НАУЧНОЕ ИЗДАНИЕ МГТУ ИМ. Н. Э. БАУМАНА

НАУКА и ОБРАЗОВАНИЕ

Эл № ФС77 - 48211. Государственная регистрация №0421200025. ISSN 1994-0408

электронный научно-технический журнал

Исследование спектральных свойств социального графа сети LiveJournal

 # 09, сентябрь 2013

DOI: 10.7463/0913.0603441

Ключарёв П. Г., Чесноков В. О.

УДК 519.177; 51-77

Россия, МГТУ им. Н.Э. Баумана

pk.iu8@yandex.ru

v.o.chesnokov@yandex.ru

Введение

В настоящее время практически каждый пользователь интернета имеет аккаунт в одной из социальных сетей. Став одной из важнейших площадок для межличностных коммуникаций, социальные сети приобрели привлекательность для различного рода исследований. Исследование структуры социальных сетей может найти применение во многих сферах деятельности человека. На основе анализа структуры социального графа можно, например, выявить пути распространения информации или выделить тесно связанные группы пользователей.

Как известно, структура социальной сети может быть представлена социальным графом. Одной из основных характеристик графа является его спектр. Спектральная теория графов, включающая в себя теорию расширяющих графов, нашла множество применений в математике, информатике, криптографии и других науках. Применение результатов этой теории для исследования социальных графов позволяет надеяться на получение новых результатов в социологии.

В данной работе исследуются спектральные свойства социального графа LiveJournal — площадки для блоггинга с возможностями социальной сети. Согласно данным рейтинга Alexa [12], этот сайт является 11-м по популярности в России и 132-м в мире. Несмотря на это, LiveJournal имеет относительно скромную аудиторию: в русскоязычном сегменте совсем недавно был зарегистрирован пятимиллионный аккаунт [1].

По сведениям руководителя LiveJournal Ильи Дронова [2], в Живом Журнале зарегистрировано более 30 миллионов пользователей, из которых активны около 2 миллионов.

1. Основные определения

В литературе (см., например, работы [13, 16]) обычно приводится довольно общее определение термина «социальная сеть»: множество социальных или межличностных отношений, объединяющее индивидуумов в социальные группы. Схожее определение дает Оксфордский словарь английского языка для словосочетания «social network». Однако он дает и другую трактовку, определяя социальную сеть как веб-сайт или другое приложение, позволяющее пользователям общаться друг с другом посредством публикации информации, комментариев, сообщений, изображений и др. [9]. Хорошее определение социальной сети дано в работе [6], на основе которой мы дадим следующее определение.

Под социальной сетью будем понимать ресурс (сайт), содержащий данные пользователей и связи между ними, позволяющий пользователю [6, 9]:

• создавать публичный или частично публичный профиль, в котором он может указывать личную информацию;

• задавать и поддерживать список других пользователей, с которыми у него имеются некоторые отношения (например, дружбы, родства, деловых и рабочих связей и т.п.);

• просматривать свой и чужие списки связей и посещать профили из этих списков;

• создавать публичные записи и комментировать записи других пользователей;

• обмениваться с другими пользователями сообщениями, изображениями и другой мультимедийной информацией.

Под социальной связью будем понимать любую связь между двумя пользователями, указанную в социальной сети и свидетельствующую о наличии регулярного взаимодействия между этими пользователями. Природа связей между пользователями может меняться в зависимости от сайта [6]. Простейшими примерами социальных связей могут быть родственные или дружественные отношения. Социальные связи могут быть направленными, т.е. связь может быть односторонней.

Социальный граф определим как граф, вершинами которого являются пользователи социальной сети, а ребрами — социальные связи между ними.

Приведем теперь определения некоторых понятий из спектральной теории графов.

Спектр неориентированного графа — это набор собственных значений матрицы смежности графа, упорядоченный по убыванию:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n.$$

Коэффициентом реберного расширения неориентированного регулярного мультиграфа $G(V, E)$ называют величину

$$h(G) = \min_{\{S \subset V | 0 < |S| \leq \frac{|V|}{2}\}} \frac{|\partial S|}{|S|},$$

где ∂S — множество ребер, каждое из которых инцидентно ровно одной вершине из S [10]. Коэффициентом вершинного расширения неориентированного регулярного мультиграфа

$G(V, E)$ называют величину

$$h_{\text{out}}(G) = \min_{\{S \subset V | 0 < |S| \leq \frac{|V|}{2}\}} \frac{|\partial_{\text{out}} S|}{|S|},$$

где $\partial_{\text{out}} S$ — множество вершин, у каждой из которых есть ровно одна соседняя вершина из множества S [4].

