Графовый подход при составлении характеристики социального объекта

О.С. Смирнова, В.В. Шишков

***Аннотация* – В статье рассмотрен графовый анализ как подход, применяющийся для получения знаний о составном объекте – тематической группе пользователей социальной сети. С этой целью описаны средства сбора и обработки данных, алгоритмы, форматы, варианты использования инструментария социально-сетевого анализа. Приведены примеры задач, решаемых с помощью применения графов в анализе социальных данных. Затронута тема визуализации дискретных и континуальных характеристик объекта на графе, а также методология оценки активности бот-профилей в распространении инфоповода. Представлено средство сбора социальных данных и визуализации пользовательских данных социальных групп.**

***Ключевые слова* – графовый анализ; социальные сети; социальный анализ; социограмма; бот-аккаунт; Twitter; социограмма; распространение контента.**

Говоря об анализе социальных структур мы подразумеваем, что имеется набор объектов, связанных между собой определенным образом и обладающих некоторыми характеристиками. Примеры таких связей легко прослеживаются в современном обществе – ими могут выступать дуальные социальные связи, такие как:

– связи родства;

– дружеские связи;

– связи, образованные односторонними информационными потоками (т.н. подписки);

– иерархические связи – к примеру, являющиеся частью должностной иерархии.

Описывая такой социальный объект в терминах теории графов мы получаем структуру, условно называемую социограммой, опираясь на характеристики которой мы можем делать выводы об объекте исследования. Такой подход моделирования с применением графов используется в антропологии, биологии, экономике, географии [1].

Рассмотрим более прикладной вариант применения, а именно – анализ социальных групп на основе открытых данных социальных сетей. К тому же, использование социальной сети в качестве готового источника позволит нам опустить такой этап социального исследования как оперативный сбор данных.

Изучение социальных тематических групп и особенностей экспансии информации внутри них позволяет выявлять закономерности в распространении новостей, трендов, различных информационных вбросов. Только за прошедший год было не менее 3 скандалов политической тематики, широко освящающихся в СМИ и связанных с действиями бот-программ, «фейковых» аккаунтов, «фабрики троллей». Анализ графовых сетей позволяет вычислять характер и объем бот-аккаунтов, участвующих в распространении того или иного инфоповода. В терминах анализа социальных сетей легко решаются прикладные задачи восстановления связей (поиск родственников, общих знакомых и т.д.) [2].

При работе с графами используется, в первую очередь, некие информационные программные средства, которые представляют нам возможности сбора данных – скриптовые языки программирования, различные API внешних сервисов, программные библиотеки для сбора данных, эмуляторы браузеров для имитации естественного механизма доступа к ресурсам. А также средства аналитики данных, которые могут быть разделены на группы, в соответствии с характером использования и предназначения – программное обеспечение, заточенное на обработку графов с количеством узлов порядка 10^7, часто играющее роль «бэкэнда». Представителями такого софта являются Gephi, Cytoscape [3, 4]. JavaScript библиотеки – браузерные скрипты, позволяющие отображать интерактивные графы, используя примитивы SVG или рендер в canvas – SigmaJS, AlchemyJS [5, 6]. Если решаются задачи, в которых для доступа к связанным данным одного типа реализуется связи многие-ко-многим и задачи выборки

Статья получена 30.05.2017 г.

Исследование выполнено федеральным государственным бюджетным образовательным учреждением высшего образования «Московский технологический университет» (МИРЭА) за счет гранта Российского фонда фундаментальных исследований (проект №16-37-00492).

О.С. Смирнова, МИРЭА (e-mail: mail.olga.smirnova@yandex.ru).

В.В. Шишков, МИРЭА (e-mail: shishkov61@gmail.com).

объектов по связям в приоритете, то целесообразно использование графовой СУБД – Neo4j [7].

Для сериализации социограмм используются множество форматов. Самые популярные из них – GEXF – подмножество XML, предназначенное для описания графов с параметрами узлов, допускающими вложенность, описание узлов и связей между ними в нотации JSON, и даже создание таблицы связности в виде CSV файла [8].

