УДК 519.25

М.Ю. Катаев, В.В. Орлова

Анализ данных событий социальных сетей

Анализ данных открытых социальных сетей (контент) может быть выполнен на количественном и качественном уровне. Этот контент является богатым источником данных для построения и анализа взаимодействия пользователей социальных сетей при формировании различных групп, что используется не только для статистических расчетов, социальных направлений анализа, но и в торговле, для разработки рекомендательных систем. Большое количество пользователей социальных сетей приводит к огромному объему неструктурированных данных (по времени, типу общения, типу сообщения и географическому месту). Данная статья направлена на обсуждение проблемы анализа социальных сетей и получения информации из неструктурированных данных. В статье обсуждаются методы извлечения информации, известные программные продукты и наборы данных.
Ключевые слова: социальные сети, методы анализа, наборы данных, динамика сетей.
doi: 10.21293/1818-0442-2020-23-4-71-77

Как следствие изменений в технических и вычислительных решениях повысилась доступность крупномасштабных проектов, реализующих социальные сети (social network), которые позволили общаться всему населению Земли между собой. С другой стороны, эти же решения открыли новую эру исследований в области анализа данных, которые генерируются социальными сетями [1]. Количество разнородных данных (текст, звук, изображения и др.), которые ежедневно генерируются традиционными Интернет, почтовыми службами и социальными сетями через стационарные компьютеры и мобильные устройства решают потребности в общении населения. Кроме того, передаваемая информация обеспечивает отличную возможность для глубокого анализа, составления прогнозов, аналитики методами искусственного интеллекта и больших данных.

В этом направлении важными являются исследования подходов к вычислительной сложности социальных сетей (телекоммуникации), анализ потоков сообщения и групп, структуры и динамики групп. Эти исследования решают задачи не только технического развития социальных сетей, безопасности, но и практические задачи торговых направлений, маркетинга и т.д. Уже фундаментальной стала задача анализа социальных сетей (social network analysis – SNA) [2] для обнаружения сообщества пользователей сети, действия которых являются однородными по каким-то признакам (кластеризация). Далее возникает необходимость не только обнаружения, но и определения структуры сообщества, включая динамику и взаимодействие с другими сообществами (группами). Понимание процессов, которые определяют зарождение, развитие и динамику сообщества, может быть использовано для поиска знаний, необходимых для решения практических задач.

Социальные сети являются, с одной стороны, генераторами одних данных и, с другой стороны, потребителями других данных в разных масштабах не только текста, изображений, звуков, но и структурированной, сжатой информации в виде лайков, пиктограмм и т.д. Проблема понимания текстовой информации связана с возможностью двусмысленности выражения разных тем, представления слов на разных языках, использования нетиповых сокращений [3]. Это делает возможным легкость общения, но сложность в анализе этой информации. Кроме того, источники сообщений пересекаются между собой, например ленты новостей и социальные сети, что усложняет понимание передачи информации внутри группы и между группами. Обеспечение закрытости персональных данных приводит к возможным пересечениям событий разных членов группы. При передаче большого количества сообщений наблюдается нестабильность качества пользовательского контента (проблема спама и пустые аккаунты). Также изменения интерфейса сетевого общения приводят к обновлению пользовательской модели данных, что ограничивает временные сроки анализа данных (от одной версии до другой) [4]. Тем не менее имеющиеся наборы данных являются однородными по выбранным параметрам и возможными для получения качественного анализа событий социальной сети.

Надо отметить, что существуют открытые и закрытые сообщения, что не позволяет оценить в полной мере сущность сообщества в заданном информационном направлении, однако метаинформация о каждом сообщении является открытой. Эта часть сообщения позволяет видеть локализацию точки отправления и точки назначения, оценивать частоту сообщения и объем, что также важно для анализа социальных сетей. Данная статья представляет собой обзор информации, которая позволяет разобраться в таком направлении, как анализ данных социальных сетей.

Представление социальной сети в виде набора данных

Так как социальные сети представлены пользователями, которые обмениваются сообщениями между собой, образуют некоторые сообщества, связанные с некоторым единым направлением сообщений, то описание этого на уровне данных достаточно сложно и требует пояснения. На рис. 1 показан

простейший вариант графического представления социальной сети, которая состоит из набора узлов (пользователей) и связей (сообщений).

