DOI: 10.46548/CIT-2021-0034-0002

А.И. Мартышкин

СОСТОЯНИЕ ИССЛЕДОВАНИЙ И ОСНОВНЫЕ ТЕНДЕНЦИИ РАЗВИТИЯ СОЦИАЛЬНЫХ СЕТЕЙ

Россия, г. Пенза, Пензенский государственный технологический университет

The article deals with the analysis of the state of research and the main trends in the development of social networks. Social networks and mass media have been developing rapidly in the last decade: new forms of information interaction on the Internet are emerging. Today, the development of the Internet has more closely intertwined people and information: users simultaneously play the role of producer and consumer of information. The structure of the network becomes more complicated, the number of its nodes increases, and the data carrying useful information becomes more. These massive social data have great value. Due to the increase in the number of users of social networks on the Internet, huge streams of user data containing a variety of information are generated daily, so the analysis of a large-scale network is the focus of research conducted today.

According to the support of sociological theory and data analysis, "influence" is the main driving force behind the dissemination of information in social networks. Modeling social influence is one of the main studies of social networks, and maximizing influence is the key to solving problems related to finding users on social networks and maximizing the transmission of certain information on a social network.

Введение. Социальные сети и средства массовой информации в последнее десятилетие быстро развиваются: постоянно появляются новые формы информационного взаимодействия в Интернете. Сегодня развитие Интернета более тесно сплело людей и информацию: пользователи одновременно играют роль производителя и потребителя информации. Структура сети усложняется, возрастает количество ее узлов, а информационных данных становится больше. Эти массивные социальные данные имеют большую ценность [1]. Из-за увеличения числа пользователей социальных сетей в Интернете ежедневно генерируется большое количество пользовательских данных, содержащих много информации, поэтому анализ масштабной сети находится центре внимания проводимых исследований. Постепенно акцент исследований перешел на анализ социальных сетей, что стало важной темой для многих ученых и экспертов; с момента первого введения проблемы максимизации влияния социальных сетей в компьютерную область в 2001 году, она стала актуальной темой проводимых исследований. Проблема максимизации влияния социальных сетей заключается в том, чтобы найти пользователей с наибольшим влиянием в сети, и при определенной модели распространения их влияние будет продолжать распространяться в социальной сети. Целью задачи максимизации влияния социальной сети является получение максимального покрытия воздействия в кратчайшие сроки и как можно меньшее количество начальных узлов.

Исследования в рамках предметной области в последние годы проводились по следующим направлениям: Домимгос, Ричардсон и соавторы (Domimgos, Richardson et al.) [2] впервые предложили проблему максимизации влияния. Они смоделировали задачу, представив ее как случайный Марковский процесс, и использовали эвристический алгоритм для решения задачи. Кемпе и соавторы (Kempe et al.) [3] изучили проблему максимизации влияния как дискретную задачу оптимизации, впервые ввели в задачу максимизации влияния социальных сетей модель независимого

каскада (НК) и модель линейного порога (ЛП). Они доказали, что функция распространения влияния имеет субмодульность и монотонность в этих двух моделях распространения и что задача максимизации влияния социальных сетей является NP-жесткой задачей в этих моделях. Также ими предлагается «жадный алгоритм» решения этой задачи, при помощи которого можно получить приближенное оптимальное решение (1-1/e). Публикация этой работы заложила основу для будущих исследований ученых по максимизации влияния социальных сетей.

Исследования по максимизации влияния социальных сетей в основном сосредоточены на двух аспектах, один из которых – улучшение модели, другой – улучшение алгоритма. С точки зрения моделей, это в основном улучшения и расширения на основе моделей НК и моделей ЛП. В терминах алгоритмов Кемпе и соавторы (Kempe et al.) [3] формализовали задачу максимизации влияния социальных сетей как дискретную задачу оптимизации, доказав, что задача максимизации влияния социальной сети является NP-жесткой задачей в модели НК и модели ЛП; функция распространения влияния в этих двух моделях имеет монотонность и субмодульность, поэтому они могут быть решены с помощью «жадного алгоритма», и может быть получено приблизительное оптимальное решение (1-1/e). В жадном алгоритме Кемпе и соавторы использовали метод Монте-Карло для исследования модели распространения. Благодаря множеству циклов повторений (20000 раз), чтобы точно оценить диапазон распространения влияния, алгоритму потребовалось много времени для реализации. Чтобы решить эту проблему, некоторые ученые оптимизировали «жадный алгоритм». В 2007 году Лесковец и соавторы (Leskovec et al.) [4] предложили оптимизацию CELF (Cost-Effective Lazy Forward). Этот метод оптимизации был предложен с использованием субмодульной функции распространения влияния, то есть убывающего усиления. Принцип оптимизации CELF таков: если предельное усиление в предыдущем раунде меньше предельного усиления в текущем раунде, оно больше не будет рассчитываться в следующем раунде в соответствии с декрементом усиления. Этот метод уменьшает количество вычислений и достигает значительного эффекта оптимизации. Эффективность работы в 700 раз выше, чем в известных ранее алгоритмах. Кимура и соавторы (Kimura et al.) [5] предложили «жадный алгоритм», основанный на теории графов на основе модели ЛП и модели НК. Время работы этого алгоритма короче, чем у оптимизации CELF. Позднее Гоял и соавторы (Goyal et al.) [6] предложили алгоритм CELF++, основанный на оптимизации CELF. Этот алгоритм уменьшает повторяющееся вычисление предельного выигрыша путем добавления некоторых маркеров. Его время работы на 35%-55% быстрее, чем оптимизация CELF. Однако многие современные методы являются слишком медленными для сетей миллиардного масштаба [7, 8], и специальные эвристики не могут гарантировать производительность [9, 10].

