АНАЛИЗ ХАРАКТЕРИСТИК СОЦИАЛЬНЫХ ГРАФОВ, ПОСТРОЕННЫХ ПО ДАННЫМ СОЦИАЛЬНОЙ СЕТИ TWITTER

М.Э. Грачева, Е.В. Якоби, В.В. Степаненко, Е.Е. Лунева
Научный руководитель: Е.Е. Лунева
Томский политехнический университет
gramar_98.98@mail.ru

Введение

Социальные сети, получившие широкое распространение в наше время, являются ресурсом, использующимся для выражения мнения относительно различных тем, событий, фактов, продуктов и т.п., а также часто содержащий социально-демографические данные многих своих пользователей в открытом доступе. Изучение таких данных в контексте определенных тем является одним из способов анализа тенденций изменения общественного мнения в широком спектре вопросов, а результаты анализа могут быть использованы в различных областях для решения задач практической направленности, включая задачи антитеррористической направленности, прогнозирование потребности, политические прогнозы, маркетинговые исследования, оценка репутационных рисков компании или физического лица.

Первостепенной задачей такого анализа является выявление характеристик реальных графов, т.е. графов, построенных на основе данных, взятых из социальной сети (например, Twitter) за некоторый промежуток времени в заранее определенной предметной области. Эти характеристики могут быть использованы для генерации моделей случайных графов, использование которых полезно на этапах экспериментального анализа при оценке эффективности математического и программного обеспечения при решении задачи идентификации пользователей-экспертов в социальных сетях в заданной предметной области.

Таким образом, цель данной работы: анализ выборки реальных графов, описывающих данные в рамках одной (необязательно одинаковой) предметной области из социальной сети Twitter за неделю.

Описание модели

При исследовании данных из социальных сетей было принято решение придерживаться следующих принципов: вершинами графа являются пользователи социальной сети Twitter. Вершины соединяются ребрами в четырех случаях:

1. Репост с комментарием
2. Комментарий
3. Репост
4. Упоминание другого пользователя в твите

Более подробно методика построения социального графа описана в [1].

При выполнении настоящей работы вес ребер не учитывался, т.к. на исследуемые характеристики данный параметр никак не влияет.

Граф также является ориентированным: входящие в вершину ребра соответствуют активности этого пользователя по отношению к другому, а значит, чем популярнее пользователь, тем больше ребер будет исходить из соответствующей ему вершины.

Анализ реальных графов

Для данного исследования взяты выборки реальных данных из социальной сети Twitter по десяти предметным областям, в соответствии с которыми построены графы (количество вершин варьируется в диапазоне от 200 до 902). Для каждого из графов высчитан ряд характеристик и определены их средние значения.

Особый интерес среди всех характеристик представляют такие, как распределение вершин по количеству входящих и исходящих ребер, коэффициент кластеризации и диаметр графа. [2]

1. Распределение вершин по количеству исходящих ребер позволяет выявить самых популярных пользователей в данной выборке. Можно предположить, что мнение таких пользователей имеет большее влияние на общественные массы, чем мнение «непопулярных» пользователей.
2. Распределение вершин по количеству входящих рёбер выявляет наиболее активных пользователей сети.
3. Диаметр графа показывает максимально возможное расстояние между двумя его вершинами.
4. Коэффициент кластеризации позволяет оценить, насколько плотно сгруппирован граф вокруг нескольких вершин. Другими словами, в предметной области, которой соответствует граф с высоким коэффициентом кластеризации, «управляют» общественным мнением несколько популярных пользователей.

В таблице 1 представлены параметры: коэффициенты кластеризации и диаметр сети для исследуемых графов.

Следует отметить, что диапазон значений для коэффициента кластеризации достаточно широкий, однако все значения находятся в пределах 0,1. Низкий показатель этой характеристики объясняется следующими причинами: во-первых, максимальная глубина выборки данных для каждого графа ограничивается семи днями, что не даёт возможности в полной мере оценить реальное количество взаимодействий между пользователями, во-вторых, в каждой выборке присутствует большое количество вершин, связанных попарно, либо не

имеющих связей ни с одной из других вершин, что сильно влияет на коэффициент кластеризации, понижая его.