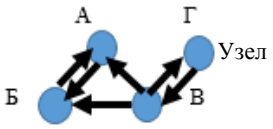

Рис. 1. Простейший вариант графического представления социальной сети

Для описания функционирования социальной сети [4, 5] вводят термины и понятия, такие как узлы, связи, отношения, края, мера центральности, меры на уровне сети, меры на уровне пути, мосты, диады и клики и т.д. Для построения графиков социальной сети (graphics social network – GSN) требуются два ключевых компонента сети, это пользователи (узлы) и связи (отношения между ними в виде сообщений). При этом сообщение может иметь ссылки на другие сообщения и веб-страницы в Интернете, которые тоже, в свою очередь, ссылаются на другие веб-страницы. Эти ссылки принято также включать в отношения между участниками процесса обмена сообщениями. Есть два типа ребер – направленные и ненаправленные. Например, если два человека (А и Б) являются друзьями в социальной сети, эти отношения не ориентированы (явно ненаправленные), так как человек А дружит с человеком Б, но также равнозначно, что человек В дружит с человеком А. Другой пример ненаправленного ребра связан с тем, что человек А находится в одной группе с человеком Б и при этом человек Б находится в группе с человеком А. Такой параметр, как вес ребра, показывает, какое количество раз ребро появляется между двумя конкретными узлами (число сообщений).

Центральность – это набор показателей, используемых для количественной оценки важности и влияния конкретного узла на сеть в целом [6]. Степень узла – это количество ребер, которое он имеет. Параметр «близость» показывает, насколько хорошо конкретный узел связан со всеми остальными узлами в сети, или, иначе, среднее количество переходов, необходимых для достижения каждого другого узла в сети. Переход – это путь ребра от одного узла к другому. Например, из рис. 1 видно, что узел А связан с узлом Б, а узел В связан с узлом Г, и, чтобы информация от узла А достигла узла Г, потребовалось бы два перехода (АВ)→(ВГ). Размер сети – это количество узлов в сети и не учитывает количество ребер (например, сеть на рис. 1 с узлами А, Б и В имеет размер 3). Плотность сети – это количество ребер, деленное на общее количество возможных ребер (например, сеть на рис. 1 с узлами А, В, Г имеет плотность сети равную 2/3, потому что есть два ребра из возможных 3). Длина – это количество ребер между начальным и конечным узлами, и расстояние – это количество ребер или переходов между начальным и конечным узлами по кратчайшему пути. Диады и клики – это пары узлов, соединенных ребрами, где диада представляет собой пару из двух узлов, а клика – это пара из трех или более узлов. Данное описание показывает, что исследование данных социальных сетей представляет собой непростой вычислительный процесс, включающий разнообразные алгоритмы.

Динамический анализ социальных сетей

В настоящее время динамика событий общества является достаточно высокой, и в этом плане социальные сети позволяют события одной страны и даже человека превращать в информацию для всего мира. Это делает возможным использование этой информации не только в научных, но и коммерческих целях, что позволило достигнуть значительного прогресса в области, связанной с анализом социальных сетей. Однако большинство известных работ сосредоточено на изучении статических ситуаций в социальных сетях или оценке динамики в глобальном масштабе (например, примером является распространение заболевания Covid). За последние годы существенно выросла доступность больших динамических наборов данных социальных сетей, что подогревает интерес к разработке автоматических подходов анализа временных событий социальных сетей.

Динамический подход в изучении поведения структуры узлов и связей социальных сетей позволяет выявить скорость роста или уменьшения размера сети, перераспределение связей между и т.д. Количественная мера оценки этих показателей позволяет определить закономерность изменений и соответственно строить прогнозные ситуации формирования тех или иных связей в социальных сетях. Понятно, что для выявления динамики изменений важным является оценка временных интервалов, определяющих четко обозначенное изменение. Разработка методик визуализации структуры сети в текущий момент времени и сравнение с прошлыми временными промежутками предоставляет возможность для более точного понимания тенденций.

Динамический анализ социальных сетей (ДАС) является новой областью, где имеется существенный потенциал для исследований и разработок аналитических программных приложений. ДАС направлен на анализ поведения социальных сетей в различных масштабах времени [1], обнаружение повторяющихся паттернов [2], структуру сообщества (формирование, развитие, существование или роспуск) [3].