2. Постановка задачи

Целью работы является получение наибольшей компоненты связности социального графа социальной сети LiveJournal (Живой журнал), вычисление двух первых компонентов спектра полученного графа и оценка значений его коэффициентов расширения. При этом граф рассматривается как неориентированный.

Для достижения поставленной цели решены следующие задачи:

1) разработка и реализация методов получения информации о социальном графе;

2) выбор и реализация алгоритмов обработки информации.

3. Методы получения информации о социальном графе

Для получения социального графа сети LiveJournal был использован следующий комбинированный подход. Большая часть данных была собрана с использованием доступа через специализированный интерфейс для ботов. Поскольку LiveJournal определяет ботов по полю HTTP-заголовка User-Agent [3], для обхода блокировки по этому критерию было введено динамическое поле User-Agent. Для этого была использована база таких полей с форума techpatterns.com [11], сконвертированная для удобства использования из XML в json.

Для ботов в Живом Журнале также существует ограничение в виде числа соединений — не больше 5 в секунду [3]. При нарушении этого ограничения боты будут заблокированы. Для обхода блокировок по IP-адресу была использована анонимная сеть Tor, поскольку она имеет наибольшее число узлов в сравнении с другими подобными сетями.

В процессе работы программы сбора данных было обнаружено, что сервис Живого Журнала для ботов выдает неполные данные. Экспериментальным путем было выяснено, что о пользователях, имеющих 2000 и более друзей, сервис будет выдавать некорректные данные, ограничивая выдачу связей и показывая лишь некоторую часть друзей, не менее 2000, но не более 2500. При этом сервис не позволяет получить данные об остальных связях, не попавших в ответ на запрос, и никаким образом не указывает, что выдача была сокращена. В связи с этим для пользователей, имеющих более 2000 друзей, был произведен повторный сбор данных, с использованием метода парсинга HTML-страниц их профилей.

В качестве алгоритма обхода вершин был использован классический алгоритм поиска в ширину. Первая вершина была выбрана произвольно, в качестве нее выступил Живой

Журнал пользователя $lytdybr$. Далее обрабатывались его друзья, друзья друзей и т.д., пока не был совершен полный обход компоненты связности социального графа.

4. Метод анализа социального графа

Для вычисления первого собственного числа был использован алгоритм итерации степени (Power Iteration) [14].

Для получения второго собственного числа был использован метод понижения степени Виландта [14]. С его помощью можно получить из матрицы A с собственными значениями $\lambda_1, \lambda_2, \ldots, \lambda_n$ матрицу W с собственными числами $\lambda_1 - \sigma, \lambda_2, \ldots, \lambda_n$. Таким образом, подставляя вместо σ полученную оценку λ_1 и применив метод Power Iteration к матрице W, можно получить ее наибольшее собственное значение, равное λ_2. Для того, чтобы получить матрицу W из A, необходимо применить следующее преобразование:

$$W = A - \lambda_1 v_1 x_1^T, \tag{1}$$

где v_1 — собственный вектор, соответствующий наибольшему собственному значению, x_1 — произвольный вектор, удовлетворяющий условию $v_1^T x_1 = 1$. Наиболее простой метод построения x_1 — составить его на основе вектора v_1:

$$x = \frac{1}{\|v_1\|^2} v_1. \tag{2}$$

Для оценки диаметра графа было использовано отношение, предложенное Чанг [7]:

$$D(G) \leq \left\lceil \frac{\log((1-w^2)/w^2)}{\log(|\lambda_1|/|\lambda_2|)} \right\rceil,$$

где w — наименьшая по модулю ненулевая координата нормированного собственного вектора, соответствующего наибольшему собственному значению.

Для оценки коэффициента реберного расширения было использовано неравенство Чигера [8]:

$$\frac{1}{2}(\lambda_1 - \lambda_2) \leq h(G) \leq \sqrt{2\lambda_1(\lambda_1 - \lambda_2)}$$

Для оценки сверху коэффициента вершинного расширения была использована оценка, полученная Бобковым [5]:

$$h_{\text{out}}(G) \leq \left(\sqrt{4(\lambda_1 - \lambda_2)} + 1\right)^2 - 1.$$

Нижняя оценка была получена из отношения $h \leq h_{\text{out}} \cdot \lambda_1$:

$$h_{\text{out}}(G) \geq \frac{h(G)}{\lambda_1}.$$

Строго говоря, оценки коэффициентов расширения справедливы только для регулярных графов, однако мы будем считать, что эти оценки приближенно выполняются и для наибольших связных компонент графов социальных сетей. В общем случае нахождение этих коэффициентов представляет собой NP-трудную задачу.