После этого появились новые методы с гарантией производительности IM, TIM/TIM+ [11] и последний IMM [12], они приняли новый метод выборки, предложенный Боргсом и др. [13], названный RIS. Все предыдущие методы надеются иметь наименьшую выборку RIS и достичь $(1 - 1 / e - \varepsilon)$ гарантии приблизительного решения. Они используют очень сложные методы оценки, чтобы приблизить количество образцов RIS к некоторому теоретическому порогу θ [11, 12]. Их методы действительно эффективно сокращают время расчета, но все они имеют два недостатка: 1) нестабильный размер образцов и 2) порог не является минимальным.

Чтобы устранить эти недостатки, Нгуен и соавторы (Nguyen et al.) [14] предложили два новых алгоритма: SSA и DSSA. Они заявили, что могут оптимизировать количество выборок, обеспечивая при этом ту же точность, что и IMM.

Социальная сеть относится к относительно стабильной системе отношений, сформированной в результате взаимодействия между отдельными людьми, например, YouTube, Facebook, Twitter и т.д., которые распространены в повседневной жизни (рисунок 1).

Рисунок 1 – Тип социальных сетей

Как основной источник распространения информации социальная сеть здесь относится не только к узкому определению веб-сайта онлайн-социальной сети, но и к общему термину для формирования сети информационного взаимодействия между людьми, которая может включать различные сервисы, такие как электронная почта, мобильные сети и т.д. Эти службы могут быть равномерно представлены графической моделью и обладать некоторыми особыми сетевыми характеристиками.

Исследования сети следует проследить с самых ранних лет. Социальная сеть аналогична определению "сети", упомянутому выше. "Социальная сеть" – это совокупность отношений между несколькими точками социальных участников и связанных участников между точками. Точки в социальной сети – участники, которыми могут быть отдельные лица, компании или коллективные социальные единицы, или школы, колледжи, организации, города, страны и т.д.

Отношения между участниками разнообразны: это могут быть дружеские отношения, торговые отношения между странами, а также отношения сотрудничества между отдельными людьми, интерактивные отношения и т.д.

Социальные сети связаны с взаимодействием и связью между людьми, а социальное взаимодействие влияет на социальное поведение людей:

1. Формирование социальной сети обусловлено различными факторами, такими как географические отношения, кровное родство, академические и карьерные отношения. Многие сети формируются в жизни естественным образом. Например, формирование сети соседей и сети родного города обусловлено географическими отношениями, сеть выпускников – академические отношения, а сеть коллег – карьерные отношения. В связи с широким использованием и распространением Интернета в будущем все больше людей будут иметь свои профили в социальных сетях.

2. Социальные сети отражают сущность индивидуальных и социальных отношений. Социальная сеть – отражение межличностного взаимодействия. С одной стороны, она подчиняется правилам и ограничениям социальных отношений. С другой стороны, это также формирует широкую, косвенную и более сложную основу для личных и социальных отношений. Чтобы понять суть отношений между индивидом и обществом, необходимо проанализировать социальную сеть, к которой принадлежит индивид, и социальное взаимодействие между ними.

3. Социальные сети формируются посредством социальных взаимодействий между отдельными людьми. Использование различных интерактивных средств массовой информации и символов для общения является необходимым условием для формирования социальной сети. Если это одностороннее действие, оно не может представлять собой социальную сеть.

4. Социальные сети эффективны для отдельных людей. Люди могут получать необходимую им информацию из социальной сети, к которой они принадлежат, получать эмоциональную поддержку, удовлетворять различные потребности и обогащать свою жизнь.

5. Социальная сеть относительно стабильна. Различные социальные сети могут быть сильно связаны или слабо связаны, но, вообще говоря, как только социальная сеть сформирована, она обладает относительной стабильностью.