Обзор баз данных, содержащих примеры данных социальных сетей

Один из способов сбора данных из социальных сетей [6] связан с использованием инструмента веб-скрапинга [7], который помогает извлекать данные из каналов социальных сетей, таких как Facebook, Twitter, LinkedIn и Instagram. Надо отметить, такой способ получения информации для некоторых сайтов социальных сетей является нарушением условий конфиденциальности, например для интернет-магазинов. Другой способ извлечения данных связан с применением API (application programming interface –

интерфейс прикладного уровня доступа к данным) для сайтов социальных сетей, таких как Facebook [8] или Twitter [9]. Третий способ получения данных основан на заранее подготовленных тестовых базданных, например, крупнейшая по объему и систематизации Stanford Large Network Dataset Collection [10] (или SNAP). Для примера в этой базе представлены данные известных социальных сетей с характеристиками, показанными в табл. 1.

Таблица 1

Данные о социальных сетях в базе-данных SNAP

Имя набора данных	Узлов	Связей	Описание
ego-Facebook	4039	88234	Социальные круги из Facebook (анонимно)
ego-Gplus	107614	13673453	Социальные круги из Google+
ego-Twitter	81306	1768149	Социальные круги из Twitter
soc-Epinions	75879	508837	Кто кому доверяет в сети Epinions.com
soc-LiveJournal	4847571	68993773	LiveJournal онлайн-социальная сеть

Другой известный набор данных The Network Data Repository with Interactive Graph Analytics and Visualization [11] представляет данные, которые накоплены в соответствии ребер, максимальную и среднюю степень, количество треугольников, средний коэффициент локальной и глобальной кластеризации и др. Также можно выделить Datasets соглашений с различными социальными сетями и группами. При получении данных, пользователь сразу видит некоторые характеристики наборов данных, например количество узлов. Значимым является ресурс, содержащий большие наборы данных Social Network Analysis. Структура и объем информации показаны на примере Twitter в табл. 2.

Таблица 2

Структура и объем информации на примере Twitter

Имя набора данных	Узлов	Связей	Описание
Twitter-Dynamic-Net	90908	443399	10^7 твитов (tweets) связанных с 156487 пользователями в динамическом режиме
Twitter-Dynamic-Action	7514	304275	Тексты пользователей по конкретной теме «Землетрясение в Гаити»
Twitter-Competitor	87603		Контент Twitter, связанный с компаниями
Twitter-Net-Tweet	$4 \cdot 10^7$	$1{,}47 \cdot 10^9$	Весь сайт Twitter в 2010 г.
Weibo-Net-Tweet	$1{,}8 \cdot 10^6$	$3{,}1 \cdot 10^8$	Пользователи Sina Weibo [12], отношения, их твиты и ретвиты

Обзор программного обеспечения анализа данных социальных сетей

Одним из распространенных программ визуализации данных социальных сетей является Gephi [13], так как не требует знаний программирования, при этом позволяя создавать разнообразные типы графиков. Входными данными могут быть разнообразные форматы, в которых записываются данные социальных сетей: узлы, связи, степень, центральность и т.д. В этой программе есть функция, которая автоматически обновляет набор данных выбранной социальной сети.

Пакет программ «sna», написанный на языке программирования R [14], предназначен для статистических вычислений и анализа данных. Этот пакет программ является полезным инструментом в области анализа социальных сетей, однако требует владения навыками программирования. Приложение UCINet [15] создано для анализа и визуализации данных социальных сетей. Важной особенностью приложения является решение задачи кластеризации для больших наборов данных, стандартных типов визуализации и формирования файлов, совмещенных с форматом Excel для Microsoft.

Программа NodeXL [16] встраивается в среду Excel, что упрощает обнаружение закономерностей и визуализации полученных результатов. Графический инструмент Graphviz [17] поддерживает графовые модели, кластеризацию, вычисляет статистические критерии, реализует стандартные топологические алгоритмы (минимальное остовное дерево и др.). Программное обеспечение NetMiner [18] предназначено для анализа и визуализации сетей передачи данных, в том числе и социальных сетей.

