Математическое, алгоритмическое и программное обеспечение

DOI 10.24412/2221-2574-2025-1-41-47 УДК 004.942; 303.732.4

АЛГОРИТМ ПОИСКА СПОРАДИЧЕСКОГО КОНТЕКСТНОГО СООБЩЕСТВА В СОЦИАЛЬНЫХ СЕТЯХ ИНТЕРНЕТА

Монахов Михаил Юрьевич
доктор технических наук, профессор, заведующий кафедрой информатики и защиты информации ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых».
E-mail: mmonakhov@vlsu.ru

Толокнов Егор Альбертович
аспирант кафедры информатики и защиты информации ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых».
E-mail: tolegork@mail.ru

Матвеева Екатерина Александровна
студент кафедры информатики и защиты информации ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых».
E-mail: eamatveeva16@mail.ru

Адрес: 600000, Российская Федерация, г. Владимир, ул. Горького, д. 87.

Аннотация: В статье предложен подход к поиску спорадических контекстных сообществ в социальных сетях Интернета. Разработаны алгоритм и программное обеспечение поиска агентов сообщества и их взаимосвязей, сочетающий в себе атрибутивный метод идентификации агентов, контекстно-ориентированный анализ сообщений, графовый метод оценки их взаимодействия. Реализация предложенного подхода позволит автоматизировать процесс выделения целевых групп пользователей в Интернете для аналитических нужд, распространения таргетированной информации, идентификации конструктивного и деструктивного влияния.

Ключевые слова: социальные сети, спорадические сообщества, идентификация пользователей, контекстный поиск, алгоритм.

Введение

Объектом в данной работе является социальная сеть (СС) в Интернете — веб-сайт, предназначенный для построения, отражения и организации взаимодействия агентов (пользователей) друг с другом [1].

Агент — субъект/пользователь СС, независимый источник информационных сообщений. Агент взаимодействует с другими агентами (передаёт сообщения, содержащие мнение, которое отражает суждение, высказывание, оценку об интересующем его объекте или явлении), проявляет независимое поведение, которое может рассматриваться как следствие его знаний, взаимодействия с другими агентами и целей, которых он желает достичь. Технически агент представлен в сети *узлом* — человеко-машинным компонентом СС, который хранит и отображает информационные сообщения как агента, которому принадлежит данный узел и сообщения других агентов.

Предметом анализа является «информационное поле», создаваемое узлами СС, доступное наблюдателю в виде потока информационных сообщений, генерируемых источниками информации — агентами.

Спорадические (единичные, непостоянные, случайные, проявляющиеся нерегулярно) сообщества в социальных сетях Интернета обозначают группы пользователей (агентов), которые не объединяются в явной форме, но проявляют общие характеристики и поведение. Эти сообщества могут быть обнаружены путём анализа данных пользователей, которые выявляют схожие шаблоны и структуры. Характеристики для определения неявных сообществ могут включать общие интересы, предпочтения, поведение в онлайн-среде и другие аспекты.

Спорадическое контекстное сообщество (СКС) — структурная единица социальной сети, состоящая из множества однородных (возраст, место жительства, пол, образование и др.) агентов, взаимодействующих путём обмена информационными сообщениями по определённой теме (контексте) и обладающая свойством связности.

Пользователи, взаимодействуя друг с другом, как правило, не знают, что являются агентами СКС, это сообщество существует, как структурная единица, только в видении наблюдателя.

Для изучения и анализа СКС могут использоваться различные методы, среди которых — анализ графовой структуры и атрибутов вершин. В предложенном алгоритме реализован гибридный подход, который сочетает особенности ранее известных методов, что покрывает их недостатки в виде неучтённого контекста, отсутствия динамики, неполноты атрибутов, различных ограничений. В гибридном подходе формируется граф СКС с учётом его структуры (взаимодействие), отбора агентов по параметрам (атрибутивность), контекста сообщений (семантика).

Реализация предложенного подхода позволит автоматизировать процесс выделения таких целевых групп людей для аналитических нужд, распространения таргетированной информации, идентификации конструктивного и деструктивного влияния в сообществах.

Анализ публикаций

В [2, 3] предложены подходы к анализу социальных сетей, включая стратегии сбора данных. Анализ социальных сетей с использованием теории графов, включая характеристики близости, уровня доверия, кластеризации с учётом различных метрик содержатся в работах [4–6]. В работе [7] был представлен «метод Галактик», который основан на последовательном выделении пересекающихся сообществ на исходном взвешенном графе, дальнейшем построении нового графа, в котором вершинами являются выделенные на первом шаге сообщества, называемые авторами «метавершинами». Выделение пересекающихся сообществ и сообществ с использованием атрибутов вершин исследовалось в работе [8]. Проблема отсутствия атрибутов вершин исследовалась в работе [9]. В работах [10–13] представлены эвристические подходы к идентификации сообществ, основанные на алгоритмах модульной оптимизации. Алгоритмы применяются к «кольцам», состоящим из кликов (групп узлов).

