ВИЗУАЛЬНОЕ ПРЕДСТАВЛЕНИЕ И КЛАСТЕРНЫЙ АНАЛИЗ СОЦИАЛЬНЫХ СЕТЕЙ

М.И. Хотилин, А.В.Благов

Самарский государственный аэрокосмический университет имени академика С.П. Королёва (национальный исследовательский университет) (СГАУ), Самара, Россия

Статья посвящена анализу социальных сетей, представленных графом. В работе приводятся подходы к моделированию распределений социальных сетей, а также алгоритмы, используемые для отыскания сообществ, а также аккаунтов, оказывающих наибольшее влияние на сообщества.

Ключевые слова: bigdata, графы, визуализация данных, анализ данных, кластеризация, modularity, SCAN.

Введение

В современном мире непрерывно генерируется огромное количество данных, будь-то данные поступающие от спутника, либо датчиков в самолете, банковских транзакций, диагностические данные пациентов и т.д. Особое место занимают социальные сети, генерируемый объем данных растет с каждым днём. Значимость социальных сетей обусловлена тем, что , с одной стороны они являются предметом социализации людей, а с другой – наиболее мощным и доступным политическим, идеологическим и экономическим инструментом [1]. Исследованию социальных сетей как систем, содержащих данные сверхбольшого объёма, посвящен ряд работ в этой области [2, 3].

Большие объёмы данных, а также зависимости (связи) между ними необходимо представить в виде удобном для восприятия. Данные социальных сетей могут быть представлены в различных видах: облако тегов, диаграммы, исторические потоки [4], однако чаще всего для этой цели используют графы.

1. Представление сети в виде графа

В основном, когда речь идет об объектах представляющих собой сеть, например социальную, понятие визуализации данных тесно связано с понятием графов. Важной задачей является представление связей в социальных сетях для выявления различного рода зависимостей.

Граф представляет собой совокупность непустого множества вершин и множества ребер: $G(X,U)$ (X -множество вершин, U -множество ребер). Вершинами в графе, описывающем социальную сеть, являются аккаунты пользователей, а ребрами – связи между ними, например, подписка в сетях типа twitteri отношение типа «дружба» в социальных сетях типа Facebook (рисунок 1).

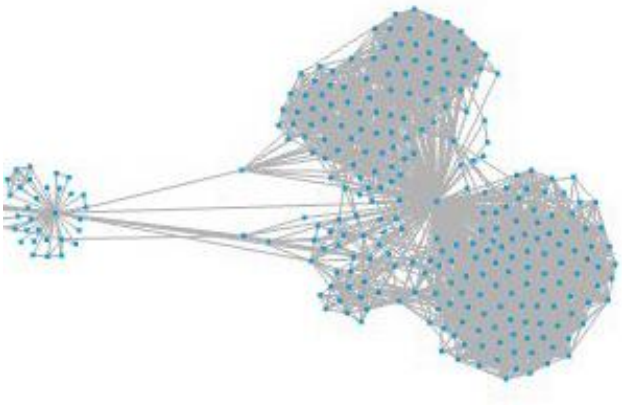

Рис. 1. Фрагмент графа социальной сети «ВКонтакте»

Одной из важнейших связанных характеристик, которую следует рассмотреть, является метрика. Метрика графа основана на понятии расстояния.

Например, назовем расстоянием $d(x_i, x_j) = d_{ij}$ между вершинами x_i и x_j графа $G(X,U)$ длину кратчайшей цепи, соединяющей эти вершины. Под длиной цепи понимается число входящих в нее ребер. Тогда функция $d(x_i, x_j)$, определенная на множестве ребер U графа G, называется метрикой графа.

Степенью вершины $x_i \in X$ графа является количество ребер инцидентных данной вершине - $d(x_i)$.

Опытным путем было доказано, что степенное распределение различных сегментов подавляющего большинства в социальных сетей имеет следующий вид (рисунок 2). f_k - это доля вершин графа $G(X,U)$, имеющих степень $d(x_i) = k$.