Особенностью программы является использование шаблонов для распознавания сетевой инфраструктуры на основе подходов интеллектуального анализа данных. Программа AutoMap [19] позволяет извлекать текст и выполнять его интеллектуальный анализ. Программное обеспечение Cytoscape [20] выполняет анализ и визуализацию данных социальных сетей, включая и семантические сети. Приложение GraphChi [21] предназначено для анализа и визуализации данных социальных сетей на основе алгоритма обработки графов. Программа NetWorkit [22] обеспечивает анализ и визуализацию данных социальных сетей с высокой производительностью и визуализацией, использует многоядерную архитектуру процессоров и видеокарт (например, NVidia [23]).

Методы анализа данных социальных сетей

Традиционный анализ данных социальных сетей выполняется на серии узлов и ребер [9], обычно получаемых из метаданных о взаимодействиях между несколькими участниками сети, без фактического анализа содержимого этих взаимодействий (сообщений). Для этих целей можно использовать информацию из баз данных, описанных выше (см. табл. 1), или из текущего набора данных, полученных по соответствующим программам. Если есть такая возможность (открытые социальные сети), то возможно объединение метаданных с данными информационного содержания каждого сообщения. Далее, применяя указанные выше программные продукты, можно перейти к выполнению анализа данных социальных сетей. Анализ позволяет полу-

чить признак, описывающий поведение субъектов сети (пользователей и групп), их настроений, а также изменения той или иной тенденции во времени. Кроме того, имея исторические данные о сети, появляется возможность анализировать ее динамику, а также предсказывать скрытые взаимосвязи, существующие в наборе данных. Кластеризация сообществ на основе поведения с течением времени может быть осуществлена путем анализа только метаданных или совместного анализа с содержанием сообщений [24].

Одной из проблем традиционного анализа социальных сетей является то, что часто рассматриваются только отношения между участниками, а не то, о чем они на самом деле отправляют друг другу сообщения. При этом не учитывается частота передачи сообщений (например, несколько раз в день, в неделю или другой промежуток времени). Часто подходы игнорируют информацию о направлении сообщений, т.е. сколько сообщений отправлено участником А для участника Б и сколько раз, участник Б ответил А. Однако заметим, что вся эта информация требуется для разных направлений исследований, например выделения тем сообщений: семейные, научные, технические и т.д. К тому же, потоки сообщений могут совмещать несколько тем для одних и тех же участников. Сложным является проблема, когда, например, два человека не являются друзьями в социальной сети, но у них есть общие друзья, поэтому они могут узнать друг друга после некоторого времени общения, а могут и не узнать, если список претендентов достаточно большой.

Самый простой способ анализа социальной сети – это отобразить сеть в виде матрицы. По столбцам и строкам матрицы расположены участники социальной сети, и тогда в каждой ячейке «1» указывает, что два человека знают друг друга, и «0», если нет. Частота их общения обозначается числом больше «1». Задача представления данных социальной сети состоит в том, чтобы была возможность использования разных математических подходов (графы, кластеризация, метод главных компонент, Марковские цепи и т.д.). Далее каждую ячейку можно превратить в другое множество данных, учитывая ключевые слова сообщений, типичное время сообщений, пространственную локацию и т.д. Как правило, эти многомерные матрицы являются разреженными. Анализ многомерных матриц содержания сообщений может быть выполнен с помощью технологии «автор–получатель–тема» (или модель ART) [25].

Так как каждый пользователь социальной сети – это узел, то в зависимости от их «дружбы» или пересылки сообщений (дуга) строится граф. К данному графу можно задать меру модулярность (modularity) [26], которая показывает, что чем эта мера меньше, тем меньше будет размер сообщества. Таким образом, можно определить понятие сообщества как определенную группу вершин, связи между которыми более плотные, чем вне группы. Величина модулярности лежит в диапазоне [−1, 1], и считается, что значения, превышающие 0,5, определяют структуру связей, связанную с сообществом.

Методы выделения сообществ социальных сетей

Рассмотрим несколько методов выделения сообществ социальных сетей. Первый метод основан на анализе близости [27] расстояний между узлами в некотором радиусе контроля. Мониторинг структуры сети позволяет определить точки перехода, т.е. процесс появления новых связей во времени в динамической развивающейся социальной сети. Для определения этих точек применяют несколько метрик: плотность графа и центральность по посредничеству (Betweenness Centrality), близости (Closeness Centrality) и радиальности (Radiality Centrality). Центральность по посредничеству показывает значимость субъекта при распространении информации в социальной сети и вычисляется как число кратчайших путей между всеми парами субъектов, которые связаны с рассматриваемым субъектом. Центральность по близости позволяет вычислить скорость распространяется информации в сети. Центральность по радиальности связана с расстоянием между узлами в некоторой окрестности заданного диаметра.