Выделим отдельные особенности и закономерности сообществ: сужающийся диаметр и транзитивность сети, существование небольшого количества влиятельных участников с большим количеством связей и большого количества обычных пользователей с примерно равномерным распределением связей между ними. В рассмотренных работах предлагались методы анализа, касающиеся какого-то одного аспекта сообществ — его графовой структуры, либо параметров.

Алгоритм поиска спорадического контекстного сообщества

Целью данной работы является решение задачи поиска спорадического контекстного сообщества в социальной сети Интернета.

Исходными данными для задачи являются:

– перечень параметров для профильной идентификации агентов СКС;

– контекстно-ориентированный словарь для анализа сообщений.

Задача состоит в поиске множества пользователей – агентов с характеристиками: профиль, взаимодействие в заданной тематике (контексте), связность между ними.

Для решения данной задачи предложен алгоритм, сочетающий в себе атрибутивный метод идентификации агентов, контекстно-ориентированный анализ сообщений, графовый метод оценки их взаимодействия.

Профиль $P_0 = \{p_1, p_2, \ldots, p_K\}$ агента синтезируемого (формируемого) СКС. Здесь $\{p_1, p_2, \ldots, p_K\}$ — множество критериев (параметров) отбора пользователей в агенты. В контексте решаемой задачи такими параметрами могут быть: возраст (p_1), место жительства (p_2), образование (p_3). Кроме того, чтобы стать полноправным (действительным) агентом, необходимо наличие *встреч* с другими пользователями (друзьями).

Под *встречей* понимается одиночное (элементарное) событие, отражающее взаимодействие двух пользователей (u_i, u_k), так что u_i выполняет действие d_s, а u_k в ответ на d_s выполняет действие d_r. Здесь $d_s, d_r \in D$, D — множество допустимых действий. $\propto(u_i, u_k)$ — результат встречи, обозначающий факт взаимодействия $\propto(u_i, u_k) = 1$ или его отсутствие $\propto(a_i, a_k) = 0$. Например, u_i опубликовал пост (или другое любое сообщение) на своей странице, u_k ответил на него комментарием или лайком (d_r), такое событие считается одной встречей. Если было несколько ответов на d_s, то имеем несколько встреч.

Информационная модель пользователя.

1. Данные профиля пользователя социальной сети. При создании профиля, например, ВКонтакте пользователю предлагается заполнить следующие разделы: основное, контакты (друзья), интересы, образование, карьера, военная служба, жизненная позиция. Ряд позиций являются обязательными, другие — нет.

2. Сообщения пользователя (способы обмена информацией)

- пост. Пост в социальных сетях — это запись, оставленная на стене личного аккаунта, то есть, это короткое сообщение. Посты содержат мысли автора, передают его идеи и эмоции по интересующему его вопросу. Опубликовав пост на своей странице в социальной сети, агент высказывает своё мнение всем своим друзьям;

- репост. Репост — возможность поделиться чужой публикацией на своей странице, оставляя её в первоначальном виде с сохранением ссылки на первоисточник. В рассматриваемой задаче понятия «пост» и «репост» эквивалентные;

- комментарий на пост. В этом случае пользователь социальной сети вступает с автором в дискуссию по поводу содержания публикации, которую разместили;

- комментарий на комментарий. То же самое, что и в предыдущем пункте, но ответная реакция уже не на пост, а на предыдущий комментарий;

- лайк на пост (репост), лайк на комментарий. Лайк — форма одобрения, представляющая собой кнопку с изображением руки и поднятым вверх большим пальцем, сердечка, звёздочки. В рассматриваемой задаче считается ответным сообщением.

Алгоритм

Шаг 1. Задаются:

- профиль P_0 агента синтезируемого (формируемого) СКС;

- контекстный словарь слов и выражений.

Шаг 2. Поиск первого кандидата в агенты СКС — пользователя (узла сети) с заданным профилем. Поиск случайный, вручную. Если такой найден, то он становится текущим кандидатом в агенты, переход к шагу 3, иначе повторять шаг 2.

Шаг 3. Проверяется наличие друзей у текущего кандидата в агенты.

Получить из информации на пользователя множество (список) пользователей – друзей этого агента. Если друзей нет, то заменить текущего агента, перейти к шагу 2.

