ИДЕНТИФИКАЦИЯ УЧАСТНИКОВ ОБЩЕСТВЕННЫХ МЕРОПРИЯТИЙ НА ОСНОВЕ ДАННЫХ СОЦИАЛЬНЫХ СЕТЕЙ

Якубович А.В., Лежнёв А.В.
Кафедра информационных технологий автоматизированных систем, кафедра систем управления,
Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: yakubovichandreiv@gmail.com, anton.lezhneiv@yandex.by

В статье предлагается анализ данный пользователей социальных сетей для сбора информации о реальных событиях на основе определения местоположения, схожести контента и социальных связей источников. Рассматривается регрессия гауссовского процесса для определения расстояния между источником сообщения о событии и предполагаемым событии.

Введение

Широкое использование социальных сетей, особенно их услуг на основе определения местоположения, превратило социальные сети в важный источник информации о реальных событиях. В последние годы было разработано множество систем обнаружения событий, использующие геотегированные посты из социальных сетей. Данная технология востребована правительственными чиновниками, полицией, и средствами массовой информации так как позволяет определить участников мероприятия, от которых можно получить ценную информацию о самом событии или расследовать причастность подозреваемых если событие связано с преступлением или террористом. Однако из-за высокой неопределенности мобильности людей,и низкой вероятности того, что пользователи делятся своей информацией о местоположении, трудно идентифицировать пользователей на месте события.

В этой статье мы предлагаем модель на основе функции сплавленной регрессии гауссовского процесса, которая использует три влиятельных фактора в социальных сетях для идентификации источников сообщения на месте: влияние мобильности, сходство контента и социальные отношения источников. Объединив эти факторы, мы можем оценить расстояние между источником сообщения и событием даже если профиль местоположения источника неизвестен [1].

I. Постановка задачи

Для каждого социального события E, которое происходит в окно времени T, обычно существует группа связанных с событиями источников $U_{all}(u1, u2, ...,)$, которые размещают соответствующую информацию о событии, включая активные источники U_{act} и обычные источники U_{nor}. Однако сведения о местоположении могут быть неопределёнными, учитывая конфиденциальность участников. Поэтому все абсолютные данные по GPS будут преобразованы в относительные расстояния как в моделировании признаков, так и в высчете результатов. И расстояние между каждым источником U_i в U_{all} и событием E определяется как $Dis(U_i, E)$. Задача исследования - оценка $Dis(U_i, E)$ для содействия выявлению пользователей на местах. Предполагается, что малое $Dis(U_i, E)$ означает более высокую возможность что U_i присутствует на месте. Поэтому, основная проблема - как определить $Dis(U_i, E)$ точно и эффективно.

В частности, социальное событие E может быть представлен как совокупность трёх величин $E < \Omega, \Theta, \Phi >$, где Ω -множество источников, Θ тема и Φ картина социальных связей извлеченная из Ω. Каждый источник U_i в U_{all} также моделируется тремя параметрами $U < \omega, \theta, \phi >$, где ω - недавние упоминания в сообщениях источника, которые указывают на его/её социальные связи, θ содержит историю тем поднимаемых источником и ϕ представляет индивидуальную картину мобильности этого источника. В соответствии с этими двумя определения, извлекаются три параметра: Влияние мобильности, Сходство содержания и Социальные отношения $(F_{MI}, F_{CS}F_{SR})$ между источником и событием. Предположим, что существует функция подобия $Sim(x, y)$, которая измеряет сходство между двумя данными формата x и y.

Можно предположить, что $F_{MI} \propto Sim(\Phi, \phi)$, $F_{CS} \propto Sim(\Theta, \theta)$ и $F_{SR} \propto Sim(\Omega, \omega)$. И задача сводится к оценке $Dis(U, E)$, что на основе модели регрессии гауссовского процесса (GPR) окончательно сформулировано как

$$Dis(U, E) = GPR(F_{MI}, F_{CS}, F_{SR}) \qquad (1)$$

В этой функции, мы предполагаем, что наиболее высокое сходство между источником и целевым событием по всем трём функциям - наименьшее расстояние между источником и событием, таким образом это приводит к более высокой вероятности того, что пользователь может быть на месте. На рисунке 1 представлена предлагаемая структура состоящая из четырех компонентов:

1. сбор данных и предварительная обработка;

2. моделирование социальных событий;
3. моделирование пользователя;
4. оценка расстояния между социальными событиями.