Другой метод связан с расчетом модулярности на основе жадного алгоритма оптимизации (Fastgreedy) [28, 29], одной из идей которого является перебор не всех пар из сообщества, а только тех, между вершинами которых существуют связи. Это делает алгоритм не настолько точным, но существенно быстрым, чтобы использовать для анализа больших графов социальных сетей. Следующий метод кластеризации [30–33] на графах связан с решением двух противоположных задач, связанных с поиском максимальных и минимальных расстояний внутри кластеров, и выполняет перебор всех пар из сообщества. Не менее известным является метод, основанный на многоуровневой оптимизации функции модулярности (Multilevel) [34, 35], где для каждой вершины рассматриваем изменение модулярности при перемещении данной вершины в другое сообщество вершины.

На принципах работы метода случайных деревьев основывается метод LabelPropogation, когда каждой вершине присваивается индекс. Необходимо перебрать все индексы и найти те, которые имеют максимальную встречаемость среди смежных вершин [36–38]. Реализация метода проста, однако результаты являются неустойчивыми при решении задач оценки динамики. Имея большой граф по объему, возникает сложная задача перебора, которая может быть решена методом случайного блуждания (Walktrap) [39]. Идея очень простая и связана с оценкой расстояний суммарного блуждания, которая для сообщества является минимальной. В настоящее время известно достаточно много методов решения задачи выделения сообществ, и представить их в одной статье не представляется возможным.

Заключение

В статье рассмотрены основные положения задачи, связанной с анализом данных событий социальных сетей. Социальные сети во всем своем многообразии и во многом определяют информационное пространство, которое окружает человека. Разные поколения, и особенно молодое, полностью доверяют этому источнику информации. Внутри каждой социальной сети образуются заданные сообщества (например, любители путешествий, питания, досуга и др.), а также стихийные, по каким-то направлениям. Эти задачи ставят не только технические проблемы организации бесперебойных и быстрых коммуникаций людей, но и проблемы безопасности, научные в математических и социальных науках. Поэтому существует задача, которая заключается в автоматическом поиске сообществ пользователей, оценке динамики этих сообществ. В данной статье предлагается краткое описание формализации данных социальных сетей, их записи и распространения через базы данных, методики анализа и выделения сообществ.

Работа выполнена при финансовой поддержке РФФИ 20-011-3154 опн.

Литература

1. Girvan M. Community structure in social and biological networks / M. Girvan, M. Newman // Proceedings of the National Academy of Sciences. – 2002. – Vol. 99 (12). – P. 7821–7826.

2. Гусарова Н.Ф. Анализ социальных сетей. Основные понятия и метрики. – СПб.: Ун-т ИТМО, 2016. – 67 с.

3. Фролов Ю.Н. Социальные сети: теория и практика / Ю.Н. Фролов, Л.К. Габышева. – Тюмень: ТюмГНГУ, 2012. – 140 с.

4. Додонов А.Г. Компьютерные сети и аналитические исследования / А.Г. Додонов, Д.В. Ландэ, В.Г. Путятин. – Киев: ИПРИ НАН Украины, 2014. – 486 с.

5. Watts D.J. Identify and search in social networks / D.J. Watts, P.S. Dodds, M. Newman // Science. – 2002. – Vol. 296. – P. 1302–1305.

6. McCallum A. Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email / A. McCallum, X. Wang, A. Corrada-Emmanuel // Journal of Artificial Intelligence Research. – 2007. – Vol. 30. – P. 249–272.

7. Webscraper [Электронный ресурс]. – Режим доступа: https://webscraper.io, свободный (дата обращения: 25.11.2020).

8. Developers.facebook [Электронный ресурс]. – Режим доступа: https://developers.facebook.com, свободный (дата обращения: 25.11.2020).

9. Developer.twitter [Электронный ресурс]. – Режим доступа: https://developer.twitter.com, свободный (дата обращения: 25.11.2020).

10. Snap.stanford [Электронный ресурс]. – Режим доступа: https://snap.stanford.edu/data, свободный (дата обращения: 25.11.2020).