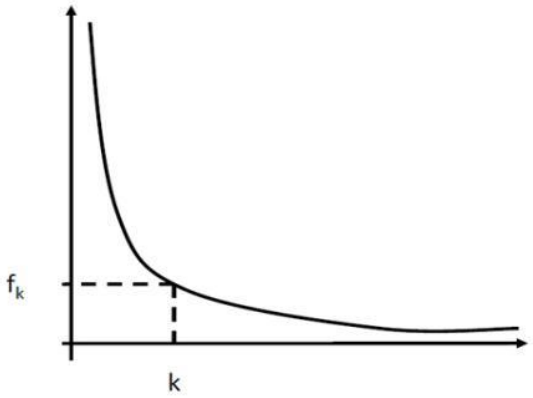

Рис. 2. Распределение степени вершин графа

Для моделирования данного распределения как правило подходят функции вида[5]:

$$p(k) = Ck^{-\alpha},$$

$$p(k) = \frac{z^k}{k!}e^{-z}.$$

Коэффициенты α, C и z находятся того или иного сегмента социальной сети.

2. Кластеризация и отыскание сообществ, алгоритмы основанные на modularity

В целях упрощения графа, а также для отыскания так называемых «сообществ» в социальной сети, описанной графом применяется кластеризация.
Существует ряд алгоритмов и подходов для обеспечения кластеризации. Одним из которых является modularity[6-7].
Данный функционал был предложен Ньюманом и Гирваном в процессе разработки алгоритма кластеризации вершин графа [6]. Под модулярностью подразумевается скалярная величина из отрезка [-1, 1], выражаемая следующей формулой:

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{d_i d_j}{2m}\right)\delta(C_i, C_j),$$

где A - матрица смежности графа, A_{ij}- (i,j) элемент матрицы A, d_i- степень i вершины графа, C_i - метка вершины (номер сообщества, к которому относится вершина), $m -$ общее количество ребер в графе, $\delta(C_i, C_j)$- дельта-функция (единица, если $C_i = C_j$, ноль иначе).
Задача поиска выделения сообществ в графе сводится к поиску такихC_i, которые будут максимизировать значение модулярности.

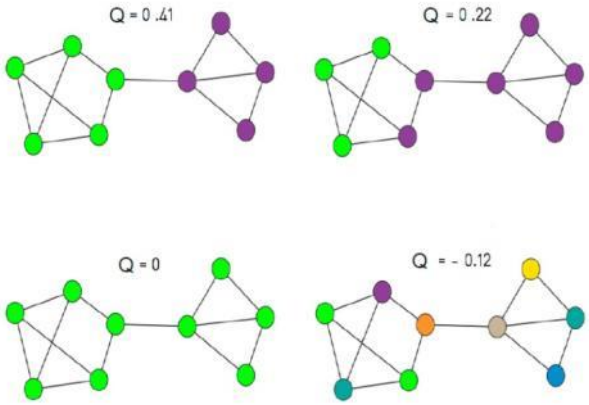

Рис. 3. Значение коэффициента модулярности при выделении различных кластеров

К достоинствам модулярности можно отнести следующее:

- модулярность достаточно просто интерпретируется. Ее значение равно разности между долей ребер внутри сообщества и ожидаемой доли связей, если бы ребра были размещены случайно;
- модулярность возможно эффективно пересчитывать при небольших изменениях в кластерах.

Однако имеются также и недостатки:

- функционал не является непрерывным, и задача его оптимизации — дискретная. Для поиска глобального оптимума используют приближенные схемы. Некоторые из них действительно оптимизируют значение функционала, другие же по значению модулярности выбирают наилучшее решение из найденных, то есть без гарантий локальной оптимальности решения;
- существует проблема разрешающей способности (грубо говоря, функционал не видит маленькие сообщества). Эта проблема решается путем использования модифициро-

ванного функционала, который сохраняет все достоинства и добавляет параметр масштаба [7].

