пользователей, сохраняются краулером в базу данных СУБД MySQL.

Еще одной сложностью при разработке краулера стала особенность поведения пользователей, которая заключается в том, что около 2/3 всех пользователей (по нашей статистике) Сети закрывают свои страницы для неавторизированных пользователей. Для обхода этой проблемы краулер авторизуется на сервере vk.com как зарегистрированный пользователь.

Выходные данные сборника.

Однако, при чрезмерной активности пользователь блокируется сервером vk.com, поэтому перед переходом на следующую сканируемую страницу краулер делает паузу в 1 секунду, что заметно снижает скорость работы. Чтобы ускорить работу, программа использует многопоточность, где каждый поток обращается к серверу как отдельный зарегистрированный пользователь, а список страниц распределяется между потоками. Программа может работать в нескольких режимах, с использование авторизации и без, а также из-под прокси-сервера.

Режим без авторизации используется для предварительного сканирования страниц пользователей, которые открыли свои страницы, что позволяет сэкономить время.

В первой версии программы список пользователей был получен парсингом выдачи стандартного поиска vk.com/search (в качестве выборки использовалось местоположение город Пермь). Оказалось, что стандартный поиск vk.com ограничен выдачей в 1000 человек.

Во второй версии это ограничение было преодолено и количество пользователей расширено по требованию задачи. Ограничение было преодолено за счет того, что краулер получал список друзей пользователя, которые проживают в городе Перми, сохранял их в базу данных и заносил в очередь, после чего процедура повторялась с каждым из них, таким образом, удалось выявить необходимое количество пользователей.

При сканировании лишь сообщений со стены пользователей мощности краулера сейчас хватает для того, чтобы охватить один не крупный город. Время обработки линейно зависит от количества параллельных потоков, имитирующих работу пользователя.

Индексация и полнотекстовый поиск

В первой версии краулера, собранная им коллекция документов составила 23318 сообщения, принадлежащих 1000-ти пользователям, что занимает 9.6 MB в базе MySql.

Полученная таким образом коллекция проиндексировалась системой полнотекстового поиска Sphinx [1]. В результате чего к коллекции добавился индекс, который с помощью системы Sphinx позволяет очень быстро осуществлять полнотекстовый поиск по всем сообщениям коллекции. Система Sphinx была выбрана вследствие ее эффективности при работе с небольшими сообщениями, каковыми являются сообщения со стены пользователей.

Время индексации составило 11.966 секунды, а скорость соответственно 804320 байт/сек или 1948.54 документов/сек.

Во второй версии коллекция была расширена до 1152586 сообщений 63770 пользователей, и занимает 493.6 MB. Время индексации составило 2163.567 секунд, а скорость 228142 байт/сек или 532.72 документов/сек.

Доступ к системе поиска по данным сообщениям может быть осуществлен по адресу http://78.47.43.6/ [2]. Представленная система поиска целенаправленно ограниченна по количеству результатов поиска и функционалу работы с ними.

Обработка больших объемов данных

При дальнейшей работе была поставлена задача получения данных обо всех пользователях пермского сегмента Сети. При этом требовалось получить не только сообщения со стены пользователя, а полную информацию с его страницы, включая "лайки", комментарии, интересы и так далее.

Эта задача потребовала бОльших ресурсов и более сложной архитектуры. Поэтому требует и более интеллектуального подхода к сканированию информации со страниц пользователей, нежели одномоментного индексирования. В связи с этим, для решения задачи анализа динамического состояния социальной сети разработана четырехуровневая модель сканирования.

На первом, низшем, уровне модель включает модель данных социальной сети ВКонтакте и библиотеку для извлечения данных со страниц пользователей. На втором уровне модель включает средства, позволяющие на базе данных первого уровня получить анализ характеристик пользователей и материалов с учетом их динамики. Третий уровень включает средства, позволяющие на базе данных второго уровня получить данные о пользователях, сообществах и материалах, обобщенные с точки зрения масштабов всей сети. И, наконец, четвертый уровень включает модели и соответствующие алгоритмы, определяющие пути и способы сканирования сети.

