Подходы к определению основного места проживания пользователей социальных сетей на основе социального графа

[1]Ю.С. Трофимович <integral@ispras.ru>
[1]И.С. Козлов <kozlov-ilya@ispras.ru>
[1,2,3]Д.Ю. Турдаков <turdakov@ispras.ru>
[1]Институт системного программирования РАН
109004, Россия, г. Москва, ул. А. Солженицына, д. 25
[2]Московский государственный университет имени М.В. Ломоносова,
119991, Россия, Москва, Ленинские горы, д. 1
[3]Национальный исследовательский университет «Высшая школа экономики»
101000, Россия, Москва, ул. Мясницкая, д. 20

Аннотация. В статье рассматриваются подходы к определению основного места проживания пользователей социальных сетей по графу образуемому в результате установления двунаправленной связи — “дружбы”. Предложен подход, базирующийся на векторном представлении вершин графа и последующем применении алгоритма классификации на основе обучения с учителем. Приведены результаты экспериментов и сравнение с референсными подходами. Показано, что предложенный подход сопоставим по качеству с другими подходами.

Ключевые слова. определение местоположения; социальные сети; социальный граф; векторное представление вершин графа.

DOI: 10.15514/ISPRAS-2016-28(6)-13

Для цитирования: Трофимович Ю.С., Козлов И.С., Турдаков Д.Ю. Подходы к определению основного места проживания пользователей социальных сетей на основе социального графа. Труды ИСП РАН, том 28, вып. 6, 2016 г., стр. 185-196. DOI: 10.15514/ISPRAS-2016-28(6)-13

1. Введение

Современные социальные сети являются не только средством общения, но так же инструментом маркетинга, рекламы и социальных исследований. Так, в статье Sakaki, Okazaki и Matsuo [1] описывается система быстрого обнаружения и оповещения о землятресениях основанная на анализе сообщений “Твиттера”

в реальном времени. В статьях Lamb, Paul и Dredze [2] и Sadilek, Kautz и Silenzio [3] используются сообщения той же социальной сети для обнаружения эпидемий гриппа и моделирования его распространения.

Для многих из этих приложений необходима информация о местоположении пользователей, которая не всегда доступна. Среди пользователей “Твиттер” лишь около 26%[4] указывают в профиле название города, около 30% пользователей “ВКонтакте” оставляют это поле пустым. Так появляется задача определения места проживания пользователей социальных сетей.

Подходы к решению задачи удобно классифицировать по информации о пользователях, на которую эти подходы опираются. Большинство подходов сконцентрированы на анализе 1) социального графа 2) пользовательского контента (фотографии, текст сообщений), либо 3) используют гибридный подход, комбинируя первые два [5].

В то время как подходы, опирающиеся на анализ контента довольно широко используют методы машинного обучения, среди подходов основанных на анализе социального графа машинное обучение не столь популярно.

Одна из сложностей, препятствующих этому — необходимость конструирования признаков на основе социального графа. Как правило методы машинного обучения с учителем принимают в качестве признаков числовой вектор конечной размерности и не могут работать с графом непосредственно. Для извлечения признаков из графа были разработаны методы, сопоставляющие вершинам графа вектора конечной размерности.

В данной статье мы приводим краткий обзор работ в области определения местоположения пользователей социальных сетей и предлагаем подход, основанный на векторном представлении вершин графа и методах машинного обучения с учителем. Стоит отметить, что предложенный подход не имеет какой-либо специфики, связанной с местоположением. Он может так же использоваться для оценки любых других демографических атрибутов, оказывающих влияние на формирование связей в обществе. Так, схожий подход использовался Perozzi и Skiena [6] для определения возраста пользователей.

Также мы приводим результаты экспериментов и сравнения с выбранными референсными подходами.

2. Обзор существующих работ

2.1 Определение места проживания по социальному графу

Начало исследованиям в области определения местоположения пользователей социальных сетей по социальному графу положила работа Backstrom, Sun и Marlow [7]. В ней авторы исследуют уже отмеченное ранее влияние расстояний на структуру социальных связей и предлагают вероятностную модель для

получения предполагаемого места проживания пользователей социальной сети “Facebook”.