Поскольку модулярность описывает качество разделения графа на группы, то к решению задачи отыскания оптимального разбиения графа можно подойти, решая задачу максимизации модулярности. Однако простым перебором решить эту задачу практически невозможно, так как число вариантов разделения n узлов на k групп растет, экспоненциально с ростом n. Для решения поставленной задачи был предложен жадный алгоритм оптимизации функции модулярности, имеющий в своем основании пошаговое объединение двух групп, дающих наибольший прирост модулярности.

Рассмотрим некое разбиение узлов из N на k групп (N – множество узлов с числом элементов n) [8]. Функция модулярности будет равна:

$$Q_1 = \frac{1}{m}\sum_{l=1,l\neq i,l\neq j}^{k}\left(m_l - \frac{(d(N_l))^2}{4m}\right) + \frac{1}{m}\left(m_i + m_j - \frac{(d(N_i))^2 + (d(N_j))^2}{4m}\right).$$

Теперь объединим группы i и j в одну, которую обозначим как $N_{i\cup j} = N_i \cup N_j$. Функция модулярности для нового графа будет иметь следующий вид:

$$Q_2 = \frac{1}{m}\sum_{l=1,l\neq i,l\neq j}^{k}\left(m_l - \frac{(d(N_l))^2}{4m}\right) + \frac{1}{m}\left(m_{i\cup j} - \frac{(d(N_{i\cup j}))^2}{4m}\right).$$

Число дуг внутри группы $N_{i\cup j}$равно сумме дуг внутри групп N_iи N_jплюс число дуг между ними. Иными словами:

$$m_{i\cup j} = m_i + m_j + m_{i,j}.$$

Степень объединённой группы $N_{i\cup j}$равна сумме степеней групп N_iи N_j, то есть:

$$d(N_{i\cup j}) = d(N_i) + d(N_j),$$

Следовательно

$$(d(N_{i\cup j}))^2 = (d(N_i))^2 + (d(N_j))^2 + 2d(N_i)d(N_j).$$

Учитывая это, получаем:

$$\Delta Q = Q_2 - Q_1 = \frac{1}{m}\left(m_{i,j} - \frac{2d(N_i)d(N_j)}{4m}\right) = \frac{1}{m}\left(m_{i,j} - \frac{d(N_i)d(N_j)}{2m}\right).$$

Отсюда получаем, что наибольший рост модулярности происходит при объединении таких группN_iи N_j, для которых величина

$$\Delta(N_i, N_j) = m_{i,j} - \frac{d(N_i)d(N_j)}{2m}$$

максимальна.

Также видно, что объединение групп, между которыми нет дуг ($m_{i,j}$=0), не может дать увеличения модулярности.

3. Классификация вершин, нахождение наиболее значимых,SCAN алгоритм

Помимо отыскания сообществ в социальной сети представленной графом значительный интерес представляет классификация вершин в графе.

В статье [7] была представлена следующая классификация вершин (см. рисунок):

— core(ядро) – это вершина, содержащая в ε − окрестности, по крайней мере μ вершин

– hub (хаб) – это отдельная вершина, соседи которой принадлежат двум или более различным кластерам;
– outlier (посторонний) – это отдельная вершина, все соседи которой принадлежат одному и тому же кластеру, или не принадлежат никакому кластеру.

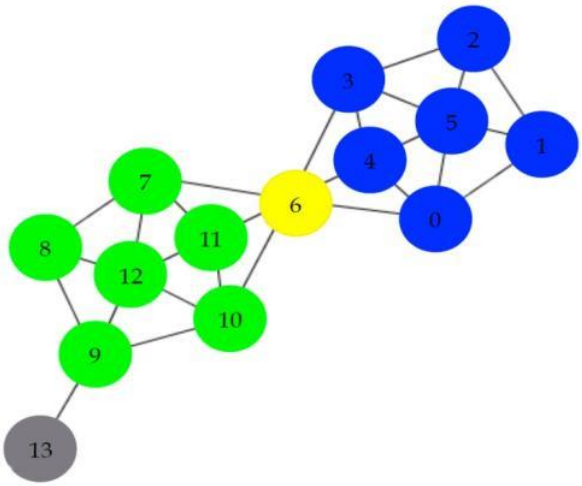

Рис. 4. Пример графа с различными типами вершин

Для осуществления подобной классификации предлагается SCAN алгоритм [7].Принцип его работы описывается следующим образом.

