СТРУКТУРА СООБЩЕСТВ СТРАН, НАБЛЮДАЕМАЯ ЧЕРЕЗ ДАННЫЕ ПЛАТФОРМЫ FLICKR

А. Б. Белый, Л. В. Рудикова, С. Л. Соболевский, А. Н. Курбацкий
Факультет прикладной математики и информатики, Кафедра технологий программирования, Белорусский государственный университет
Кафедра современных технологий программирования, Учреждение образования «Гродненский государственный университет имени Янки Купалы»
Центр городской науки и прогресса, Нью-Йоркский университет
Минск, Гродно, Республика Беларусь; Бруклин, Соединенные Штаты Америки
E-mail: {alexander.belyi, rudikowa}@gmail.com, sobolevsky@nyu.edu, kurb@unibel.by

В статье рассматривается глобальная сеть перемещений людей между странами, построенная с использованием данные о цифровых гео-локализованных фотографиях и видео, размещенных в открытом доступе на интернет-сервисе Flickr. Рассматриваемый набор данных открывает новые возможности для понимания мобильности, проливая свет на краткосрочные поездки из одной страны в другую. В работе демонстрируется, как применение метода поиска сообществ к сети перемещений между странами позволяет выявить интересные пространственные закономерности.

Введение

Всё чаще, путешествуя из одной страны в другую, люди оставляют за собой цифровой след в различного рода сервисах. Соответствующие данные открывают огромные возможности для исследований, с их помощью мы можем восстановить перемещения людей, проанализировать их и, возможно, обнаружить интересные и важные закономерности. Рассматривая мобильность в глобальном масштабе, крайне важно учитывать различные аспекты человеческих перемещений, которые состоят из разных видов мобильности и включают как релокацию на постоянное место жительства, так и краткосрочные посещения. А анализ данных из различных источников позволяет отдельно рассматривать международные миграции и краткосрочные путешествия.

Предыдущие исследования миграционных данных и данных сервиса Twitter показали, что определение сообществ в сетях мобильности обычно приводит к географически связным сообществам (даже при том, что никакие пространственные характеристики методом определения сообществ не учитываются), выявляя важные географические закономерности. В своей работе мы применили похожий подход к сети основанной на данных сервиса Flickr. Сеть перемещений между странами, полученная из этих данных, в основном представляет краткосрочные перемещения, т.к. в большинстве случаев данные Flickr отражают туристическую активность.

I. Набор данных

Используемый набор данных Flickr, состоит из 130 миллионов фотографий и видео-файлов. Он был собран из двух наборов, размещенных в открытом доступе [1–2]. Набор содержит данные за десять лет: с 2005 по 2014 год. Используя эти данные, мы построили ориентированную взвешенную сеть краткосрочных перемещений, в которой вершины соответствуют странам, а веса рёбер равны количеству пользователей из одной страны, посетивших другую страну. Для этого сначала мы определили постоянную страну пребывания пользователей (когда это было возможно). Затем, если пользователь имел фотографии или видео, сделанные в других странах, мы трактовали это как посещение им данной страны. Для определения страны постоянного пребывания мы использовали один из наиболее консервативных методов, используемых в подобных исследованиях [3]: мы считали такой страной ту, в которой пользователь сделал наибольшее количество фотографий (не менее 10) и провел наибольшее количество времени (не менее полугода). Используя этот критерий, мы смогли определить страну постоянного пребывания для более чем 500 тысяч пользователей, которые сделали более 80% всех фотографий и видео, т. е. более 90 миллионов. Мы также исключили из рассмотрения страны, соответствующие вершинам с входящей или исходящей силой менее 10. В результате мы получили сеть из 201 страны.

II. Определение сообществ

Для определения сообществ мы использовали один из наиболее популярных и хорошо установившихся подходов к разбиению сетей, основанный на максимизации функции модулярности. Но, т.к. в рассматриваемой нами сети отсутствуют петли, нам пришлось поправить классическую функцию модулярности. В частности, мы изменили способ, которым нуль-модель, используемая в модулярности, оценивает ожидаемый вес каждого ребра. В своей классической форме модулярность использует $\frac{s_i t_j}{m}$ как ожидаемый вес ребра из начальной вершины i в конечную вершину j, где s_i и t_j – входящая и исходящая сила вершин i и j, соответственно, и m – суммарный вес всех рёбер, т. е. $s_i = \sum_j w_{ij}$, $t_j = \sum_i w_{ij}$ и $m = \sum_k s_k = \sum_k t_k = \sum_{ij} w_{ij}$,

в то время как w_{ij} – набдюдаемый вес ребра из i в j. Однако, если петли не участвуют в этом распределении, то ожидаемый вес скорее должен быть $\frac{s_i t_j}{\sum_{k \neq i} t_k}$ или $\frac{s_i t_j}{\sum_{k \neq j} s_k}$, в зависимости от того, рассматривать его как распределение исходящих сил s_i между всеми конечными вершинами, за исключением самой i, или как распределение входящих сил t_j между всеми начальными вершинами, за исключением самой j. В качестве окончательного значения мы использовали среднее этих двух значений, что привело к выражению $\frac{1}{2}\left(\frac{s_i t_j}{m - t_i} + \frac{s_i t_j}{m - s_j}\right)$.