Социальная сеть представляет собой графовую структуру. Каждый узел в графе представляет индивидуум, а ребро может представлять отношения между индивидами. Информационное взаимодействие между индивидами формирует социальную сеть и абстрактное описание сети осуществляется в виде графической модели. Пользователи – это узлы v, а социальные связи между пользователями – границы между узлами e = (v,u); V представляет собой множество всех точек в сети, V ={v1,v2, ..., vn}; E представляет собой множество всех рёбер, E = {(v, u)|v, u ∈ V}; социальная сеть – структура узлов и ребер. Граф представляется как G, G = (V, E).

Основными элементами сети являются узлы и ребра. Статистические свойства узлов и ребер, такие как степень, расстояние, посредник и прочность соединения, будут показаны ниже. Исследования процесса коммуникации и других аспектов неотделимы от понимания основных свойств, поэтому понимание этих статистических свойств имеет большое значение для изучения социальных сетей.

Ранжирование – весьма распространенное требование в обществе. Во многих случаях необходимо указать, какие узлы в сети являются более важными, чтобы предоставить
дополнительные рекомендации или информацию о принятии решений. Далее приведены некоторые классические показатели важности или центральной роли узлов в социальных сетях. Эти показатели могут быть объединены с проблемой максимизации влияния и наиболее важным элементом при выборе начального узла.

Определение степени: число всех соседних узлов узла v_i называется его степенью d_i, в ориентированных графах сумма всех соседних узлов, указывающих на v_i, называется степенью d_i^{in}, а сумма соседних узлов, указывающих на v, называется степенью d_i^{out}. Большинство реальных сетей не являются случайными. Несколько узлов часто имеют много соединений, в то время как большинство из них имеют несколько узлов. Распределение узлов по степеням соответствует распределению по степенному закону, и это называется безразмерной характеристикой сети (без масштаба). Сложная сеть с распределением мощности в соответствии с распределением по степенному закону называется сетью без масштаба. Степень централизации узла: очевидно, что важность узла связана с количеством соседей узла, то есть, чем больше степень узла, тем более важным становится узел.

$$C_D(v_i) = d_i = \sum_j A_{ij} . \quad (1)$$

Нормализованная степень центральности:

$$C_D^*(v_i) = d_i / (n-1) . \quad (2)$$

С точки зрения расстояния, чем важнее узел, тем быстрее он достигает других узлов сети. Центральная близость используется для оценки близости узла ко всем другим

узлам и выражается как обратная величина среднего значения расстояния кратчайшего пути узла.

$$C_C(v_i) = \left[\frac{1}{n-1} \sum_{j \neq i}^{n} g(v_i, v_j) \right]^{-1} = \frac{n-1}{\sum_{j \neq i}^{n} g(v_i, v_j)}. \quad (3)$$

Промежуточность узла представляет собой число кратчайших путей через узел в сети.

$$C_B(v_i) = \sum_{v_s \neq v_i \neq v_t \in V, s<t} \frac{\delta st(v_i)}{\delta st}, \quad (4)$$

где sδt – число кратчайших путей, которые существуют между узлами v_s, v_t и δ_{st} (v_i) является число путей через точку v_i среди этих кратчайших путей.

Взаимосвязь отражает роль и влияние соответствующего узла или ребра во всей сети и имеет большое практическое значение. Межличностные отношения включают в себя сильные и слабые связи. Важность различения этих связей имеет значение для понимания механизма распространения информации. Соединение является ребром в сети, сила соединения отражается в весе ребра в сети и в модели распространения информации, то есть в настройке вероятности или порога распространения информации для разных ребер.

Статистические характеристики социальных сетей. Распределения по степенному закону в сети: узлы с большой степенью в сети составляют лишь небольшую часть от общего числа узлов, в то время как узлы с небольшой степенью составляют большинство.