Поиск начинается с начального посещения каждой вершины один раз, с целью нахождения структурно-связных кластеров, а затем посещения изолированных вершин, чтобы идентифицировать их (hubили outlier).

SCAN выполняет один проход сети и находит все структурно-связанные кластеры для заданного параметра. В начале все вершины помечены как неклассифицированные. Алгоритм SCAN классифицирует каждую вершину либо как являющуюся членом кластера, либо как не являющуюся. Для каждой вершины, которая еще не классифицирована, SCAN проверяет, является ли эта вершина ядром. Если вершина является ядром, новый кластер расширяется из этой вершины. В противном случае вершина помечается как не являющаяся членом кластера.

Чтобы найти новый кластер, SCAN начинается с произвольной ядраV и ищет все вершины, которые структурно-достижимы из V. Этого вполне достаточно, чтобы найти полный кластер, содержащий вершину V. Генерируется новый ID кластера, который будет назначен всем найденным вершинам.

SCAN начинается, постановкой всех вершин в ε-окрестности вершины V в очередь. Для каждой вершины в очереди вычисляются все непосредственно достижимые вершины, и в очередь вставляются те вершины, которые до сих пор не классифицированы. Это повторяется до тех пор, пока очередь не опустеет.

Вершины, не являющиеся членами кластеров, могут быть дополнительно классифицированы как хабы или посторонние. Если отдельная вершина имеет ребра на два или более кластеров, она может быть классифицирована как хаб. В противном случае, это посторонний.

Отличительной особенностью является наличие параметров μ и ε, которые могут задаваться пользователем или экспертом. При этом нахождение оптимального значения дан-

ных параметров можно провести при помощи машинного обучения системы, используя определённые сегменты сети.

Заключение

Представление социальных сетей в виде графа и дальнейший его анализ, включающий кластеризацию и отыскание зависимостей, является актуальной задачей в области BigData. Использование описанных в статье методов и подходов позволяет производить классификацию сегментов социальной сети, а также находить элементы, представляющие наибольший интерес, например, пользователей, влияющих на несколько отдельных сообществ (в графовом представлении – вершины типа “hub”). При нахождении степенного распределения вершин графа, описывающего социальную сеть, можно осуществлять моделирование социальных сетей с заданным распределением.

Представленные в данной статье алгоритмы планируются к доработке и использованию в исследовании сегментов социальных Самарского региона. Авторами разрабатывается необходимый инструментарий для графовой визуализации необходимых сегментов социальных сетей, а также распределенные методы обработки графов высокой размерности.

Литература

1. Tan, W. Social-network-sourced big data analytics/ W. Tan, M.W. Blake, I. Saleh., S. Dustdar //IEEE Internet Computing. – 2013. – №. 5. – P. 62-69.
2. 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. 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.
4. Иванов, П.Д.Технологии BigData и различные методы представления больших данных / П.Д. Иванов, А.Г. Лопуховский// Инженерный журнал: наука и инновации, вып. 9. – 2014.
5. Gastner, M. T. Optimal design of spatial distribution networks / Michael T Gastner, M. E. J. Newman // Phys. Rev. E. – 2006. – T.74. – C. 016117.
6. Newman, M. E. J. Finding and evaluating community structure in networks / M. E. J. Newman, M. Girvan // Phys. Rev. E. – 2004. – T. 69. – C. 026113.
7. Scan: a structural clustering algorithm for networks / Xu X. et al. //Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2007. – P. 824-833.
8. Newman, M. E. J.Fast algorithm for detecting community structure in networks / M. E. J. Newman // Phys. Rev. E. – 2004. – T. 69. – C.066133.