Поскольку известно, что модулярность имеет определенные недостатки, такие как, например, предел разрешающей способности, не позволяющий определять достаточно мелкие сообщества, мы также использовали подход предложенный ранее в [4], который вводит так называемый параметр разрешающей способности. Таким образом, окончательная формула для скорректированной модулярности, используемая в нашем случае для сети без петель выглядит следующим образом:

$$\frac{1}{2m} \sum_{i \neq j} \left(2w_{ij} - a\frac{s_i t_j}{m - t_i} - a\frac{s_i t_j}{m - s_j}\right) \delta(C_i, C_j),$$

где a обозначает параметр разрешающей способности, i, j – вершины, C_i, C_j – сообщества, которым они принадлежат, $\delta(x, y) = 1$, если $x = y$, иначе 0. Наконец, для нахождения наилучшего разбиения, мы оптимизировали эту версию модулярности, используя точный и эффективный алгоритм Combo [5], способный максимизировать различные типы целевых функций.

III. Структура сообществ сети перемещений между странами

Максимизация модулярности при значении разрешающего параметра 1.0 приводит к выявлению *пяти* сообществ, в то время как для значения 2.0, число найденных сообществ возрастает до *семнадцати*, делая рассмотрение больших значений затруднительным, поскольку становится сложно различить визуально и анализировать разные сообщества на карте. На Рис. 1 приведены разбиения для значений разрешающего параметра 1.0 и 2.0; страны, попавшие в одно сообщество, окрашены одним цветом.

Из рисунка видно, что основные географические регионы, такие как Северная и Южная Америки, Восточная Азия, страны СНГ, объединены в отдельные сообщества. Для разных значений параметра можно наблюдать интересные особенности. Например, Египет и Турция попадают в одно сообщество со странами СНГ, что может быть объяснено их популярностью среди туристов из СНГ. Если рассматривать структуру сообществ большей гранулярности, то можно заметить сильную связь Ирака и Афганистана со странами Северной Америки, также как европейских стран с их бывшими африканскими колониями. В то же время, только Ирландия попадает в одно сообщество с Великобританией, некогда могущественным доминионом с колониями по всему миру. Полученные кластеры указывают на то, что, хотя люди чаще путешествуют в близлежащие страны, общий язык и история играют важную роль в выборе страны назначения очередного путешествия.

Также интересным результатом является то, что структура сообществ в сети мобильности, построенной по данным сервиса Flickr, обладает свойством, присущим большинству других изученных сетей мобильности: сообщества географически связаны и отражают устоявшиеся регионы.

1. SFGEO.ORG [Electronic resource] – Mode of access: http://sfgeo.org/data/tourist-local. – Date of access: 02.09.2016.
2. Thomee, B. YFCC100M: The New Data in Multimedia Research / B. Thomee [et al.] // Communications of the ACM. – 2016. – Vol. 59, № 2. – P. 64–73.
3. Bojic, I. Choosing the Right Home Location Definition Method for the Given Dataset / I. Bojic [et al.] // Social Informatics. – Springer International Publishing, 2015. – P. 194–208.
4. Arenas, A. Analysis of the structure of complex networks at different resolution levels / A. Arenas, V. Fernandez, S. Gomez // New Journal of Physics. – 2008. – Vol. 10 – P. 053039.
5. Sobolevsky, S. General optimization technique for high-quality community detection in complex networks / S. Sobolevsky [et al.] // Physical Review E. – 2014. – Vol. 90, № 1. – P. 012811.

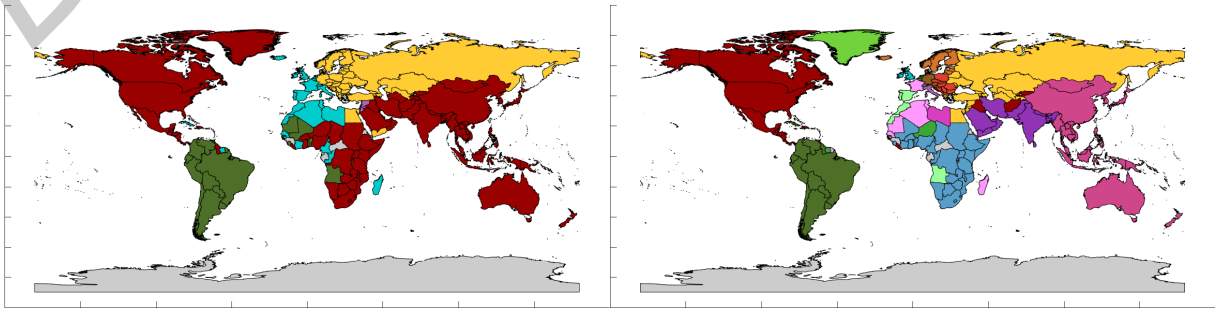

Рис. 1 – Разбиение стран на сообщества разной гранулярности.