Рассматривая социальные группы как предмет аналитики важно уделить внимание расположению узлов в разложенном на плоскости графе – оно позволяет визуально оценить граф. Существует порядка нескольких десятков алгоритмов, производящих раскладку графа в двух измерениях, нас интересуют в первую очередь те, которые легко параллелятся, чтобы мы могли использовать нескольких потоков, к примеру – Force Atlas – однофазный алгоритм без условия выхода, действует «расталкивая» узлы, при этом ребра имеют ограниченную длину (рисунок 1); Openord – многофазный алгоритм определенной продолжительности, при выполнении каждой фазы, выполняется с комбинацией параметров, специфичных фазе [9, 10].

Рисунок 1 – Типичный социальный граф

Использование цвета тоже довольно очевидно при возможности разбить элементы на группы (выполнить кластеризацию). Для этого используются как специализированные алгоритмы, например, Алгоритм кластеризации Маркова, так и просто разделение объектов по классу модульности. Также в цвет может быть положена не дискретная величина, тогда соответственно это будет градиент из одного цвета в другой. К примеру, это может быть:

– коэффициент модульности, высокое значение этого показателя указывает на сложную внутреннюю структуру, которая часто называется структурой сообщества и описывает как сеть разделена на подсети или сообщества;

– центральность – одна из метрик, целью которых является количественное определение «важности» или «влияния» (в различных смыслах) конкретного узла (или группы) внутри сети;

– плотность – доля прямых связей в сети по отношению к их общему числу;

– расстояние от объекта – минимальное количество связей, необходимых для соединения двух отдельных элементов сетевого графа [11].

Для использования размера объекта в качестве отображения дополнительного параметра на визуализации целесообразно брать параметры, так же выраженные непрерывными величинами.

Известный интернет исследователь – Александр Лоуренс на основе нескольких ключевых фраз, в частности, посвященных убийству политического деятеля Немцова, а также связанных с активностью издания «Новая газета» собрал расширенные сообщества пользователей Twitter, составленные из друзей и читателей каждого аккаунта.

На основании метаданных пользователей были выбраны бот-аккаунты, основываясь на наличии указанного часового пояса в профиле и наполнении списка «избранное».

При соединении собранных групп в единый набор данных, в результате оказалось 17 590 тесно связанных Twitter-аккаунтов (рисунок 2). Следует упомянуть, что в выборках по случайной фразе связность результатов довольна низка. Метаданные подтвердили, что большая часть аккаунтов действительно является бот-профилями. 93% не указали в профиле местоположения, 96% не имели информации о часовом поясе и 97% не имели записей в «Избранном». Кроме того, несмотря на то, что каждый аккаунт в среднем опубликовал по 2 830 твитов, они почти никогда не взаимодействовали с другими пользователями. Такие исследования являются серьёзной поддержкой идеи о том, что бот-профили были созданы одним агентством с определенной целью [12].

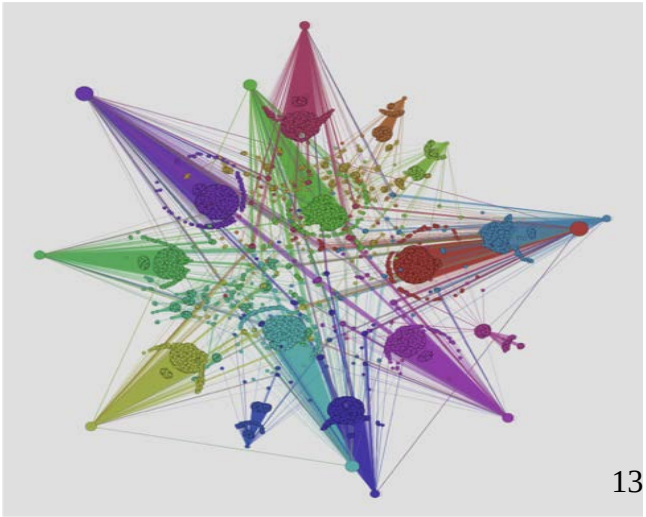

Рисунок 2 – Связанный кластер бот-аккаунтов

При визуальной оценке графового отображения социальной группы выделяются кластера в соответствии с тем, какие действия ведут пользователи – подписчики, те, кто взаимодействует с контентом, но не является подписчиком сообщества, и бот-аккаунты, образующие отдельный связанный кластер, плохо взаимодействующий с содержимым группы. Простой анализ объема этих кластеров служит показателем «качества инфоповода» – чем больше кластер бот-аккаунтов, тем более искусственной выглядит повестка группы [13].