Для оценки относительной погрешности вычислений, а также в качестве критерия сходимости было использовано следующее соотношение:

$$\varepsilon\|v^{(k)}\| \leq \|Av^{(k)} - \lambda^{(k)}v^{(k)}\|, \tag{3}$$

где ε — относительная погрешность вычислений, $\lambda^{(k)}$ и $v^{(k)}$ — наибольшее собственное число и соответствующий ему собственный вектор матрицы A на k-й итерации алгоритма.

5. Программное обеспечение

Для решения задач статьи было разработано программное обеспечение для получения и анализа социального графа. Данные через анонимную сеть Tor были скачаны в несколько потоков на нескольких виртуальных машинах, предварительно обработаны и записаны в реляционную базу данных (использовалась СУБД PostgreSQL). Задачи распределялись с помощью базы данных ключ-значение, реализованной на основе Redis. После завершения работы программы сбора данные из базы данных были переданы программе анализа спектральных характеристик.

6. Результаты работы

В результате работы программы сбора данных были собраны данные о 6 126 529 пользователях, при этом общее число направленных связей составило 148 014 932. После удаления петель число ребер сократилось на 2 597 522 до 145 417 410. Если сопоставить полученные данные со статистикой Живого Журнала, согласно которой в нем около 2 миллионов активных пользователей [2], и принять во внимание общеизвестный факт о том, что графы социальных сетей имеют одну большую компоненту связности, содержащую более 95% пользователей, можно утверждать, что были собраны данные о всех активных пользователях Живого Журнала.

Среди собранных пользователей были рассмотрены четыре категории:

1) пользователи, у которых нет друзей;

2) пользователи, которые в друзьях у ровно одного человека;

3) пользователи, которые имеют наибольшее число друзей;

4) пользователи, которые в друзьях у большого числа людей.

В результате просмотра случайных выборок данных категорий, каждая из которых имела объем 30 аккаунтов, выяснилось следующее:

1) в первую категорию попали преимущественно неактивные пользователи, «мертвые» аккаунты и боты;

2) во второй категории присутствуют как боты, так и настоящие пользователи, причем довольно часто они связаны с пользователями из первой категории;

3) почти все пользователи из третьей категории являются реальными людьми — известными блоггерами, дизайнерами, журналистами и политиками;

4) четвертая и третья категории пользователей практически совпадают.

На основе этих соображений с учетом условия связности графа было принято решение удалить пользователей первой категории из базы данных, для чего применялся итеративный способ. После 17 итераций, когда число пользователей первой категории свелось к 5, было принято решение о дальнейшей нецелесообразности удаления, так как данное число пренебрежимо мало в сравнении со всей базой данных. Итоговое число пользователей получилось равным 5 225 254, а число ребер — 138 101 942. После удаления пользователей была произведена процедура перенумерации вершин.

Общее время сбора составило чуть более 5 суток, сбор осуществлялся одним узлом в 150 потоков. Средняя скорость обработки пользователей составила 843 пользователя в минуту. Данный показатель превышает аналогичный, равный 334, полученный в [15] в 2,5 раза. При добавлении дополнительного узла или потоков скорость обработки не увеличивалась. На этом основании можно сделать вывод, что узким местом системы сбора являются сервера Живого Журнала и достигнут предел его пропускной способности. За пять суток работы было собрано 2,7 Гб «сырых» данных.

Поскольку в системе распределения задач заложена возможность ограничения глубины обхода графа, в каждой задаче указывалась ее удаленность от начальной вершины. Таким образом, можно оценить сверху диаметр графа через эксцентриситет начальной вершины, равный 10.

После обработки данные были экспортированы в двоичный файл. Поскольку экспериментальным путем было выяснено, что данные помещаются в оперативной памяти, дальнейшая их обработка производилась именно в ней.

Полное время обработки графа составило 38 минут 25 секунд. Из них 62 секунды были потрачены на заполнение памяти, 528 — на вычисление наибольшего по модулю собственного значения за 29 итераций, 1715 — на вычисление второго по величине за 99 итераций. Относительная погрешность измерений первого собственного вектора составила 0.1%, а второго — 2,2%. В результате работы программы были получены следующие данные:

а) $\lambda_1 = 989 \pm 1$;

б) $\lambda_2 = 709 \pm 16$;

в) $D(G) \leq 38$;

г) $137 \leq h(G) \leq 760$;

д) $0{,}14 \leq h_{\text{out}}(G) \leq 1212$.