Рис. 1 – Структура системы

II. Алгоритм предположения

Для оценки расстояния между источником сообщения о событии и событием предложена функция регрессии гауссовского процесса (FEGOR). Рабочий процесс FEGOR представлен на рисунке 2.

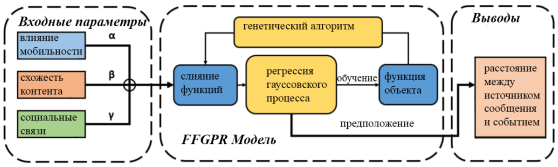

Рис. 2 – Схема процесса FAGOR

Предыдущие исследования показали, что мобильность человека может быть смоделирована на основе броуновского движения [2] с длиной шага имеющей гауссовское распределение [5]. На основе этого наблюдения мы моделируем расстояние между источником сообщения и расположением социального события как Гауссовское распределение, на среднее и дисперсию которого могут влиять различные факторы. Необходимо выявить ключевые факторы из мобильности пользователя, контента сообщения и особенности социальных отношений. Следовательно, необходимо применять регрессию гауссовского процесса (GPR) [6].

III. Заключение

Был предложение алгоритм на основе модели регрессии гауссовского процесса (FEGOR), которая эффективно включает три параметра: влияние мобильности, сходство контента и социальные отношения, которые извлекаются из данных источников, связанных с событиями. В процессе предположения местоположения события все абсолютные данные о местоположении источников преобразуются в относительные целью защиты конфиденциальности расположения источников.

В сравнении с другими алгоритмами, математическая модель на основе регрессии гауссовского процесса показывает наилучший результат. Сравнение приведено в таблице 1.

Таблица 1 – Сравнение алгоритмов

Метод	Параметры			Средняя ошибка км.
	Мобильность	Контент	Соц. связи	
FEGOR	**Да**	**Да**	**Да**	**1.7**
Ryoo [3]	Да	Нет	Да	10
OLIM [4]	Да	Да	Нет	28.5

Эта работа основывается на предыдущих исследованиях на основе местоположения [7,8] и фокусируется на поиске пользователей в местах общественных мероприятия, что может иметь огромную ценность для многих социально-ориентированных приложения. Например, данные могут быть полезны правительственными чиновниками, чтобы лучше понять модели различных общественных мероприятий (например, парад, протест) и разработать лучшие стратегии управления толпой. И на основе показания обнаруженных на месте участников, правительственные или полицейские силы могут получить более подробную и значимую информация о произошедшем. Данные алгоритмы могут использоваться для любых социальных сетей, в том числе, русскоязычных.

IV. Список литературы

1. Cheng Z, Caverlee J, Lee K. You are where you tweet: a content-based approach to geo-locating twitter users. Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010: 759-768.
2. Camp, Tracy, Jeff Boleng, and Vanessa Davies. “A survey of mobility models for ad hoc network research.” Wireless communications and mobile computing 2.5
3. Ryoo K M, Moon S. Inferring Twitter user locations with 10 km accuracy. Proceedings of the companion publication of the 23rd international conference on World wide web companion. International World Wide Web Conferences Steering Committee, 2014: 643-648.
4. Yamaguchi Y, Amagasa T, Kitagawa H, et al. Online User Location Inference Exploiting Spatiotemporal Correlations in Social Streams. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM, 2014: 1139-1148.
5. Jiang, Bin, Junjun Yin, and Sijian Zhao. “Characterizing the human mobility pattern in a large street network.” Physical Review E 80.2 (2009): 021136. (2002): 483-502.
6. Rasmussen, C. E., and Williams, C. K. I. Gaussian processes for machine learning. The MIT Press, 2006
7. Wang Y, Kankanhalli M S. Tweeting Cameras for Event Detection. Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015: 1231-1241.
8. Pan B, Zheng Y, Wilkie D, et al. Crowd sensing of traffic anomalies based on human mobility and social media. Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2013: 344-353.