Разработка и реализация сервисов по сбору данных социальных сетей в целях улучшения среды обитания человека

И.А. Рыцарев[1], А.В. Благов[1], М.И. Хотилин[1]

[1]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086

Аннотация. В статье обсуждается необходимость разработки и реализации сервисов по сбору данных социальных сетей для выявления проблем среды обитания человека. Авторы предлагают методы и инструменты для сбора и анализа данных по выявленным проблемам из социальных сетей.

1. Введение

В настоящее время существует множество способов мониторинга среды обитания человека, такие как съёмки со спутников или беспилотных летательных аппаратов, оснащение камерами транспорта, либо установка стационарных камер и т.д. Зачастую эта задача решается силами определенных муниципальных служб. Как правило, такой подход связан с большими трудозатратами. Определенную популярность приобретают такие информационные сервисы, как тематические интернет сайты, к примеру: https://rosyama.ru/, http://moyasamara.com/ и т.д. Стоит отметить, что для подачи заявки через данные сайты необходимо как минимум помнить название ресурса, его адрес. Более того, создавая заявку, её автор должен заполнить информацию о себе. Среди некоторых людей эти ограничения могут создать некоторые трудности. В то же время все большую популярность приобретают социальные сети [1]. У многих людей имеются смартфоны и иные персональные мобильные гаджеты, снабженные возможностью выхода в социальные сети. Благодаря этому каждый пользователь генерирует большое количество данных, которые могут предоставлять интерес для разных направлений и сфер деятельности. Сбору, обработке и анализу данных социальных сетей посвящено много научно-исследовательских работ [2-4]. Стоит отметить, что для пользователя социальных сетей сами сети могут являться наиболее удобным ресурсом для размещения информации, в том числе и о различных проблемах окружающей среды: свалки, аварии, ямы, пожары и т.д.

В данной статье описывается реализованный сервис для оперативного сбора информации о проблемах среды обитания в одной из самых популярных социальных сетей Twitter. Сервис позволяет осуществлять сбор необходимой информации в режиме on-line по любой необходимой геолокации.

2. Алгоритм сбора данных, их и обработка и классификация

Задача сбора необходимой информации из социальных сетей о проблемах окружающей среды может быть поделена на сбор данных, их фильтрацию, обработку и, если необходимо, классификацию. Наибольшую ценность для получения информации о различных проблемах среды обитания представляют данные генерируемые в режиме on-line. При этом фильтрация может идти как по определенной, интересующей геолокации, так и по предмету публикуемого контента.

Разработанный в рамках исследования алгоритм сбора данных, помимо обозначенной фильтрации использует и дополнительный элемент обратной связи с пользователем посредством автоматического запроса об уточнении информации об опубликованном сообщении. На рисунке 1 представлена схема алгоритма сбора необходимых данных с социальной сети.

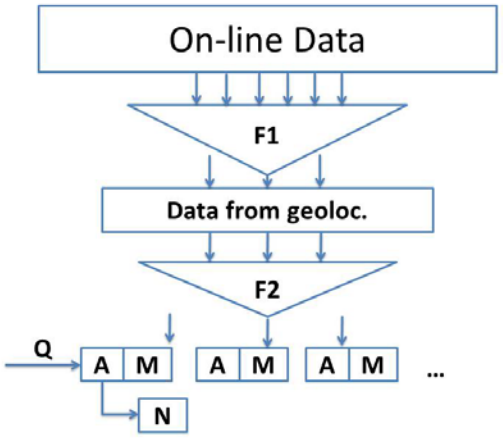

Рисунок 1. Обобщенная схема алгоритма сбора данных социальной сети.

Генерируемые в режиме on-line данные социальной сети собираются с помощью фильтра *F1* по обозначенной геолокации. Далее при помощи контентного фильтра *F2*, настроенного по ключевым словам или тематикам, собираются необходимые данные о проблемах среды обитания. Полученные при помощи двойной фильтрации данные схематично можно разделить на две части: данные об авторе – *A* и само сообщение – *M*. При этом, согласно разработанному алгоритму, автору сообщения автоматически приходит уточняющий запрос *Q*, на предоставление дополнительной информации, такой как указание точного адреса или координат возникшей проблемы, описанной в сообщении *M*, либо о её характере. Автор *A*, получая данный запрос, оставляет сообщение *N*, содержащее необходимую информацию. Можно сказать, что по разработанному алгоритму собирается следующий необходимый набор данных: $\sum_{i=1}^{k}(M_{gi}+N_{gi})$, для всех k пользователей A_i, авторов сообщений M_{gi}, сгенерированных по геолокации g, и приславших дополнительную информацию N_{gi}.