11. Networkrepository [Электронный ресурс]. – Режим доступа: http://networkrepository.com/soc.php, свободный (дата обращения: 25.11.2020).

12. Weibo [Электронный ресурс]. – Режим доступа: https://weibo.com, свободный (дата обращения: 25.11.2020).

13. Gephi [Электронный ресурс]. – Режим доступа: https://gephi.org, свободный (дата обращения: 25.11.2020).

14. Sna (Social Network Analysis) [Электронный ресурс]. – Режим доступа: https://cran.r-project.org/web/ packages/sna, свободный (дата обращения: 25.11.2020).

15. Ucinetsoftware [Электронный ресурс]. – Режим доступа: https://sites.google.com/site/ucinetsoftware, свободный (дата обращения: 25.11.2020).

16. Nodexl [Электронный ресурс]. – Режим доступа: https://nodexl.com, свободный (дата обращения: 25.11.2020).

17. Graphviz [Электронный ресурс]. – Режим доступа: https://www.aminer.org/data-sna, свободный (дата обращения: 25.11.2020).

18. Netminer [Электронный ресурс]. – Режим доступа: http://www.netminer.com, свободный (дата обращения: 25.11.2020).

19. Automap [Электронный ресурс]. – Режим доступа: http://www.casos.cs.cmu.edu/projects/automap, свободный (дата обращения: 25.11.2020).

20. Cytoscape [Электронный ресурс]. – Режим доступа: https://cytoscape.org, свободный (дата обращения: 25.11.2020).

21. GraphChi [Электронный ресурс]. – Режим доступа: https://github.com/GraphChi, свободный (дата обращения: 25.11.2020).

22. Networkit [Электронный ресурс]. – Режим доступа: https://networkit.github.io, свободный (дата обращения: 25.11.2020).

23. Nvidia [Электронный ресурс]. – Режим доступа: https://www.nvidia.com, свободный (дата обращения: 25.11.2020).

24. Коломейченко М.И. Алгоритм выделения сообществ в социальных сетях / М.И. Коломейченко, А.А. Чеповский, А.М. Чеповский // Фундамент. и прикл. матем. – 2014. – Т. 19, вып. 1. – С. 21–32.

25. Базенков Н.И. Обзор информационных систем анализа социальных сетей / Н.И. Базенков, Д.А. Губанов // УБС. – 2013. – Вып. 41. – С. 357–394.

26. Fortunato S. Community detection in graphs // Phys. Rep. – 2010. – Vol. 486. – P. 75–174.

27. Newman M. E. Modularity and community structure in networks // Proc. Natl. Acad. Sci. USA. – 2006. – Vol. 103. – P. 8577–8582.

28. Leskovec C. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters / J. Leskovec, K. Lang, A. Dasgupta // Internet Mathematics. – 2009. – Vol. 6(1). – P. 29–123.

29. Jaewon Y. Defining and evaluating network communities based on ground-truth / Y. Jaewon, L. Jure // Knowledge and Information Systems. – 2015. – Vol. 42, No. 1. – P. 181–213.

30. Пупырев С.Н. Визуализация структуры сообществ в графах // Системы правления и информационные технологии. – Воронеж, 2009. – № 2(36). – С. 21– 27.

31. Райгородский А.М. Модели случайных графов и их применения // Труды МФТИ. – 2010. – Vol. 2, № 4. – P. 130–140.

32. Watts D.J. Collective dynamics of ‘small-world’ networks / D.J. Watts, S.H. Strogatz // Nature. – Vol. 393(6684). – P. 440–442.

33. Bansal N. Correlation clustering / N. Bansal, A. Blum, S. Chawla // Proceedings of 43rd FOCS. – 2002. – P. 238–247.

34. Ellison N.B. Social network sites: Definition, history, and scholarship / N.B. Ellison // Journal of Computer-Mediated Communication. – 2007. – Vol. 13, No. 1. – P. 210–230.

35. Baller D. An Empirical Method for the Evaluation of Dynamic Network Simulation Methods / D. Baller, J. Lospinoso, A.N. Johnson // International Conference on Information and Knowledge Engineering. – Las Vegas, NV. – 2008. – P. 358–364.

36. Назарчук А.В. О сетевых исследованиях в социальных науках // Социологические исследования. – 2011. – № 1. – С. 39–51.

