Выделение неявных пересекающихся сообществ на графе взаимодействия Telegram-каналов с помощью «метода Галактик»

В.А. Попов, А.А. Чеповский

Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия

Аннотация. В работе представлен «метод Галактик» для выделения неявных сообществ на графе взаимодействующих объектов, полученном при импорте сети каналов из мессенджера Telegram. Данный метод 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». Взвешенные ребра нового графа между «метавершинами» строятся исходя из весов между каждой парой вершин исходного графа. Далее на новом графе выделяются непересекающиеся сообщества. В итоге получается разбиение исходного графа на пересекающиеся «метасообщества». Для оценки качества разбиения представленным методом проведен психолингвистический анализ текстов полученных метасообществ, выделены закономерности в зависимости от тематической направленности каналов внутри метасообщества. В итоге обработки и анализа полученных метасообществ получено подтверждение качества разбиения. Сочетание алгоритмического метода выделения сообществ и психолингвистического анализа имеет практическое применение для задач анализа информационного воздействия в социальных сетях.

Ключевые слова: *Telegram, анализ социальных сетей, импорт данных из социальных сетей, модель информационного воздействия, граф взаимодействующих объектов, выделение сообществ, психолингвистический анализ текстов.*

DOI: 10.14357/20790279220405

Введение

Тема изучения графов взаимодействующих объектов, в том числе полученных из социальных сетей, давно и активно изучается [1-5]. Но в настоящее время в мире крайне интенсивно растет популярность не только социальных сетей, но и кроссплатформенных систем обмена сообщениями, в частности – мессенджера Telegram. Он активно используется не только для обмена личными сообщениями между пользователями, но и как площадка для профессиональных каналов (например, для СМИ, медиа, ведения личных блогов), которые издают регулярные посты-публикации разной направленности. За счет удобной возможности репостов, цитирования и упоминания других каналов в Telegram осуществляется быстрое распространение информации. Все эти возможности позволяют рассматривать множество Telegram-каналов как граф взаимодействующих объектов.

Одним из методов анализа таких графов является подход по выявлению групп общения пользователей путем выделения неявных сообществ [6-9]. Стоит отметить, что многие каналы в мессенджере Telegram не имеют единую тематику постов или отражают различные взгляды и интересы, поэтому Telegram-каналы могут относиться к нескольким неявным сообществам сети. Вследствие этого необходимо применять алгоритмы для выделения именно пересекающихся сообществ на графах (одна вершина может относиться к нескольким сообществам).

В данной работе представлен «метод Галактик», позволяющий выделить метасообщества на графе взаимодействующих объектов, полученном при импорте данных из Telegram-каналов. Метод состоит в последовательном выделении пересекающихся сообществ на взвешенном графе, формировании и обработке полученного

нового графа, где вершинами являются полученные ранее сообщества, и последующем выделении непересекающихся сообществ на этом новом графе. В итоге получается разбиение исходного графа на пересекающиеся метасообщества. По аналогии с (*F,L,C,R*)-моделью для Twitter, введенной авторами в работе [10], для импорта данных из мессенджера Telegram была использована (*U,M,R*)-модель, построенная и описанная авторами в работе [11]. Для оценки качества полученного разбиения производится психолингвистический анализ текстов полученных сообществ.

1. Импорт данных из мессенджера Telegram

Представляемый «метод Галактик» применяется на взвешенных графах взаимодействующих объектов, полученных авторами при импорте реальных данных из Telegram-каналов с использованием разработанной ранее (*U,M,R*)-модели информационного взаимодействия [11]. В мессенджере Telegram были выделины три вида взаимоотношений между каналами: наличие общих внешних URL, упоминания и репосты. После импорта данных строится взвешенный граф, в котором вершинами являются Telegram-каналы, а наличие ребер между ними и веса определяются имевшими место взаимодействиями.