К сожалению, существующие математические методы не позволяют повысить точность нижней оценки коэффициента вершинного расширения. Вычисления производились на персональном компьютере с процессором Intel Core i5-660 (3.33GHz), оперативной памятью 8Гб и операционной системой Linux (дистрибутив Xubuntu 12.10 x64).

Заключение

В данной статье был использован комплексный метод для доступа к социальной сети LiveJournal, позволяющий обойти некоторые ограничения с помощью использования анонимной сети Tor и динамической смены идентификатора клиентского приложения. Программа сбора данных реализует выбранный метод доступа к социальной сети и имеет следующие основные возможности:

1) многопоточная обработка до 843 запросов в минуту;

2) динамическая смена IP-адреса средствами Tor;

3) динамическая смена идентификатора клиентского приложения;

4) возможность обхода графа с ограничением глубины;

5) возможность обработки других социальных сетей путем создания классов.

Программа анализа спектральных характеристик реализует алгоритм итерации степени и позволяет вычислять оценки двух наибольших собственных чисел из спектра, коэффициентов реберного и вершинного расширений, диаметра произвольного графа, задаваемого входным файлом, содержащим список ребер. Вычисления производятся в оперативной памяти.

В результате работы программного комплекса была собрана максимальная по размеру компонента связности социального графа Живого Журнала. После удаления всех пользователей, не имеющих друзей, как неактивных была осуществлена обработка собранного графа и получены оценки его спектральных характеристик.

Результаты данной статьи могут стать первым шагом в новом направлении исследований — применении спектральной теории графов для анализа социальных сетей.

Работа выполнена при частичной поддержке РФФИ (грант № 12-07-31012).

Список литературы

1. В русскоязычном LiveJournal — 5 миллионов аккаунтов: дайджест // LiveJournal: сайт. Режим доступа: http://www.livejournal.ru/themes/id/29599 (дата обращения 03.02.2013).
2. Дронов И. ЖЖ атакуют постоянно // РИА Новости: сайт. Режим доступа: http://ria.ru/interview/20111201/503613736.html (дата обращения: 05.04.2013).
3. Правила LiveJournal для роботов // LiveJournal: сайт. Режим доступа: http://www.livejournal.com/bots/ (дата обращения 7.02.2013).
4. Alon N., Capalbo M.R. Explicit Unique-Neighbor Expanders // Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science (FOCS 2002), 16–19 November 2002, Vancouver, BC, Canada. IEEE Computer Society, 2002. P. 73–79. DOI: 10.1109/SFCS.2002.1181884.
5. Bobkov S., Houdre C., Tetali P. λ_∞, Vertex Isoperimetry and Concentration // COMBINATORICA. 2000. Vol. 20, no. 2. P. 153–172.

6. Boyd Danah M., Ellison N.B. Social network sites: Definition, history, and scholarship // Journal of Computer-Mediated Communication. 2007. Vol. 13, no. 1. P. 210–230. DOI: 10.1111/j.1083-6101.2007.00393.x.
7. Chung F.R.K. Diameters and eigenvalues // Journal of the American Math. Soc. 1989. Vol. 2, no. 2. P. 187–196.
8. Chung F.R.K. Spectral graph theory. American Mathematical Society, 1997. 207 p. (CBMS: Conference Board of the Mathematical Sciences. Regional Conference Series in Mathematics; no. 92).
9. Definition of social network in Oxford Dictionaries (British and World English) // Oxford Dictionaries: website. Available at: http://oxforddictionaries.com/definition/english/social%2Bnetwork, accessed 26.03.2013.
10. Hoory S., Linial N., Wigderson A. Expander graphs and their applications. An Overview // Bulletin of American Mathematical Society. 2006. Vol. 43. P. 439–561.
11. Firefox UserAgent Switcher list // Tech Patterns: website. Available at: http://techpatterns.com/forums/about304.html, accessed 21.03.2013.
12. Livejournal.com Site Info // Alexan: website. Available at: http://www.alexa.com/siteinfo/livejournal.com, accessed 10.04.2013.
13. Pattison P. Algebraic Models for Social Networks. Structural Analysis in the Social Sciences. Cambridge University Press, 1993.
14. Saad Y. Numerical methods for large eigenvalue problems. Publ. of Society for Industrial and Applied Mathematics, 2011. (Classics in applied mathematics).
15. Hsu W.H., Lancaster J.P., Paradesi M.S.R., Weninger T. Structural Link Analysis from User Profiles and Friends Networks: A Feature Construction Approach // Proceedings of ICWSM. Boulder, CO, USA, 2007. P. 75–80.
16. Wasserman S., Faust K. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994. 857 p. (Structural Analysis in the Social Sciences).