37. Newman M. The structure and dynamics of networks / M. Newman, D.J Watts. – Princeton University Press, 2006. – 596 p.

38. Barabasi A.L. Emergence of scaling in random networks / A.L. Barabasi, R. Albert // Science. – 1999. – Vol. 286(5439). – P. 509–512.

39. Метафизический смысл Big Data. 2020 [Электронный ресурс]. – Режим доступа: https://www.aminer.org/data-sna, свободный (дата обращения: 25.11.2020).

Катаев Михаил Юрьевич
Д-р техн. наук, профессор каф. автоматизированных систем управления (АСУ)
Томского государственного университета
систем управления и радиоэлектроники (ТУСУР)
Ленина пр-т, д. 40, г. Томск, Россия, 634050
Тел.: 8 (382-2) 70-15-36
Эл. почта: kataev.m@sibmail.com

Орлова Вера Вениаминовна
Д-р соц. наук, доцент, профессор,
зав. каф. философии и социологии (ФиС) ТУСУРа
Ленина пр-т, д. 40, г. Томск, Россия, 634050
Тел.: +7 (382-2) 70-15-90
Эл. почта: vera.v.orlova@tusur.ru

Kataev M.Yu., Orlova V.V.
Social media event data analysis

Social media analysis has become ubiquitous at a quantitative and qualitative level due to the ability to study content from open social networks. This content is a rich source of data for the construction and analysis of the interaction of social network users when forming various groups, used not only for statistical calculations, social areas of analysis, but also in trade or for the development of recommendation systems. The large number of social media users results in a huge amount of unstructured data (by time, type of communication, type of message and geographic location). This article aims to discuss the problem of analyzing social networks and obtaining information from unstructured data. The article discusses information extraction methods, well-known software products and datasets.
Keywords: social networks, analysis methods, datasets, network dynamics.
doi: 10.21293/1818-0442-2020-23-4-71-77

References

1. Girvan M., Newman E.J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 2002, vol. 99(12), pp. 7821–7826.

2. Gusarova N.F. *Analiz sotsial'nykh setey. Osnovnye ponyatiya i metriki* [Social media analysis. Basic concepts and metrics]. SPb: Universitet ITMO, 2016. 67 p. (in Russ.).

3. Frolov Yu.N., Gabysheva L.K. *Sotsial'nye seti: teoriya i praktika* [Social networks: theory and practice]. Tyumen, TyumGNGU, 2012. 140 p. (in Russ.).

4. Dodonov A.G., Lande D.V., Putyatin V.G. *Komp'yuternye seti i analiticheskie issledovaniya* [Computer networks and analytical research]. – K.: IPRI NAN Ukrainy, 2014. 486 p. (in Russ.).

5. Watts D.J., Dodds P.S., Newman M.E.J. (2002). Identify and search in social networks. *Science*, vol. 296, pp. 1302–1305.

6. McCallum A., Wang X., Corrada-Emmanuel A. Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. *Journal of Artificial Intelligence Research*., 2007, vol. 30, pp. 249–272.

7. Webscraper.io [Electronic resource]. Access mode: https://webscraper.io, free (Accessed: November 25, 2020).

8. Developers.facebook [Electronic resource]. Access mode: https://developers.facebook.com, free (Accessed: November 25, 2020).

9. Developer.twitter [Electronic resource]. Access mode: https://developer.twitter.com, free (Accessed: November 25, 2020).

10. Snap [Electronic resource]. Access mode: https://snap.stanford.edu/data, free (Accessed: November 25, 2020).

11. Networkrepository [Electronic resource]. Access mode: http://networkrepository.com/soc.php, free (Accessed: November 25, 2020).

12. Weibo [Electronic resource]. Access mode: https://weibo.com, free (Accessed: November 25, 2020).

13. Gephi [Electronic resource]. Access mode: https://gephi.org, free (Accessed: November 25, 2020).

14. Sna (Social Network Analysis) [Electronic resource]. Access mode: https://cran.r-project.org/web/packages/sna, free (Accessed: November 25, 2020).

15. Ucinetsoftware [Electronic resource]. Access mode: https://sites.google.com/site/ucinetsoftware, free (Accessed: November 25, 2020).