Шаг 4. Просматривая информацию на пользователей — друзей текущего агента поочерёдно, оставляем из них только «профильных», формируя множество кандидатов в действительные агенты СКС. Если у текущего агента не оказалось ни одного «профильного» друга, то он не может быть кандидатом в действительные агенты (КДА). Если при этом текущий агент был найден впервые, то перейти к шагу 2 (заново ищем первого).

Шаг 5. Просматривая поочерёдно элементы множества КДА (начиная со второго), находим их «профильных» друзей, добавляя их в множество КДА (если их нет в данном множестве). Такой поиск в ширину может завершиться, когда закончатся уникальные профильные «друзья друзей». Если их слишком много, то данные действия продолжаются, пока не достигнут некий предел мощности данного множества. В результате будут найдены все «профильные» КДА.

Шаг 6. На данном этапе удаляются кандидаты в действительные агенты СКС, у которых количество встреч с другими КДА меньше наперёд заданного значения. Для каждого кандидата КДА, просматривая поочерёдно его сообщения и ответные сообщения других кандидатов, подсчитывается количество «встреч» в соответствии приведённым выше правилом. Анализируются следующие варианты встреч: «пост КДА – комментарий другого КДА на пост», «пост КДА — лайк другого КДА на пост», «комментарий КДА – на комментарий другого КДА», «комментарий КДА — лайк другого КДА на комментарий». На этом же этапе формируется матрица связей КДА друг с другом. Сила связей определяется количеством встреч.

Шаг 7. Проверяются все сообщения отобранных (оставшихся) действительных агентов на соответствие тематике (контексту), оставляя лишь тех, кто посылал сообщения со словами и выражениями (совокупность слов) из заданного словаря. Задаётся минимальное количество совпадений слов и выражений из сообщений КДА словам и выражениям из словаря, на основании которого принимается решение о принадлежности сообщений контексту. Получить информацию о сообщениях КДА. Сообщения КДА могут находиться как у данного КДА, так и у других КДА, с которыми он взаимодействовал. Распаковываем каждое сообщение в слова, удаляя предлоги, знаки препинания и т.п. Каждое слово и выражение сравниваем со словарём. Если количество совпадений превышает задаваемый порог, то такой КДА остаётся элементом множества КДА, иначе КДА удаляется из множества КДА, также удаляются его связи в матрице связей КДА друг с другом.

Шаг 8. Проверяется связность вершин графа, образованного КДА. Удаляются все несвязные вершины (КДА) и их связи. Остаётся множество действительных агентов СКС.

Конец алгоритма.

Эксперимент

Экспериментальное исследование алгоритма состояло из нескольких этапов:

1. Разработано программное обеспечение реализации алгоритма (на языке python, библиотека requests). Использовалось API Вконтакте, СУБД с SQL сервером в Microsoft SQL Server Management Studio)
2. Задавался профиль СКС (возраст агентов 21–41 лет, место проживания — г. Владимир и Владимирская область, образование неполное высшее (учится в вузе) и высшее). Контекстный словарь включал около 200 слов и выражений по тематике «Встречаем Новый год». Количество совпадений по словарю — не менее 3, число встреч между КДА — не менее 2.

Эксперимент в части сбора исходных данных продолжался в течение месяца. Анализировалась СС Вконтакте.

На первом этапе (шаги 2–4) было отобрано 131 КДА. Из них «профильных» оказалось 43 (шаг 5). КДА с требуемым количеством встреч (шаг 6) осталось 28, а участвующих в тематической переписке (шаг 7) 21. Связность графа была обеспечена у 17 агентов.

На основе полученных в ходе эксперимента данных, был сформирован граф СКС (рис. 1). Узлы имеют номера КДА, отобранные в агенты СКС, связи означают выявленные встречи.

В результате применения алгоритма к экспериментальной выборке кандидатов, были получены связи в сообществе, состоящем из 21 обнаруженного агента, где между агентами имеется как минимум 2 взаимодействия.

Для оценки точности алгоритма шаги 6–8 были проделаны вручную. Результат оказался — 25 агентов СКС. Не найденные встречи изображены пунктирной линией.

Анализ результатов

1. Все вместе найденные агенты не принадлежали ни одному известному открытому сообществу.
2. Точность выделения СКС (около 70%) невысокая. Повлияли факторы поиска выражений по словарю: в реальных сообщениях содержится молодёжный сленг, который программа не распознает, а аналитик делает это весьма уверенно.
3. Не было «пропуска цели». Все найденные программой агенты, как подмножество, принадлежали множеству найденных вручную.
4. Среди КДА были отмечены «одноразовые» встречи, видимо, из-за краткосрочности наблюдения. Следует ожидать повышения точности с увеличением времени наблюдения, хотя здесь могут добавляться не учтённые КДА.