В работе Jurgens [8] было предложено использовать метод распространения меток, суть которого заключается в итеративном выполнении следующих шагов:

1. определение множества пользователей, чье место проживания неизвестно;
2. выполнение метода select для пользователей, определенных на этапе 1;
3. добавление в список пользователей с известным местоположением тех пользователей, чье местоположение было определено на этапе 2.

Метод select() возвращает предполагаемое место проживания для данного пользователя основываясь на информации о его друзьях. В работе [8] рассмотрено несколько реализаций этого метода, за основной референсный подход мы взяли незначительно модифицированную реализацию под названием “Traditional Label Propagation” — ее возвращаемым значением является самое часто встречаемое место проживание среди друзей данного пользователя.

Методы машинного обучения активно используются для определения местоположения по тексту сообщений, но не нашли широкого применения в анализе социальных графов.

Jia и др. [9] в своей работе отмечают что эффективность подхода основанного на анализе социального графа во многом зависит от подмножества друзей, выбранного для анализа и предлагают использовать Random Walk with Restart (RWR) для сортировки друзей по степени “влияния”. Там же приводятся теоретические оценки максимальной точности подходов на основе анализа социального графа и их сравнение с имеющимися подходами.

Chen, Liu и Zou [10] так же используют RWR для оценки “похожести” пользователей основываясь на социальном графе, и предлагают графовую модель — Social Tie Factor Graph (STFG).

2.2 Векторное представление вершин графа

2.2.1 PCA на графах

Ранние работы по представлению вершин графа основывались на разложении матрицы Кирхгофа соответствующего графа [11]. Такой подход получил название PCA на графах. Однако, было показано что проекции не всегда подходят для использования в качестве векторного представления — в то время как проекции сильно различающихся вершин находятся далеко друг от друга, проекции похожих вершин оказываются слишком близко [12].

2.2.2 DeepWalk

Perozzi, Al-Rfou и Skiena [13] предложили модель DeepWalk. DeepWalk использует случайные блуждания для получения векторного представления вершин графа. Работа состоит из двух этапов:

1. Случайные блуждания по графу порождают последовательности вершин.
2. На основе полученных последовательностей вершин получают векторное представление вершин. Для этого используют неглубокую нейронную сеть, по аналогии с моделью word2vec [14].

2.2.3 BLM

В другом методе, предложенном в работе [12], получение векторного представления вершин графа базируется на вероятностной модели рёбер. В данном подходе каждая вершина представлена двумя векторами, $\vec{In} \in \mathbb{R}^D$ и $\vec{Out} \in \mathbb{R}^D$ (где D — размерность векторного пространства), которые являются векторным представлением входящих и исходящих рёбер этой вершины соответственно. Вероятность ребра между вершинами $v \to u$ вычисляется по формуле

$$p(v|u) = \frac{\exp(In_u^T Out_v)}{\sum_{w \in V} \exp(In_u^T Out_w)}$$

Векторное представление может быть получено за счёт максимизации вероятностей (принцип максимума правдоподобия), однако непосредственное получение векторного представления затруднено необходимостью вычислять сумму $\sum_{w \in V} \exp(In_u^T Out_w)$. Вместо этого авторы работы используют метод контрастной оптимизации [15].

3. Референсные подходы

3.1 “Случайный друг”

Местоположение каждого пользователя считается равным местоположению одного случайно выбранного друга.

3.2 Подход на основе распространения меток

Подход основан на работе Jurgens [8], используемая модификация описана в Алгоритме 1. Пусть: U — множество пользователей социальной сети, N — отображение пользователей в списки друзей, M — отображение пользователей в местоположения, k — количество друзей, при котором определение местоположения по друзьям считается возможным. Метод select() возвращает самую часто встречающуюся метку. Алгоритм возвращает в качестве результата своей работы E — итоговое отображение пользователей в местоположения. Основным отличием от подхода, описанного в разделе 2, является параметр k. В экспериментах k задавалось равным 5.

Data: U, N, M, k
Result: E
$E = M$;
$E' = \emptyset$;
while $E \neq E'$ **do**
  $E' = E$;
  for $u \in U$ - *domain*(E) **do**
    $L :=$ пустое множество;
    for $n \in N(u)$ **do**
      if $n \in domain(E)$ **then**
        добавить $E(n)$ в L;
      end
    end
    if $|L| > k$ **then**
      $E(u) = \text{select}(L)$;
    end
  end
end

Алгоритм 1. Алгоритм определения места проживания на основе распространения меток.

