ВЕСТНИК ПЕРМСКОГО УНИВЕРСИТЕТА

2012 Информационные системы и технологии Вып. 13 (40)

УДК 004 Информационные технологии. Компьютерные технологии. Теория вычислительных машин и систем

Выявление феноменов этнической агрессии в пермском сегменте социальной сети vk.com

Е. А. Рабчевский

Пермский государственный национальный исследовательский университет, 614990, Пермь, ул. Букирева, 15. evgeny@rabchevsky.name

Рассматривается задача формирования корпуса сообщений пермского сегмента социальной сети vk.com, содержащих феномены этнической агрессии. Описывается методика создания паука, индексирующего пермский сегмент социальной сети. Описываются технические сложности, связанные с получением репрезентативной выборки по теме этнической агрессии, и пути их решения. Рассматривается использование векторной модели поиска по полученному индексу с пулом запросов по теме этнической агрессии как средство получения репрезентативной выборки. Обсуждаются результаты применения первоначального пула запросов и методика его модернизации. Проводится анализ полученного корпуса на предмет его репрезентативности и соответствующего пула запросов.

Ключевые слова: этническая агрессия в Интернете, поиск в Интернете, поиск по vk.com, формирование корпуса по заданной теме, поиск по социальным сетям.

1. Введение

В настоящее время Интернет полноценно отражает социальный срез нашего общества и представляет собой открытую площадку для высказываний и обмена информацией любого рода. В частности, Интернет применяется и для агитации и манипулирования общественным мнением, что нередко используется злоумышленниками при подготовке террористических актов и преступлений на почве межнациональной и межрелигиозной розни.

В целях профилактики таких преступлений необходимо проводить ряд мероприятий по анализу интернет-ресурсов, в которых проявляются феномены этнической агрессии.

Во-первых, необходимо провести классификацию феноменов этнической агрессии по следующим признакам:

- источнику информации (представитель террористической организации, представитель правоохранительных органов, частное лицо и др.);
- типу изложения информации (повествование о произошедшем событии, агитация, призыв к действиям, формирование общественного мнения и т.д.);
- тем действиям, к которым подталкивают источники информации;
- тем действиям, которые могут совершать потребители полученной информации;
- корреляции представленной информации с другими источниками и в чем именно;
- другим.

Во-вторых, необходимо проанализировать, какую именно информацию критически важно выявить для профилактики преступлений.

Все это позволит:

- более оперативно выявлять потенциальных преступников;
- контролировать информационный фон в группах риска;
- своевременно получать оперативную информацию.

Эти задачи можно решить только при наличии определенного корпуса, содержащего все подобные виды источников информации, которык можно будет проанализировать для обобщения вышеуказанных признаков.

Следует отметить, что наиболее эффективное практическое решение указанных задач возможно в контексте определенного региона государства. Привязка информации к региону даст правоохранителям дополнительную информацию при проведении оперативных мероприятий.

Таким образом, первым шагом на пути создания средства, позволяющего

автоматизировать процесс поиска и анализа феноменов этнической агрессии в Интернете, является создание корпуса интернет-ресурсов, содержащих данные феномены в определенном регионе страны.

2. Постановка задачи

Формирование корпуса интернет-ресурсов, содержащих феномены этнической агрессии в определенном регионе страны, включает задачу собственно выявления феноменов агрессии и задачу локализации полученного на первом этапе источника информации по конкретному региону.

Необходимо отметить, что первую задачу в большей степени можно решить путем полнотекстового поиска по ключевым словам. Однако задача локализации источника информации более комплексная и требует для своего решения более глубокого анализа текстов на естественном языке, в том числе и семантического.

Сложности задачи локализации обусловлены разнородностью источников информации, однако для определенного рода источников эта задача решается очень просто. Речь идет о социальных сетях, которые благодаря своей модели данных изначально представляют информацию о месте нахождения источника информации. Так, например, просматривая владельцев персональных страниц социальной сети «Вконтакте» [1], пользователь изначально может видеть место жительства интересующего его человека или группы.

Таким образом, для упрощения более глобальной задачи по формированию корпуса из разнородных источников в качестве источников информации мы осознанно будем рассматривать лишь сообщения, расположенные на страницах пермских пользователей социальной сети «Вконтакте».

В рамках данной статьи будет описываться решение задачи по выявлению феноменов этнической агрессии на страницах пермских пользователей социальной сети Вконтакте. Это позволит контролировать распространение нежелательной информации, выявлять его центры, выявлять опасных и потенциально опасных пользователей и групп, находить информацию, способную помочь при борьбе с этнической агрессией.