Указанная модель сканирования полностью реализована лишь на первом уровне, остальные уровни реализованы частично. В частности, разработана модель данных социальной сети ВКонтакте. Она предоставляет возможности для последующего анализа индексированных данных в различных разрезах. Данная модель оперирует простейшими фактами, такими как: отправка сообщения, лайк, репост и т.д. Факты складываются в триплеты: «пользователь А» написал «текст сообщения»; «текст сообщения» находится на «странице пользователя А». Таким образом, можно описать все действия пользователей. Удобство такого

Выходные данные сборника.

представления заключается в возможности получать данные по сложным запросам, например: получить все сообщения, которые «А» написал на стене «Б» и у которых имеются лайки. Модель связывает понятия, с которыми система оперирует при навигации по персональной странице одного пользователя (например, ссылки «Мне нравится», «Друзья», и т.д.), с характеристиками этих элементов навигации, которые определяют, какую часть страницы пользователя и каким образом, следует обрабатывать, для того чтобы извлечь из его страницы необходимую информацию.

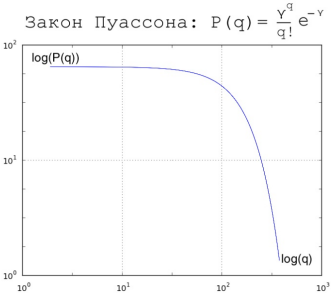
Вопрос о физическом хранении RDF графа, хранящего в себе такого рода информацию, которая указаны выше, на данный момент остается открытым.

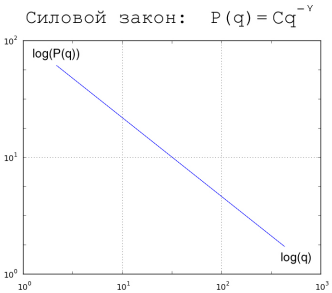
Создана библиотека, которая в совокупности с моделью данных социальной сети, позволяет в режиме реального времени получать полную информацию со страницы одного пользователя.

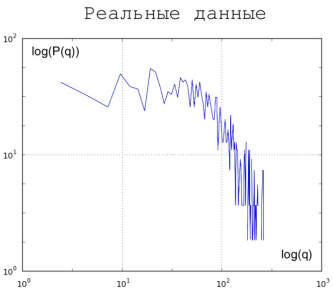
При обработке таких больших объемов данных будут достигнуты ограничения СУБД MySql по производительности. Поэтому было приятно решение перейти на распределенную базу данных в системе PostgreSQL.

Исследование общественного мнения в сети

Используя возможности программы – краулера для фрагмента Сети, состоящего из 1000 пользователей, были проведены исследования, в ходе которых получены данные о структуре фрагмента Сети. А также определена его топология. Исследованный фрагмент Сети был представлен в виде графа. На графике ниже с меткой «реальные данные» представлено распределение степеней P(k), где k является числом связей, выходящих из данного узла графа (пользователя), а P(k) - это вероятность того, что степень (число связей) случайно выбранного узла равняется k. Видно, что график распределения для исследуемого фрагмента Сети близок к распределению Пуассона. Если увеличить объем исследуемого сегмента, то можно предположить, что в предельном случае распределение совпадет с Пуассоновским. Таким образом, Сеть, как и ожидалось, является безмасштабной [4].







На полученном фрагменте Сети Проведено имитационное моделирование, в ходе которого показано, что мнения агентов стабилизируются [5].

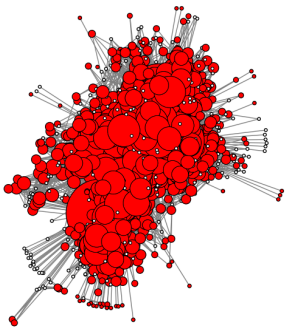
На начальном этапе моделирования мнения пользователей распределялись случайным образом. Каждому пользователю присваивалось значение в диапазоне [-50..50]. На последующих шагах пользователи обменивались мнениями, в результате чего мнения стабилизировались и приняли значение равное 50.