Algorithm 1. Algorithm for inferring location of residency based on label propagation

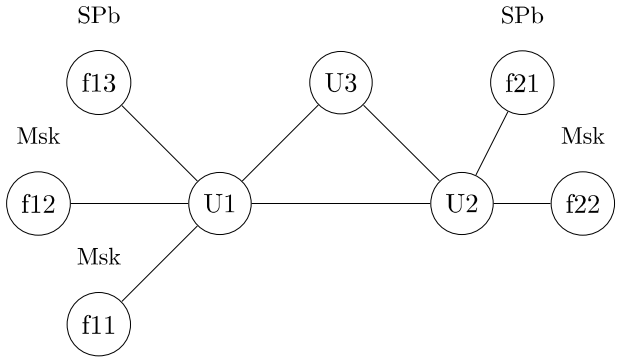

Рис. 1. Пример социального графа

Fig. 1. Example of a social graph

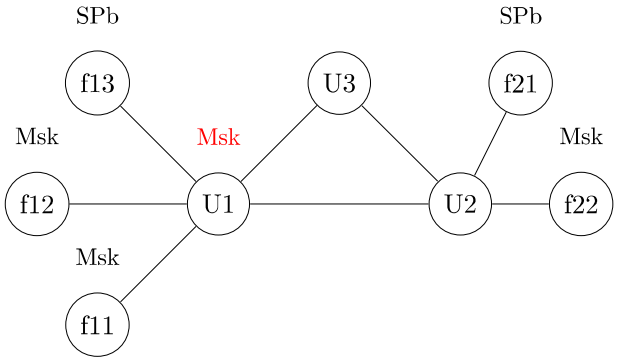

Рис. 2. Социальный граф после первой итерации подхода на основе распространения меток

Fig. 2. Social graph after first iteration of the approach based on label propagation

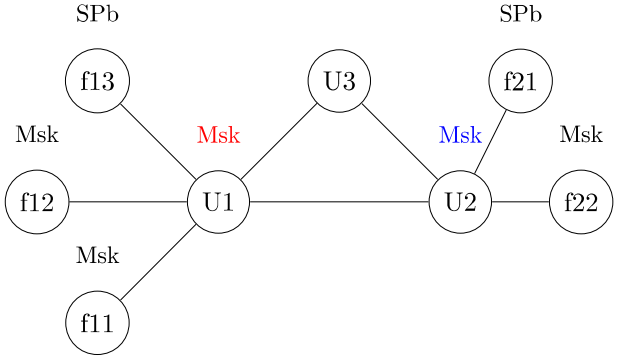

Рис. 3: Социальный граф после второй итерации подхода на основе распространения меток

Fig. 3. Social graph after second iteration of the approach based on label propagation

Небольшая иллюстрация приведена на рисунках 1-3. Для простоты предлагается взять k равным 3. На рисунке 1 приведен пример социального графа. Для пяти пользователей место проживания считается известным, для трех (U1, U2 и U3) — неизвестным. На каждой итерации выбираются пользователи, у которых минимум k друзей с известным местом проживания. При k равным 3 для данного социального графа на первой итерации будет отобран только пользователь U1. В качестве места проживания будет взято наиболее популярное местоположение среди друзей — для U1 это Msk 2. После этого начнется вторая итерация. Если считать место проживания U1 известным, становится возможным предсказать место проживания U2 — у него получается ровно три друга с известным местом проживания. После присвоения U2 Msk в качестве места проживания вторая итерация заканчивается 3. На момент начала

третьей итерации только U3 остается с неизвестным местоположением, но у него меньше k друзей с известным местоположением, поэтому в ходе третьей итерации граф остается неизменным и работа алгоритма заканчивается.

4. Подход на основе векторного представления вершин графа

Предлагаемый нами подход для определения места постоянного проживания пользователя состоит из следующих этапов:

1. Построение векторного представления вершин социального графа.
2. Обучение классификатора с использованием полученных на первом этапе векторов в качестве признаков.

В качестве метода получения векторного представления вершин был использован метод BLM, описанный в разделе 2.2.3. В социальной сети “Вконтакте” преобладают ненаправленные рёбра(связь типа дружба). По этой причине, а также в целях экономии памяти было принято $\vec{In} \equiv \vec{Out}$.