Таблица 1. Значения коэффициента кластеризации и диаметра сети реальных графов

№	Коэффициент кластеризации	Диаметр сети
1	0,083	4
2	0,011	3
3	0,021	2
4	0,002	4
5	0,047	2
6	0	2
7	0,104	5
8	0,022	2
9	0,006	2
10	0,082	3
Диапазон	0 – 0,104	2 - 5
Ср. знач.	0,0378	2,9

Распределение вершин по количеству входящих рёбер показало, что подавляющее большинство пользователей проявили активность в данной предметной области по отношению только к одному пользователю (рис. 1). Также из графика видно, что велико количество и неактивных пользователей. Можно предположить, что часть из вершин, соответствующая данным пользователям, изолирована, а также в это количество могут входить популярные пользователи.

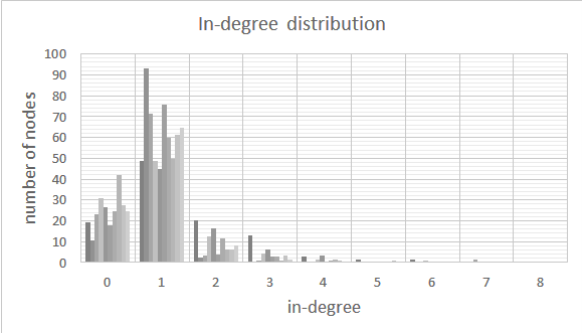

Рис. 1. Распределение вершин по количеству входящих рёбер

Несмотря на невысокий коэффициент кластеризации можно заметить, что распределение по количеству выходящих ребер убывает по экспоненте, что подтверждает предположение о том, что существует небольшое количество вершин, соответствующих популярным пользователям, в отношении которых люди проявляют наибольшую активность.

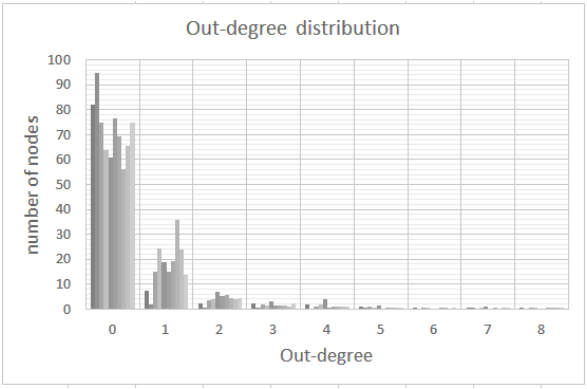

Рис 2. Распределение вершин по количеству исходящих рёбер

Заключение

В результате исследования характеристик графов, построенных на основе данных, полученных из социальной сети Twitter, выявлены следующие особенности, характерные для каждого из построенных графов:

1. Низкий коэффициент кластеризации (не более 0,1)
2. Относительно небольшой диаметр (2,9 – среднее значение)
3. Распределение вершин по количеству исходящих рёбер убывает по экспоненте (3 и более исходящих ребер имеют около 5% вершин).
4. Распределение вершин по количеству входящих рёбер показало, что количество активных пользователей очень мало, однако велико количество пользователей, не проявляющих активность вообще.

В дальнейшем, данные свойства могут быть использованы для построения моделей усреднённых графов.

Работа выполнена при финансовой поддержке РФФИ (проект №17-07-00034 А).

Список использованных источников

1. Luneva E.E., Zamyatina V.S., Banokin P.I., Yefremov A.A. Estimation of social network user's influence in a given area of expertise // Journal of Physics: Conference Series. – 2017. Vol. 803, № 1. – С.1-6.
2. Гусарова Н.Ф. Анализ социальных сетей. Основные понятия и метрики. – СПб: Университет ИТМО, 2016. – 67 с.