16. Nodexl [Electronic resource]. Access mode: https://nodexl.com, free (Accessed: November 25, 2020).

17. Graphviz [Electronic resource]. Access mode: https://www.aminer.org/data-sna, free (Accessed: November 25, 2020).

18. Netminer [Electronic resource]. Access mode: http://www.netminer.com, free (Accessed: November 25, 2020).

19. Automap [Electronic resource]. Access mode: http://www.casos.cs.cmu.edu/projects/automap, free (Accessed: November 25, 2020).

20. Cytoscape [Electronic resource]. Access mode: https://cytoscape.org, free (Accessed: November 25, 2020).

21. GraphChi [Electronic resource]. Access mode: https://github.com/GraphChi, free (Accessed: November 25, 2020).

22. Networkit [Electronic resource]. Access mode: https://networkit.github.io, free (Accessed: November 25, 2020).

23. nvidia [Electronic resource]. Access mode: https://www.nvidia.com, free (Accessed: November 25, 2020).

24. Kolomeychenko M.I., Chepovskiy A.A., Chepovskiy A.M. Algoritm vydeleniya soobshchestv v sotsial'nykh setyakh [Algorithm for highlighting communities in social networks]. *Fundament. i prikl. matem*., 2014, vol. 19, no. 1, pp. 21–32 (in Russ.).

25. Bazenkov N.I., Gubanov D.A. *Obzor informatsionnykh sistem analiza sotsial'nykh setey* [Review of information systems analysis of social networks]. *UBS*, 2013, vol. 41, pp. 357–394 (in Russ.).

26. Fortunato S. Community detection in graphs. *Phys. Rep*., 2010, vol. 486, pp. 75–174.

27. Newman M.E. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 2006, vol. 103, pp. 8577–8582.

28. Leskovec J., Lang K., Dasgupta A. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. *Internet Mathematics*, 2009, vol. 6(1), pp. 29–123.

29. Jaewon Y., Jure L. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 2015, vol. 42, no. 1, pp. 181–213.

30. Pupyrev S.N. *Vizualizatsiya struktury soobshchestv v grafakh* [Community structure visualization in graphs]. *Sistemy pravleniya i informatsionnye tekhnologii*. Voronezh, 2009, vol. 2(36), pp. 21–27 (in Russ.).

31. Raygorodskiy A.M. *Modeli sluchaynykh grafov i ikh primeneniya* [Random graph models and their applications]. *Trudy MFTI*, 2010, vol. 2, no. 4, pp. 130–140 (in Russ.).

32. Watts D.J., Strogatz S. H. Collective dynamics of «small-world» networks. *Nature*, 1998, vol. 393(6684), pp. 440–442.

33. Bansal N., Blum A., Chawla S. Correlation clustering *Proceedings of 43rd FOCS*, 2002, pp. 238–247.

34. Ellison N.B. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 2007, vol. 13, no. 1, pp. 210–230.

35. Baller D., Lospinoso J., Johnson A.N. An Empirical Method for the Evaluation of Dynamic Network Simulation Methods. *International Conference on Information and Knowledge Engineering, Las Vegas, NV*,2008. pp. 358–364.

36. Nazarchuk A.V. *O setevykh issledovaniyakh v sotsial'nykh naukakh* [About network research in the social sciences]. *Sotsiologicheskie issledovaniya*, 2011, no. 1, pp. 39–51 (in Russ.).

37. Newman E.J., Watts D.J. *The structure and dynamics of networks*. Princeton University Press, 2006, 596 p.

38. Barabasi A.L., Albert R. Emergence of scaling in random networks. *Science*, 1999, vol. 286(5439), pp. 509–512.

39. Metaphysical meaning of Big Data. [Electronic resource]. https://www.aminer.org/data-sna, Access mode free (date of access: 25.11.2020) (Accessed: November 25, 2020) (in Russ.).

Mikhail Yu. Kataev
Doctor of Engineering Sciences, Professor,
Department automated control systems (ACS)
Tomsk State University of Control Systems
and Radioelectronics (TUSUR)
40, Lenin pr., Tomsk, Russia, 634050
Phone: +7 (382-2) 70-15-36
Email: kataev.m@sibmail.com

Vera V. Orlova
Doctor of Sciences in Sociology, Associate Professor,
Head of the Department of Pholosophy and Sociology,
TUSUR
40, Lenin pr., Tomsk, Russia, 634050
Phone: +7 (382-2) 70-15-90
Email: vera.v.orlova@tusur.ru