Данный подход отличает автоматическое извлечение векторов признаков из графа, что позволяет отойти от ручного конструирования признаков. После получения векторного представления вершин задача сводится к стандартной задаче машинного обучения — классификации. Каждый пользователь представлен конечномерным числовым вектором(вектор признаков), вектору признаков необходимо сопоставить один из конечного числа классов.

Нами были испробованы различные методы клас- сификации: логистическая регрессия, случайный лес, XGboost, а также многослойные нейронные сети. Результаты работы системы были оценены с помощью кросс-валидации. Наилучшие результаты показала многослойная нейронная сеть. Нами была использована библиотека Keras [16].

5. Эксперименты

5.1 Данные

Все эксперименты проводились на данных социальной сети “ВКонтакте”. За “золотой стандарт” бралось местоположение указанное самими пользователями в профиле. Данные о местоположении пользователя, предоставляемые сетью ВКонтакте, содержат в себе:

1. идентификационный номер пользователя;
2. идентификационный номер местоположения;
3. строковое описание местоположения (например, название города).

В качестве единиц местоположения для РФ были выбраны субъекты РФ, для зарубежных пользователей — страны. Каждой имеющейся паре {Идентификационный номер местоположения, Строковое описание местоположения} был сопоставлен регион РФ или страна. Для получения

региона РФ по строковому описанию местоположения использовался сервис Яндекса “Геокодер”, для получения названия страны по строковому описанию местоположения использовался сервис openstreetmap.org. Всего данные состояли из 156057700 профилей пользователей. Место проживания было известно для 99055808 из них.

5.2 Метрики

Для оценки подходов были выбраны следующие метрики:

1. доля верно определенных регионов (accuracy);
2. F1-мера с усреднением “Macro” — невзвешенное среднее значение F1-мер;
3. F1-мера с усреднением “Weighted” — взвешенное пропорционально присутствию в датасете среднее значение F1-мер;
4. точность с усреднением “Macro” — невзвешенное среднее значение точностей;
5. точность с усреднением “Weighted” — взвешенное пропорционально присутствию в датасете среднее значение точностей;
6. полнота с усреднением “Macro” — невзвешнное среднее значение полноты;
7. полнота с усреднением “Weighted” — взвешенное пропорционально присутствию в датасете среднее значение полноты.

5.3 Результаты

Результаты тестирования приведены в таблицах I и II. В таблицах используются следующие обозначения: RN — подход “Случайный друг”, LP — подход на основе распространения меток, GE — подход на основе векторного представления. В таблице 1 приведены значения метрик для всех предсказанных пользователей, вне зависимости от их активности в социальной сети. В таблице 2 приведены значения метрик только для пользователей, оставивших хотя бы один комментарий.

Для подхода на основе распространения меток использовалось сокрытие 85% известных меток. То есть местоположение 15% (15003815) выбранных случайным образом пользователей считалось известным, остальные же 85% требовалось предсказать.

Определенный интерес представляют метрики замеренные с макроусреднением, и большая разница в значениях одних и тех же метрик с усреднением “Macro” и “Weighted”. Одна из особенностей используемого датасета в неравномерности представленных классов — регионов. Редкие и слабо представленные регионы сложны для предсказания, и при неверных ответах классификатора существенно занижают значения метрик с макроусреднением. “Weighted”-метрики напротив, используют долю пользователей из определенного региона в качестве веса этого региона, и больше подходят для оценки подходов. Небольшое отставание подхода на

основе векторного представления от подхода с распространением меток по взвешенной точности, и значительно более высокие значения полноты и F1-меры позволяют заключить что предложенный подход на основе векторного представления может быть эффективно использован для определения места проживания пользователя как сам по себе, так и в комбинации с классификаторами на основе пользовательских данных.

6. Заключение

В данной статье мы привели краткий обзор работ в области определения места проживания пользователей социальных сетей, в том числе с использованием машинного обучения. Так же нами был предложен подход на основе векторного представления вершин графа и нейронной сети и приведены результаты экспериментов, показывающие применимость предложенного подхода.