Во втором варианте моделирования мнения пользователей устанавливались также как и в первом, но у выбранной подгруппы мнение было целенаправленно установлено в значение (-50). Ниже представлен конечный результат моделирования двух сценариев – со случайным первоначальным распределением мнения, и с целенаправленно смещенным распределением. Результат представлен на изображении цветом.

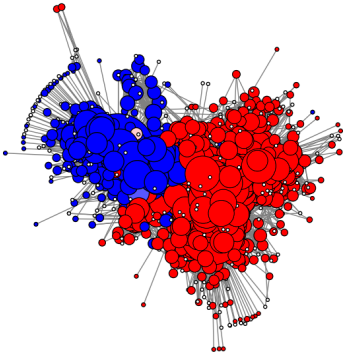
На графиках представлены зависимости значения среднего мнения от шага.

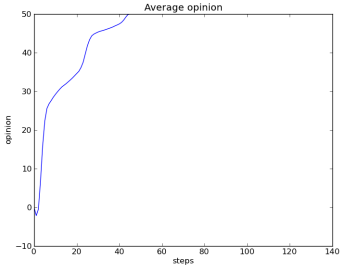
Выходные данные сборника.

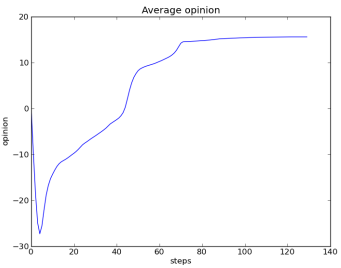
случайное распределение



смещение мнения подгруппы







Данное имитационное моделирование показывает, что, манипулируя начальными мнениями группы агентов, можно эффективно влиять на итоговое среднее мнение всех членов сети.

В ходе экспериментов выявлено, что сеть устойчива к случайным воздействиям, а целенаправленные воздействия, могут вызвать лавинообразный процесс распространения информации [6]. Данные результаты соответствуют теории безмасштабных сетей.

Для проведения реальных экспериментов по анализу изменения общественного мнения в Сети по определенному вопросу, потребуется формальный механизм преобразования вербальных оценок пользователей Сети определенных тезисов в точное количественное представление. Что планируется реализовывать в контексте определенной предметной области.

Используя систему поиска, расширенную критериями поиска по времени и граф сети, можно будет производить оценку скорости рас-

пространения информации и ее охват, выявлять характерные пути распространения информации, и т.д.

Выводы

1. Исследование топологии фрагмента Сети показало, что Сеть можно считать безмасштабной.

2. Для исследования распространения информации в Сети для больших объемов данных требуется согласовать свои действия с администрацией сайта ВКонтакте.

3. При расширении объема хранимой информации требуется переход с MySql на систему, обладающую более высоким быстродействием и дающую широкие возможности для хранения, доступа и архивирования информации в распределенном виде, например PostgreSQL.

4. Для исследования распространения реальной информации необходимо выбрать предметную область и сформулировать методику по количественной формализации вербальных оценок пользователей.

Список источников

1. Платформа полнотекстового поиска http://sphinxsearch.com/

2. Демонстрация поиска по пермскому сегменту социальной сети Вконтакте http://78.47.43.6/

3. Бизнес-анализ в социальной сети Одноклассники http://habrahabr.ru/company/odnoklassniki/blog/149391/

4. Barabasi A.L. Scale Free Networks : Scientific American - http://www.barabasilab.com/pubs/CCNR-ALB_Publications/200305-01_SciAmer-ScaleFree/200305-01_SciAmer-ScaleFree.pdf

5. Губанов Д.А., Новиков Д.А., Чхартишвили А.Г. Социальные сети модели информационного влияния, управление и противоборство: Под ред. чл.-корр. РАН Д.А. Новикова. – М.: Издательство физико-математической литературы, 2010. – 288 с. ISBN 9785-94052-194-5

6. Barabasi A.L. Social consensus through the influence of committed minorities - http://arxiv.org/PS_cache/arxiv/pdf/1102/1102.3931v2.pdf

Выходные данные сборника.