После сбора необходимых данных часто требуется произвести их классификацию по отнесению к тем или иным проблемам: свалки, ямы, аварии и т.д. Для решения это задачи можно использовать ключевые слова «хештеги» данного сообщения [5]. Для более детального анализа может использоваться разработанный в рамках данного исследования алгоритм коллективного принятия решения.

По любому собранному сообщению каждое слово из текста сообщения $T = t_1 t_2 t_3 ... t_n$ сопоставляется со словарем $U = u_1 u_2 u_3 ... u_m$ с целью получения списка категорий:

$$\delta(u_i) = \{\underset{1}{k_j} \underset{2}{...} \underset{3}{k_h}\}_z,$$

где $\delta(u_i)$ - функция извлечения категорий слова u_i из словаря;

$\{k_{j_1}, k_{h_2}, k_{3_z}\}$ - вектор из z элементов k (элемент списка категорий $K = k_1 k_2 k_3 ... k_l$), к которым принадлежит слово u_i;

$1 \le j, h, z \le l$.

Если $\gamma(t_i) = u_i$ ($\gamma(t_i)$ - функция стемминга текста), то $\delta(t_i) = \delta(u_j) = \overline{k}$. В случае отсутствия слова в словаре оно отдается эксперту на классификацию с последующим добавлением в словарь.

Отличительной особенностью данного алгоритма коллективного принятия решения является наличие динамического порогового значения: при определении порогового значения учитывается количество и вес категорий слов, упоминающихся в тексте.

В результате обработки текста мы имеем перечень категорий (к которым относятся слова) на осонове которого алгоритм высчитывает пороговое значение «веса» тематики и классифицирует исходный текст.

3. Результаты и обсуждения

Сбор данных социальной сети Twitter может осуществляться посредством программных продуктов Apache Ambari и Flume, подробнее данный метод описан в [6] Однако для сбора данных с применением ряда фильтров зачастую удобнее разработать свой программный продукт с использованием стандартных библиотек (twitter4j, tweepy и т.п.) [7].

В рамках данного исследования был разработан программный продукт на языке программирования Python, содержащий модуль авторизации, модуль сбора данных и модуль фильтрации. Данный программный продукт позволяет собирать данные по геолокации, по ключевым словам, по пользователю, а также кэшировать все медиафайлы пользователя. Для исключения перебоев в работе программного продукта связанных с превышением лимитов установленных социальной сетью Twitter в программный продукт вшито множество ключей авторизации. Программный продукт работает в режиме real-time мониторинга, а также может делать запросы на получение лежащей на серверах информации.

Для каждого пользователя социальной сети Twitter реализованным интерфейсом для взаимодействия является бот-аккаунт @contriol63. Геолокационный фильтр настроен на шестьдесят третий регион Российской Федерации, которым является Самарская область.

Пользователь, собирающийся написать в своем сообщении о возникшей проблеме окружающей среды, может добавить в своё сообщение либо имя бота-аккаунта, либо его адрес, либо набор ключевых слов, представленных в таблице №1.

Таблица 1. Ключевые слова, необходимые для употребления в сообщении.

Проблема	Spacing
несанкционированная свалка	#свалка63, свалка63, #svalka63, #свалка63, #dump63, dump63, #мусор63, мусор 63, #musor63, musor63, #trash63, trash63
пожар	#пожар63, пожар63, #pozhar63, pozhar63, #fire63, fire63
ямы на дорогах	#яма63, яма63, #yama63, yama63, #pit63, pit63
авария	#авария63, авария63, #avariya63, avariya63, #breakdown63, breakdown63
опасность	#опасность63, опасность63, #opasnost63, opasnost63, #danger63, danger63
нарушение	#нарушение63, нарушение63, #narushenie63, #narushenie63, #disturb63, disturb63
без обозначения по имени или адресу бота	control63, @control63, контроль63, @контроль63

Список, представленный в таблице №1, может дополняться.