Табл. 1. Результаты тестирования: все пользователи (вне зависимости от наличия комментариев)

Обозначения: RN — подход "Случайный друг", LP — подход на основе распространения меток, GE — подход на основе векторного представления

Table 1. Test results: all users (independent of comment presence)

Notation: RN — "Random friend" approach, LP — approach based on label propagation, GE — approach based on vector representation

Метрика		Метод		
		RN	LP	GE
Accuracy		0.417	0.463	**0.516**
Macro	Precision	0.357	**0.522**	0.308
	Recall	0.180	0.150	**0.167**
	F1-score	0.231	0.192	**0.203**
Weighted	Precision	0.654	**0.760**	0.729
	Recall	0.417	0.463	**0.516**
	F1-score	0.506	0.555	**0.595**

Табл. 2. Результаты тестирования: только пользователи оставившие комментарии

Table 2. Test results: only users who have posted comments

Метрика		Метод		
		RN	LP	GE
Accuracy		0.644	0.786	**0.839**
Macro	Precision	0.245	**0.374**	0.302
	Recall	0.225	0.241	**0.268**
	F1-score	0.234	0.255	**0.271**
Weighted	Precision	0.671	0.819	**0.823**
	Recall	0.644	0.786	**0.840**
	F1-score	0.657	0.788	**0.827**

Список литературы

[1]. Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo. "Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development". *IEEE Trans. on Knowl. and Data Eng.* 25.4 (Apr. 2013), pp. 919–931.

[2]. Alex Lamb, Michael J Paul, Mark Dredze. "Separating Fact from Fear: Tracking Flu Infections on Twitter." Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Atlanta. June 2013.

[3]. Adam Sadilek, Henry A Kautz, Vincent Silenzio. "Modeling Spread of Disease from Social Interactions." Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012.

[4]. Zhiyuan Cheng, James Caverlee, Kyumin Lee. "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*. CIKM '10. Toronto, ON, Canada: ACM, 2010, pp. 759–768.

[5]. Afshin Rahimi et al. "Exploiting Text and Network Context for Geolocation of Social Media Users". *CoRR* abs/1506.04803 (2015).

[6]. Bryan Perozzi, Steven Skiena. "Exact Age Prediction in Social Networks". *Proceedings of the 24th International Conference on World Wide Web*. WWW '15 Companion. Florence, Italy: ACM, 2015, pp. 91–92.

[7]. Lars Backstrom, Eric Sun, Cameron Marlow. "Find me if you can: improving geographical prediction with social and spatial proximity". *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 61–70.

[8]. David Jurgens. "That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships". *ICWSM* 13 (2013), pp. 273–282.

[9]. Yantao Jia et al. "Location Prediction: A Temporal-Spatial Bayesian Model". *ACM Transactions on Intelligent Systems and Technology (TIST)* 7.3 (2016), p. 31.

[10]. Jinpeng Chen, Yu Liu, Ming Zou. "Home location profiling for users in social media". *Information & Management* 53.1 (2016), pp. 135–143.

[11]. Fan RK Chung. "Spectral graph theory". CBMS regional conference series in mathematics, No. 92 (1996).

[12]. Oleg U Ivanov, Sergey O Bartunov. "Learning Representations in Directed Networks". *International Conference on Analysis of Images, Social Networks and Texts*. Springer. 2015, pp. 196–207.

[13]. Bryan Perozzi, Rami Al-Rfou, Steven Skiena. "Deepwalk: Online learning of social representations". *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2014, pp. 701–710.

[14]. T Mikolov, J Dean. "Distributed representations of words and phrases and their compositionality". *Advances in neural information processing systems* (2013).

[15]. Michael U Gutmann, Aapo Hyvärinen. "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics". *Journal of Machine Learning Research* 13.Feb (2012), pp. 307–361.

[16]. François Chollet. *Keras*. https://github.com/fchollet/keras. 2015.