Для подготовки данных разработано программное средство, служащее вспомогательным инструментом для анализа социальных групп, опубликованное под Универсальной общественной лицензией GNU. Средство позволяет скачать, просмотреть в браузере содержимое группы, связи в ней и некоторые статистические характеристики сообщества, а также выгрузить их в одном из открытых форматов для работы с графами с целью последующего более глубокого анализа и использованием вышеописанных инструментов и методов [14].

В статье рассмотрено применение теории графов к социальным объектам. Описаны преимущества использования инструментария социально-сетевого анализа. Приведен краткий разбор используемых программных средств, алгоритмов раскладки графов, варианты используемых размерностей при построении визуализации. Рассмотрены существующие кейсы использования графового анализа при анализе соц. сетей, а также представлено программное средство в качестве вспомогательного инструмента для сбора данных и предварительного анализа. Свободная лицензия рассматриваемого программного средства подразумевает возможность его использования в качестве компонента при создании комплексных систем анализа социальных данных или в качестве каркаса механизма экспорта графов.

БИБЛИОГРАФИЯ

[1] Wikipedia. Анализ социальных сетей. Эл. ресурс: https://ru.wikipedia.org/wiki/Анализ_социальных_сетей (дата обращения: 17.05.2017 г.)

[2] Wikipedia. Dynamic network analysis. Эл. ресурс: https://en.wikipedia.org/wiki/Dynamic_network_analysis (дата обращения: 17.05.2017 г.)

[3] Портал gephi.org. Документация gephi. Эл. ресурс: http://gephi.org (дата обращения: 17.05.2017 г.)

[4] Портал cytoscape.org. Документация cytoscape. Эл. ресурс: http://www.cytoscape.org (дата обращения: 19.05.2017 г.)

[5] Портал sigmajs.org. Документация Sigmajs. Эл. ресурс: http://sigmajs.org (дата обращения: 19.05.2017 г.)

[6] Портал github.io. Хостинг AlchemyJS. Эл. ресурс: http://graphalchemist.github.io/Alchemy/#/ (дата обращения: 21.05.2017 г.)

[7] Портал neo4j.com. Документация Neo4j. Эл. ресурс: http://neo4j.com (дата обращения: 21.05.2017 г.)

[8] Портал gephi.org. Документация gephi. Поддерживаемые форматы. Эл. ресурс: https://gephi.org/users/supported-graph-formats/(дата обращения: 21.05.2017 г.)

[9] Wikipedia. Force directed graph drawing. Эл. ресурс: https://en.wikipedia.org/wiki/Force-directed_graph_drawing (дата обращения: 22.05.2017 г.)

[10] Портал gephi.org. Хостинг плагинов. Openord-Layout. Эл. ресурс: https://marketplace.gephi.org/plugin/openord-layout/(дата обращения: 23.05.2017 г.)

[11] Wikipedia. Social network analysis. Эл. ресурс: https://en.wikipedia.org/wiki/Social_network_analysis (дата обращения: 23.05.2017 г.)

[12] Lawrence Alexander. Social Network Analysis Reveals Full Scale of Kremlin's Twitter Bot Campaign. Эл. ресурс: https://globalvoices.org/2015/04/02/analyzing-kremlin-twitter-bots/(дата обращения: 23.05.2017 г.)

[13] Ляпунов С.М., Шишков В.В. Алгоритм анализа связи пользователей в социальной группе с помощью графического представления. Научный альманах. 2017. № 1-3 (27). С. 107-110.

[14] Хостинг проектов bitbucket.org. Проект vk_groups_analysis. Эл. ресурс: https://bitbucket.org/evil_sneer/vk_groups_analysis (дата обращения: 23.05.2017 г.).

Graph approach in drawing up the characteristics of a social object

O.S. Smirnova, V.V. Shishkov

Annotation **– The article considers graph analysis as an approach used to gain knowledge about a composite object - a thematic group of users of a social network. To this end, the means of data collection and processing, algorithms, formats, options for using the tools of social-network analysis. Examples of problems solved using graphs in the analysis of social data are given. The topic of visualization of discrete and continual characteristics of the object on the graph was touched upon, as well as the methodology for assessing the activity of bot-profiles in the dissemination of the information guide. The tools for data collection and visualization of user data of social groups are presented.**

KeyWords **– graph analysis; social networks; social analysis; sociogram; Bot account; Twitter; distribution of content.**