Подтвердилась гипотеза о наличии небольшого количества влиятельных участников с большим количеством связей.

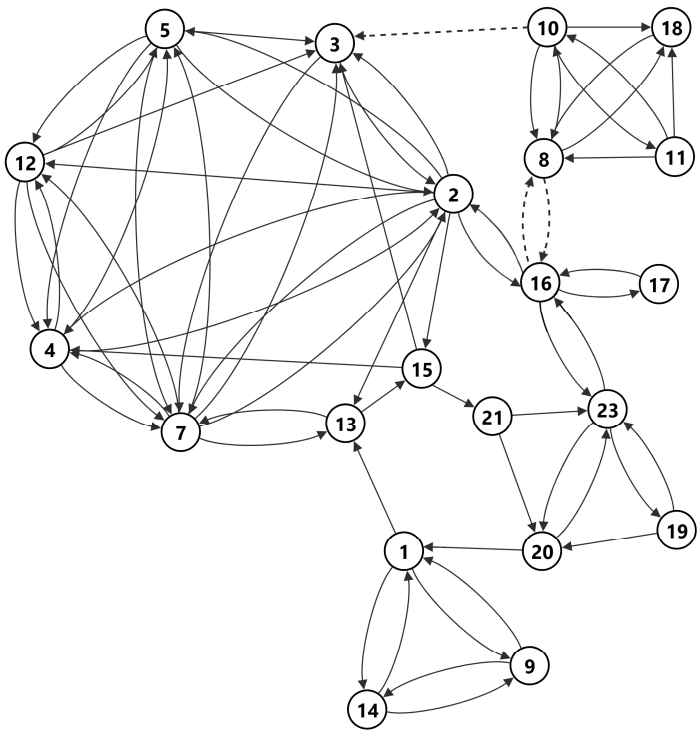

Рис. 1. Граф спорадического контекстного сообщества

Литература

1. *Губанов Д.А., Новиков Д.А., Чхартишвили А.Г.* Социальные сети: модели информационного влияния, управления и противоборства. М.: Физматлит, 2010. 334 с.

2. *Чураков А.Н.* Анализ социальных сетей // Социологические исследования (Социс). 2001. №1. С. 109–121.

3. *Батура Т.В.* Модели и методы анализа компьютерных социальных сетей // Программные продукты и системы. 2013. №3. С. 130–137.

4. *Hanneman R.A., Riddle M.* Introduction to Social Network Methods. Riverside: University of California, 2005. 322 p.

5. *Aggarwal C.C.* Social Network Data Analytics. New York: Springer, 2011. 502 p.

6. *Radicchi F., Castellano C., Cecconi F., Loreto V., Parisi D.* Defining and identifying communities in networks // Proceedings of the National Academy of Sciences. 2004. DOI: 10.1073/pnas.0400054101.

7. *Попов В.А., Чеповский А.А.*

Выделение неявных пересекающихся сообществ на графе взаимодействия Telegram-каналов с помощью «метода Галактик» // Труды Института системного анализа Российской академии наук. 2022. Т. 72, № 4. С. 39–50.

8. *Chesnokov V.O., Klyucharev P.G.* Social Graph Community Differentiated by Node Features with Partly Missing Information // Science and Education of the Bauman MSTU. 2015. No. 9. Pp. 188–199.

9. *Pun C.S.* Research on Multivariate Statistical Analysis with Missing Data. 2011. DOI: 10.13140/RG.2.2.32674.22721.

10. *Гончаров М.* Выявление неявных сообществ в социальных сетях [Электронный ресурс]. 2012. URL: http://businessdataanalytics.ru/CommunityDetection.htm (дата обращения: 17.04.2024).

11. *Blondel V.D. Guillaume J.-L., Lambiotte R., Lefebvre E.* Fast unfolding of communities in large networks // Journal of Statistical Mechanics: Theory and Experiment. 2008. DOI: 10.1088/1742-5468/2008/10/P10008.

12. *Fortunato S.* Community detection in networks: A user guide // Physics Reports. 2016. Vol. 659(12). DOI:10.1016/j.physrep.2016.09.002

13. *Barthélemy M., Fortunato S.* Resolution limit in community detection // Proceedings of the National Academy of Sciences. 2007. DOI: 10.1073/pnas.0605965104.

Поступила 17 апреля 2024 г.