Approaches to estimate location of social network users based on social graph

[1]*Y.S. Trofimovich <integral@ispras.ru>*
[1]*I.S. Kozlov <kozlov-ilya@ispras.ru>*
[1,2,3]*D.Y. Turdakov <turdakov@ispras.ru>*
[1]*Institute for System Programming of the Russian Academy of Sciences*
Moskva, Aleksandra Solzhenitsyna, 25, 109004, Russia
[2]*Lomonosov Moscow State University,*
GSP-1, Leninskie Gory, Moscow, 119991, Russia
[3]*National Research University Higher School of Economics (HSE)*
20 Myasnitskaya Ulitsa, Moscow, 101000, Russia

Abstract. Many applications require information about the geolocation of users, which is not always available. Among the users of Twitter only about 26% indicate the name of the city in their profiles, about 30% of users of VKontakte leave this field blank. So there is the problem of determining the place of residence of social network users. We investigate approaches to geolocation of social network users using their mutual bidirectional ties — social graph. At first, we present a brief overview of the work in the field of geolocating users of social networks. Then we propose an approach that relies on graph nodes' embeddings and supervised machine learning techniques. Series of experiments were conducted with proposed and baseline approaches. Experiments show that proposed approach is comparable with others. The results of experiments allow us to conclude that the proposed approach based on vector representation can be effectively used to determine the user's place of residence by itself, or in combination with classifiers based on user data It is worth noting that the proposed approach has no any specifics related to the geolocation. It can also be used to assess any other demographic attributes that influence the formation of relationships in society. Thus, a similar approach was used in Perozzi and Skiena to determine the age of the users.

Keywords: geolocation, social networks, social graph, graph embeddings

DOI: 10.15514/ISPRAS-2016-28(6)-13

For citation: Trofimovich Y.S., Kozlov I.S., Turdakov D.Y. Approaches to estimate location of social network users based on social graph. Trudy ISP RAN/Proc. ISP RAS, vol. 28, issue 6, 2016, pp. 185-196 (in Russian). DOI: 10.15514/ISPRAS-2016-28(6)-13

References

[1]. Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo. "Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development". *IEEE Trans. on Knowl. and Data Eng.* 25.4 (Apr. 2013), pp. 919–931.

[2]. Alex Lamb, Michael J Paul, Mark Dredze. "Separating Fact from Fear: Tracking Flu Infections on Twitter." Conference of the North American Chapter of the Association for

Computational Linguistics: Human Language Technologies (NAACL-HLT), Atlanta. June 2013.

[3]. Adam Sadilek, Henry A Kautz, Vincent Silenzio. "Modeling Spread of Disease from Social Interactions." Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012.

[4]. Zhiyuan Cheng, James Caverlee, Kyumin Lee. "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*. CIKM '10. Toronto, ON, Canada: ACM, 2010, pp. 759–768.

[5]. Afshin Rahimi et al. "Exploiting Text and Network Context for Geolocation of Social Media Users". *CoRR* abs/1506.04803 (2015).

[6]. Bryan Perozzi, Steven Skiena. "Exact Age Prediction in Social Networks". *Proceedings of the 24th International Conference on World Wide Web*. WWW '15 Companion. Florence, Italy: ACM, 2015, pp. 91–92.

[7]. Lars Backstrom, Eric Sun, Cameron Marlow. "Find me if you can: improving geographical prediction with social and spatial proximity". *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 61–70.

[8]. David Jurgens. "That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships". *ICWSM* 13 (2013), pp. 273–282.

[9]. Yantao Jia et al. "Location Prediction: A Temporal-Spatial Bayesian Model". *ACM Transactions on Intelligent Systems and Technology (TIST)* 7.3 (2016), p. 31.

[10]. Jinpeng Chen, Yu Liu, Ming Zou. "Home location profiling for users in social media". *Information & Management* 53.1 (2016), pp. 135–143.

[11]. Fan RK Chung. "Spectral graph theory". CBMS regional conference series in mathematics, No. 92 (1996).

[12]. Oleg U Ivanov, Sergey O Bartunov. "Learning Representations in Directed Networks". *International Conference on Analysis of Images, Social Networks and Texts*. Springer. 2015, pp. 196–207.

[13]. Bryan Perozzi, Rami Al-Rfou, Steven Skiena. "Deepwalk: Online learning of social representations". *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2014, pp. 701–710.

[14]. T Mikolov, J Dean. "Distributed representations of words and phrases and their compositionality". *Advances in neural information processing systems* (2013).

[15]. Michael U Gutmann, Aapo Hyvärinen. "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics". *Journal of Machine Learning Research* 13.Feb (2012), pp. 307–361.

[16]. François Chollet. *Keras*. https://github.com/fchollet/keras. 2015.