3. Архитектура решения задачи

Предлагается решать поставленную задачу в несколько этапов, каждый из которых является технической задачей.

На первом этапе создается краулер – программа, обходящая (имитирующая просмотр веб-страниц реальным пользователем) страницы пользователей пермского сегмента социальной сети «Вконтакте».

Краулер сканирует содержимое страниц пользователей и передает полученную информацию на сервер баз данных.

Таким образом, после работы первого этапа в базе данных сервера мы имеем определенное количество сообщений пермского сегмента социальной сети «Вконтакте». Данное количество сообщений определяется внутренней работой краулера, и о ней будет сказано в соответствующей главе.

На втором этапе полученная таким образом коллекция индексируется системой полнотекстового поиска Sphinx [2]. В результате чего к коллекции добавляется индекс, который с помощью системы Sphinx позволяет очень быстро вести полнотекстовый поиск по всем сообщениям коллекции.

На третьем этапе формируется специальный словарь терминов, употребление которых в документе может свидетельствовать о феноменах этнической агрессии. Далее по каждому из терминов словаря проводится поиск охватывающий коллекцию, и полученные результаты вручную ранжируются экспертом на наличие в документе феноменов этнической агрессии.

Таким образом термины словаря были отранжированы по уровню корреляции наличия в документе термина словаря и наличия в документе феноменов этнической агрессии.

Далее релевантные документы анализировались экспертами вручную, и в них выделялись новые термины, которые можно было занести в словарь.

Такой итеративный процесс происходил несколько раз, в результате чего была получена текущая версия словаря.

4. Краулер

Для сканирования страниц пользователей социальной сети «Вконтакте», далее Сети, на интерпретируемом высокоуровневом языке программирования Ruby [3] была разработана особая программа – краулер.

Особенность работы краулера состоит в том, что информация в Сети представляется с помощью динамических технологий. Для сканирования Сети краулеру требуется полностью имитировать работу реального пользователя. Это достигается формированием определенной последовательности запросов типа POST (полностью имитирующей, например, вертикальную прокрутку страницы), отправляемой на HTTP сервер Сети. Ответы сервера Сети, которые являются сообщениями со страниц пользователей, сохраняются краулером в базе данных СУБД MySQL [4].

Еще одной сложностью при разработке краулера стала особенность поведения пользователей, около 2/3 из них (по нашей статистике) закрывают свои страницы для неавторизированных пользователей. Для обхода

этой проблемы краулер авторизуется на сервере vk.com как зарегистрированный пользователь.

Однако при чрезмерной активности пользователь блокируется сервером vk.com, поэтому перед переходом на следующую сканируемую страницу краулер делает паузу в 1 с, 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 использованием авторизации и без, а также из-под прокси-сервера.

Режим без авторизации используется для предварительного сканирования, с целью выявления открытых страниц пользователей, что позволяет сэкономить время.

В первой версии программы список пользователей был получен парсингом выдачи стандартного поиска vk.com/search (в качестве выборки использовалось местоположение г. 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 г. Перми, сохранял их в базе данных и заносил в очередь, после чего процедура повторялась с каждым из них, таким образом удалось выявить необходимое количество пользователей.

Мощности краулера хватает для того, чтобы охватить один некрупный город. На сервере с 512 МБ RAM и полосой пропускания 100 Мб скорость краулера составляет в среднем 100 человек в секунду. Таким образом, сбор информации обо всех жителях Перми занимает около трех дней.

5. Поиск релевантных документов

В первой версии краулера собранная им коллекция документов составила 23318 сообщений, принадлежащих 1000 пользователей, что занимает 9.6 МБ в базе MySql. Время индексации составило 11.966 с, а скорость соответственно 804320 Б/с или 1948.54 документов/с. По этим данным был проведен поиск, в качестве запросов использовались термины словаря этнической агрессии. Однако релевантных результатов получить не удалось. Высказывания носили общий характер. Это говорило о том, что случайным образом выбранная сервером Сети 1000 пользователей не проявляет этнической агрессии в Сети.

Во второй версии коллекция была расширена до 1152586 сообщений 63770 пользователей и занимала 493.6 МБ. Время индексации составило 2163.567 с, а скорость 228142 Б/с или 532.72 документов/с. В этой выборке была достигнута релевантность по запросам – терминам словаря этнической агрессии. Общее количество найденных документов, содержащих феномены этнической агрессии, составило 241.