English

ALGORITHM FOR SEARCHING FOR SPORADIC CONTEXTUAL COMMUNITY IN INTERNET SOCIAL NETWORKS

Mikhail Yuryevich Monakhov — Grand Dr. in Engineering, Professor, the Head of Department of Computer Science and Information Security, Federal State Budgetary Educational Institution of Higher Education “Vladimir State University named after A.G. and N.G. Stoletovs”.
E-mail: mmonakhov@vlsu.ru

Egor Albertovich Toloknov — Postgraduate Student, Department of Computer Science and Information Security, Federal State Budgetary Educational Institution of Higher Education “Vladimir State University named after A.G. and N.G. Stoletovs”.
E-mail: tolegork@mail.ru

Ekaterina Aleksandrovna Matveeva — Student, Department of Computer Science and Information Security, Federal State Budgetary Educational Institution of Higher Education “Vladimir State University named after A.G. and N.G. Stoletovs”.
E-mail: eamatveeva16@mail.ru

Address: 600000, Russian Federation, Vladimir, Gorky St., 87.

Abstract: The presented paper describes an approach to finding sporadic contextual communities in Internet social networks. The developed algorithm and software combine an attribute-based method for agent identification, context-based message analysis, and a graph-based method for evaluating their interactions. This approach will automate the selection of target user groups on the Internet for analytical purposes, dissemination of targeted information, and identification of influence in communities. The paper highlights the importance of sporadic contextual communities in social networks that share common characteristics and behaviors but are not explicitly unified. The proposed approach relies on analyzing graph structure and vertex attributes, taking into account message context, dynamics, and other constraints. Using the hybrid approach in analyzing sporadic communities will effectively identify target user groups, enhance analytical research capabilities, and identify constructive and destructive influences in communities. The developed method is an innovative tool for automating the process of analyzing social networks and identifying key communities for further research and action. Analysis of the results showed that all agents found did not belong to known communities, the accuracy of SCS identification was about 70%. One-time meetings were noted, probably due to the short-term nature of the observation, but the hypothesis about the presence of influential participants with a large number of connections was confirmed. The experimental study identified key features of user behavior in the social network and confirmed the hypothesis of sporadic contextual communities. The results of the study can be used to improve social network analysis algorithms and increase the accuracy of key communities’ identification.

Keywords: social networks, sporadic communities, user identification, contextual search.

References

1. *Gubanov D.A., Novikov D.A., Chkhartishvili A.G.* Social networks: models of information influence, management and confrontation. Moscow: Fizmatlit, 2010. 334 p.
2. *Churakov A.N.* Analysis of social networks. Sotsiologicheskie issledovaniya (Sotsis). 2001. No. 1. Pp. 109–121.
3. *Batura T.V.* Models and methods of analysis of computer social networks. Program products and systems. 2013. No. 3. Pp. 130–137.
4. *Hanneman R.A., Riddle M.* Introduction to Social Network Methods. Riverside: University of California, 2005. 322 p.
5. *Aggarwal C.C.* Social Network Data Analytics. New York: Springer, 2011. 502 p.
6. *Radicchi F., Castellano C., Cecconi F., Loreto V., Parisi D.* Defining and identifying communities in networks. Proceedings of the National Academy of Sciences. 2004. DOI: 10.1073/pnas.0400054101.
7. *Popov V.A., Chepovsky A.A.* Identification of implicit intersecting communities on the graph of interaction of Telegram channels using the "Galaxy method". Trudy Instituta sistemnogo analiza Rossiyskoy akademii nauk (ISA RAN). 2022. Vol. 72, No. 4. Pp. 39–50.
8. *Chesnokov V.O., Klyucharev P.G.* Social Graph Community Differentiated by Node Features with Partly Missing Information. Science and Education of the Bauman MSTU. 2015. No. 9. Pp. 188–199.
9. *Pun C.S.* Research on Multivariate Statistical Analysis with Missing Data. 2011. DOI: 10.13140/RG.2.2.32674.22721.
10. *Goncharov M.* Identification of implicit communities in social networks [Electronic source]. 2012. URL: http://businessdataanalytics.ru/CommunityDetection.htm (Access date: 17.04.2024).
11. *Blondel V.D. Guillaume J.-L., Lambiotte R., Lefebvre E.* Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008. DOI: 10.1088/1742-5468/2008/10/P10008.
12. *Fortunato S.* Community detection in networks: A user guide. Physics Reports. 2016. Vol. 659(12). DOI:10.1016/j.physrep.2016.09.002
13. *Barthélemy M., Fortunato S.* Resolution limit in community detection. Proceedings of the National Academy of Sciences. 2007. DOI: 10.1073/pnas.0605965104.