SCIENTIFIC PERIODICAL OF THE BAUMAN MSTU

SCIENCE and EDUCATION

EL № FS77 - 48211. №0421200025. ISSN 1994-0408

electronic scientific and technical journal

Study of the spectral properties of LiveJournal's social graph

 # 09, September 2013

DOI: 10.7463/0913.0603441

Klyucharev P. G., Chesnokov V. O.

Bauman Moscow State Technical University
105005, Moscow, Russian Federation
pk.iu8@yandex.ru
v.o.chesnokov@yandex.ru

In this paper we compute some characteristics of the spectrum of LiveJournal's social graph and estimate the vertex and edge expansion ratios of this graph. We use the Power Iteration algorithm to compute first and second elements of the graph's spectrum. The concept of application of the spectral graph theory methods to problems of social network analysis appeared to be very promising. Methods described in this paper could be applied to analysis of social networks, study of the social interactions between people and many other problems at the intersection of sociology, computer science and information security.

References

1. *V russkoyazychnom LiveJournal — 5 millionov akkauntov: daydzhest* [The Russian-speaking segment of LiveJournal now has 5 million accounts]. LiveJournal: website. Available at: http://www.livejournal.ru/themes/id/29599, accessed 03.02.2013.
2. Dronov I. *ZhZh atakuyut postoyanno* [LJ are constantly attacked]. RIA Novosti: website. Available at: http://ria.ru/interview/20111201/503613736.html, accessed 05.04.2013.
3. *Pravila LiveJournal dlya robotov* [LiveJournal rules for robots]. LiveJournal: website. Available at: http://www.livejournal.com/bots/, accessed 7.02.2013.
4. Alon N., Capalbo M.R. Explicit Unique-Neighbor Expanders. *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science (FOCS 2002)*, 16-19 November 2002, Vancouver, BC, Canada. IEEE Computer Society, 2002, pp. 73–79. DOI: 10.1109/SFCS.2002.1181884.
5. Bobkov S., Houdre C., Tetali P. λ_∞, Vertex Isoperimetry and Concentration. *COMBINATORICA*, 2000, vol. 20, pp. 153–172.

6. boyd danah M., Ellison N. B. Boyd Danah M., Ellison N.B. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 2007, vol. 13, no. 1, pp. 210–230. DOI: 10.1111/j.1083-6101.2007.00393.x.
7. Chung F.R.K. Diameters and eigenvalues. *Journal of the American Math. Soc.*, 1989, vol. 2, no. 2, pp. 187–196.
8. Chung F.R.K. *Spectral graph theory*. American Mathematical Society, 1997. 207 p. (*CBMS: Conference Board of the Mathematical Sciences. Regional Conference Series in Mathematics*; no. 92).
9. *Definition of social network in Oxford Dictionaries (British and World English)*. Oxford Dictionaries: website. Available at: http://oxforddictionaries.com/definition/english/social%2Bnetwork, accessed 26.03.2013.
10. Hoory S., Linial N., Wigderson A. Expander graphs and their applications. An Overview. *Bulletin of American Mathematical Society*, 2006, vol. 43, pp. 439–561.
11. *Firefox UserAgent Switcher list*. Tech Patterns: website. Available at: http://techpatterns.com/forums/about304.html, accessed 21.03.2013.
12. *Livejournal.com Site Info*. Alexan: website. Available at: http://www.alexa.com/siteinfo/livejournal.com, accessed 10.04.2013.
13. Pattison P. *Algebraic Models for Social Networks*. Cambridge University Press, 1993. (*Structural Analysis in the Social Sciences*).
14. Saad Y. *Numerical methods for large eigenvalue problems*. Publ. of Society for Industrial and Applied Mathematics, 2011. (*Classics in applied mathematics*).
15. Hsu W.H., Lancaster J.P., Paradesi M.S.R., Weninger T. Structural Link Analysis from User Profiles and Friends Networks: A Feature Construction Approach. *Proceedings of ICWSM*. Boulder, CO, USA, 2007. pp. 75–80.
16. Wasserman S., Faust K. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994. 857 p. (*Structural Analysis in the Social Sciences*).