По релевантным сообщениям были найдены пользователи, публикующие видео с убийствами и расправами, демонстрацией оружия, съемками избиений в Перми, нацистскими жестами на фоне живого огня и т.д.

6. Формирование пула запросов

Изначально словарь терминов этнической агрессии составлял 80 терминов (полученных с помощью онтологического моделирования предметной области «терроризм»). Производился их поиск по собранной коллекции сообщений. Однако он не дал результатов, тогда были взяты энциклопедические статьи по теме экстремизма, и словарь стал содержать лишь 10 терминов, но узкоспециализированных.

По десяти запросам было найдено в общей сложности 241 сообщение, 40 сообщений экстремистской направленности и 201 сообщение, имеющее нейтральный характер. Таким образом, 0.062% сообщений, хранящихся в базе данных, содержат феномены этнической агрессии.

Для каждого запроса можно вычислить соотношение количества экстремистских сообщений, полученных поиском, к общему количеству полученных. Эту характеристику назовем показателем корреляции запроса с наличием в содержащем его документе феноменов этнической агрессии, или коротко показателем корреляции с феноменами агрессии. Таким образом, имеется возможность составить такой список запросов, по которым можно ожидать получение релевантной выборки, отранжированной по данному показателю.

Применительно к предметной области экстремизма запросы делятся на использующиеся для обозначения принадлежности к организации, религиозные и национальные, жаргонизмы, лозунги. Анализ сообщений позволяет расширить пул запросов новыми терминами и жаргонизмами, полученными из контекста.

7. Выводы

Сообщения, полученные с применением поиска по терминам из словаря, могут носить нейтральный характер, это происходит при цитировании новостей, репосте анекдотов или при использовании ключевых слов в любом не экстремистском контексте. Поэтому максимально релевантными (по отношению количества экстремистских высказываний к общему количеству высказываний) оказываются специальные слова, используемые только в экстремисткой среде.

Термины словаря имеют показатели корреляции с феноменами этнической агрессии на уровне до 52.6%.

Здесь в процентах и в виде простой дроби указано соотношение релевантных документов с

общим количеством документов, содержащих данные термины.

Эти исследования не окончены и имеют итеративный характер, а на данный момент представляют собой больше средство для анализа сообщений социальной сети «Вконтакте», нежели законченные исследования по получению словаря терминов этнической агрессии и соответствующего корпуса. Воспользоваться ограниченным функционалом разработанной поисковой системы можно по адресу в Интернете [5]. В дальнейшем планируется увеличить количество документов в коллекции, переработать словарь путем более глубокого анализа результатов поиска по терминам словаря и их корреляции.

Полученный таким образом словарь можно будет использовать для поиска ресурсов, содержащих феномены этнической агрессии не только в социальных сетях, но и в любых других ресурсах.

И наоборот, такими же методами можно использовать документы, про которые известно, что они содержат феномены этнической агрессии, для расширения словаря или получения особого словаря, специализированного для исследуемой коллекции.

Одним из направлений будущих исследований может быть повышение показателей корреляции запросов с феноменами этнической агрессии за счет выявления употребления терминов словаря в новостных цитатах и других вариантов употребления терминов словаря, которые снижают показатели корреляции.

Список литературы

1. Социальная сеть «Вконтакте» // URL: http://vk.com/
2. Платформа полнотекстового поиска // URL: http://sphinxsearch.com/
3. Язык программирования Ruby // URL: http://www.ruby-lang.org/en/
4. СУБД MySQL // URL: http://www.mysql.com/
5. Сервис поиска по пермскому сегменту социальной сети «Вконтакте» // URL: http://78.47.43.6/

Identifying of ethnic aggression phenomena in the Perm segment of social network vk.com

E. A. Rabchevskiy

Perm State National Research University, Bukirev st., 15, Perm, 614990, evgeny@rabchevsky.name

This paper deals with the creation of the corpus of posts from Perm segment of the social network vk.com, containing the phenomena of ethnic aggression. Describes how to create a spider indexing Perm segment of social networks. Describes the technical-cal difficulties associated with obtaining a representative sample on ethnicity of aggression, and their solutions. Discusses the use of the vector model search of the resulting index to a pool requests on ethnic aggression as a means of obtaining a representative sample. We discuss the results of the original pool of requests and the method of its modernization. The analysis of the resulting housing for its representativeness and appropriate pool requests.

Keywords: Ethnic aggression in Internet; search in the Internet; search in vk.com; creation of corpus on a given topic, search on social networks.