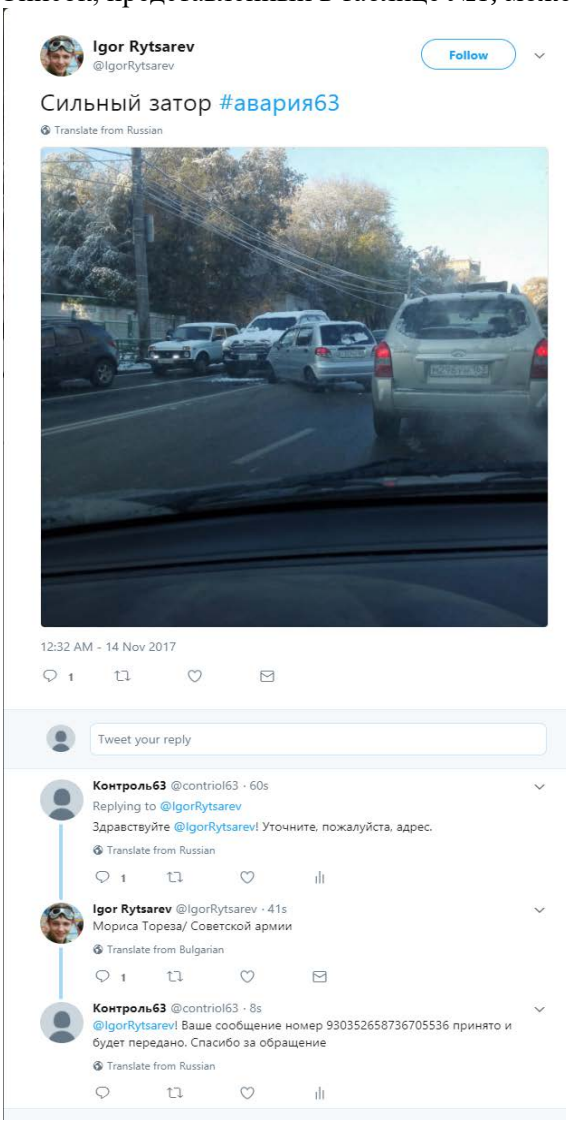

Рисунок 2. Уведомление о получении и уточняющий вопрос в сети Twitter.

Написанное сообщение, содержащее какое-либо из ключевых слов в таблице 1, с геолокацией по 63 региону собирается и классифицируется по типу. Автору сообщения автоматически направляется уведомление о сборе и задается уточняющий вопрос (рисунок 2).

Собранный пул сообщений о различных проблемах среды обитания человека представляет собой ценную информацию и может служить для оперативного адресного устранения возникших аварий и нарушений. При этом сервис представляется очень удобным для всех пользователей, которые просто продолжают пользоваться своими социальными сетями, выкладывая туда информацию о возникших проблемах. Стоит также отметить, что разработанный сервис может иметь практическую ценность лишь в случае его поддержки структурами и службами, занимающимися решением проблем среды обитания человека. Основной мотивацией для пользователей, выкладывающих информацию по данной теме, может быть лишь оперативная реакция и последующее решение проблем.

4. Выводы

Результатом работы является разработка и реализация сервиса по сбору необходимой информации о возникающих авариях и нарушениях в социальной сети Twitter по определенной геолокации. Дополнительно также реализован инструмент классификации, собранных сообщений.

В дальнейшем разработанный сервис может быть: во-первых масштабирован на другие социальные сети, к примеру, ВКонтакте, и синтерирован между ними, а во-вторых может быть дополнен обработчиком изображений (фотографий, прикрепляемых к сообщению) и видео.

5. Благодарности

The work has been performed with partial financial support from the Ministry of Education and Sciences of the Russian Federation within the framework of implementation of the Program for Improving the Samara university Competitiveness among the World's Leading Research and Educational Centers for the Period of 2013-2020s.

6. Литература

[1] Tan, W. Social-network-sourced big data analytics / W. Tan, M. Blake, I. Saleh, S. Dustdar // IEEE Internet Computing. – 2013. – Vol. 5. – P. 62-69.

[2] Semertzidis, K. How people describe themselves on Twitter / K. Semertzidis, E. Pitoura, P. Tsaparas // Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks, 2013. – P. 25-30.

[3] Xu, X. Scan: a structural clustering algorithm for networks / X. Xu et al. // Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2007. – P. 824-833.

[4] Blagov, A. Big Data Instruments for Social Media Analysis / A. Blagov, I. Rytcarev, K. Strelkov, M. Khotilin // Proceedings of the 5th International Workshop on Computer Science and Engineering, 2015. – P. 179-184.

[5] Krokos, E. A look into twitter hashtag discovery and generation / E. Krokos, H. Samet, J. Sankaranarayanan //Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks. – ACM, 2014. – P. 49-56.

[6] Рыцарев, И.А. Построение моделей активности пользователей социальных сетей / И.А. Рыцарев, А.В. Благов // Информационные технологии и нанотехнологии. – 2015. – С. 216-220.

[7] Rytsarev, I.A. Development and research of algorithms for clustering data of super-large volume / I.A. Rytsarev, A.V. Blagov // CEUR Workshop Proceedings. – 2017. – Vol. 1903. – P. 80-83.

Development and implementation of services to collect social networking data in order to improve the human environment

I.A. Rytsarev[1], A.V. Blagov[1], M.I. Khotilin[1]

[1]Samara National Research University, Moskovskoe shosse 34A, Samara, Russia, 443086

Abstract. The article discusses the need to develop and implement services for the collection of social networking data for the detection of environmental problems. The authors offer methods and tools for collecting and analyzing data on identified problems from social networks.

Keywords: Social networks, Data mining, Earth Remote Sensing, Twitter, Data analysis, Environment improving.