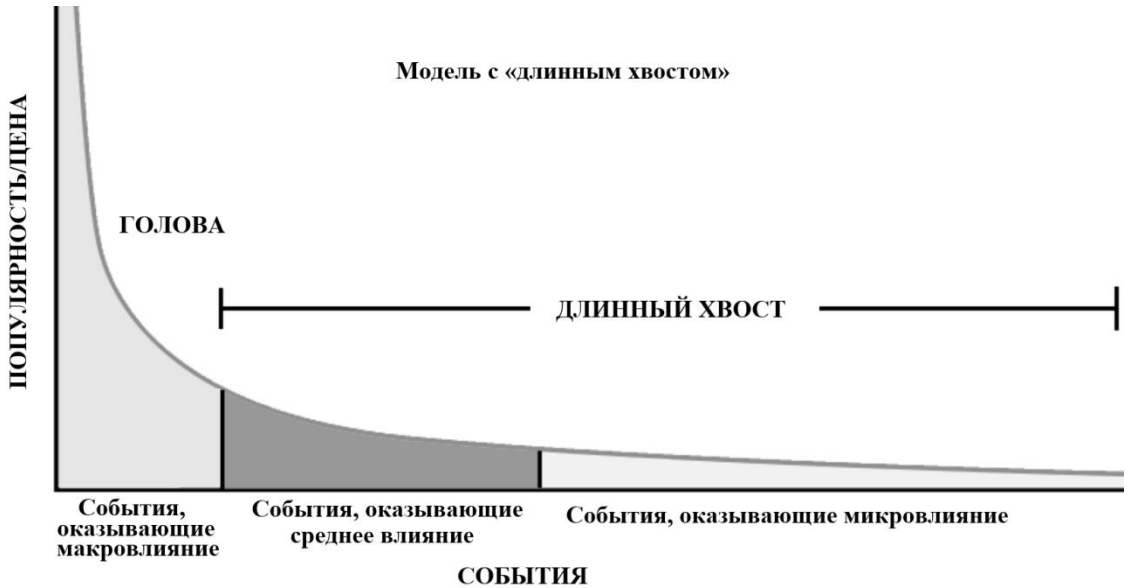

Рисунок 2 – Распределение мощности

Характеристика малого мира. Феномен малого мира означает, что каждый человек в мире может быть связан короткой цепочкой социальных отношений. Сеть малого мира расширяет феномен малого мира не только социальной сети, но и любой сети. Сеть малого мира – математическое описание явления малого мира.

Распространение без масштабирования. В теории сетей сеть без масштаба – сложная сеть с определенным классом характеристик. Типичной особенностью такой сети является то, что большинство узлов в ней подключены только к нескольким узлам, и есть несколько узлов, подключенных ко многим узлам. Наличие таких ключевых узлов делает сеть без масштабирования очень устойчивой к неожиданным сбоям, но

уязвимой для совместных атак. Особенностью сети без масштабирования является то, что ее распределение по степеням не имеет определенного среднего показателя, то есть степень большинства узлов близка к этой.

Теория шести рукопожатий. Шесть рукопожатий [15] – теория о том, что любой человек в мире может быть связан с любым другим человеком на планете через цепочку знакомых, имея не более пяти посредников. Теория показывает, что в социальных сетях существуют «слабые связи», которые делают дистанцию между людьми очень короткой.

Вывод. В статье представлены основные теоретические знания о социальных сетях и некоторые важные показатели теории графов.

1. Yu, M. Liu, W. Dou, X. Liu and S. Zhou, "Networking for Big Data: A Survey," in IEEE Communications Surveys & Tutorials, vol. 19, no. 1, pp. 531-549, Firstquarter 2017, doi: 10.1109/COMST.2016.2610963.
2. Domingos P, Richardson M. Mining the network value of customers[C]//Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. 2001: 57-66.
3. D. Kempe, J. Kleinberg, and E. Tardos, “Maximizing the spread of influence through a social network,” in KDD’03, pp. 137–146, ACM New York, NY, USA,2003.
4. Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance NS (2007) Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining (KDD’07).
5. Kimura M, Saito K, Nakano R. Extracting influential nodes for information diffusion on a social network[C]//AAAI. 2007, 7: 1371-1376.
6. Goyal A, Lu W, Lakshmanan L V S. CELF++: optimizing the greedy algorithm for influence maximization in social networks[C]. Proceedings of the 20th international conference companion on World Wide Web. ACM, 2011: 47-48.
7. A. Goyal, W. Lu, and L. Lakshmanan, “Simpath: An efficient algorithm for influence maximization under the linear threshold model,” in Data Mining (ICDM), 2011 IEEE 11th International Conference on, pp. 211–220, IEEE, 2011.
8. E. Cohen, D. Delling, T. Pajor, and R. F. Werneck, “Sketch-based influence maximization and computation: Scaling up with guarantees,” in Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 629–638, ACM, 2014.
9. Chen W, Wang Y, Yang S. Efficient influence maximization in social networks[C]. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009: 199-208.
10. Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks[C]. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010: 1029-1038.
11. Y. Tang, X. Xiao, and Y. Shi, “Influence maximization: Near-optimal time complexity meets practical efficiency,” in Proceedings of SIGMOD international conference on Management of data, pp. 75–86, ACM, 2014.
12. Y. Tang, Y. Shi, and X. Xiao, “Influence maximization in near-linear time: A martingale approach,” in Proceedings of SIGMOD International Conference on Management of Data, pp. 1539–1554, ACM, 2015.
13. C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, “Maximizing social influence in nearly optimal time,” in Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA ’14, pp. 946–957, SIAM, 2014.

14. H. T. Nguyen, M. T. Thai, and T. N. Dinh. Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks. In SIGMOD, pages 695– 710, 2016.
15. S. Milgram. The Small World Problem. Psychology Today, 1967, Vol. 2, 60– 67