Пусть $G(V,E)$ – граф, у которого V – множество Telegram-каналов, а E – множество всех возможных ребер – взаимодействий между парами каналов. На данном множестве ребер E определим весовую функцию $w(e_{AB})$ ($e_{AB} \in E;\ A,B \epsilon V$) следующим образом:

$$w(e_{AB}) = U \times \delta_{e_{AB}}^{U} + M \times \delta_{e_{AB}}^{M} + R \times \delta_{e_{AB}}^{R},$$

где $\delta_{e_{AB}}^{U}$ – количество общих уникальных внешних ссылок (URL) в постах у каналов A и B за выбранный период; $\delta_{e_{AB}}^{M}$ – количество постов, где в тексте канал A упомянул канал B плюс количество постов, где B упомянул A за выбранный период (для каждого поста смотрятся уникальные упоминания, то есть если в одном посте канал A упомянул несколько раз канал B, то вклад в коэффициент $\delta_{e_{AB}}^{M}$ будет только +1); $\delta_{e_{AB}}^{R}$ – количество репостов канала A сообщений канала B, плюс количество репостов канала A у канала B за выбранный период.

Вычисление весов ребер применяется для каждой пары Telegram-каналов из числа вершин графа. Если весовая функция $w(e_{AB})$ для двух каналов равна нулю, то ребра между вершинами нет, если больше нуля, то формируется ребро с посчитанным весом. Таким образом, на основе множества Telegram-каналов V и значений весовой функции $w(e_{AB})$ для каждой пары вершин мы формируем множество ребер $\tilde{E}$. Соответственно, полученный граф $G(V,\tilde{E})$ – итоговый взвешенный граф взаимодействующих объектов.

Сформированный на основе (*U,M,R*)-модели взвешенный граф $G(V,\tilde{E})$ взаимодействующих объектов сохраняется в специальном XML-подобном формате AVS (специальный формат для хранения графов, в котором описываются атрибуты каждой вершины и каждого ребра). Такое хранение графа позволяет сохранять отдельно атрибуты с текстами каналов, которые потом будут использованы авторами для анализа качества разбиения графа на метасообщества.

2. Алгоритм метода Галактик для выделения неявных сообществ

Представим детализированный метод «Галактик» для выделения неявных сообществ на графах. Алгоритм метода состоит из четырех шагов.

Шаг 1. На первом шаге метода с помощью одного из алгоритмов выделяем неявные пересекающиеся сообщества. После выполнения первого шага мы получаем разбиение графа $G(V,\tilde{E})$ на пересекающиеся сообщества.

Шаг 2. На втором шаге надо «убрать мусор» – удалить сообщества, состоящие из одной вершины. В итоге выполнения этой операции мы получаем граф $\widetilde{\boldsymbol{G}}$.

Шаг 3. Формируются метавершины на графе $\widetilde{\boldsymbol{G}}$ как выделенные на первом шаге и оставшиеся после второго шага сообщества. Получается новый взвешенный граф $\tilde{\tilde{G}}$, где каждая метавершина представляет одно из выделенных ранее сообществ (и не удаленных на втором шаге). Вес между двумя метавершинами при этом равен сумме всех весов между каждой парой вершин исходного графа из этих метавершин.

Шаг 4. Для графа $\tilde{\tilde{G}}$ применяется алгоритм выделения неявных непересекающихся сообществ. Получаем разбиение метавершин на метасообщества. Так как каждая метавершина – это набор исходных вершин (Telegram-каналов), то получаем новое разбиение исходного графа $G(V,\tilde{E})$ на неявные пересекающиеся сообщества для исходного графа.

Для графов из мессенджера T*elegram* на Шаге 1 был использован метод *Connected Iterative Scan* (*CIS*) [12]. Данный алгоритм основан на принципе увеличения показателя плотности. На Шаге 4 для графов из мессенджера Telegram авторы использовали алгоритм *Louvain* [13].

3. Применение «метода Галактик» на реальных данных

Рассмотрим на примере работу описанного метода. Сначала с помощью разработанного программного обеспечения были импортированы данные от финансового канала *@infernal_money* на глубину 3. Рассматриваемым периодом дат были выбраны две недели с 8.12.2021 по 22.12.2021. Далее при построении графа по (*U,M,R*)-модели были выбраны коэффициенты (1, 2, 3) информационного взаимодействия. Полученный граф был сформирован и сохранен в *AVS*-файл, также были скачены тексы постов у всех полученных каналов. В результате получен граф G_1, содержащий 625 вершин и 6137 ребер.

На первом шаге метода к графу был применен алгоритм выделения неявных пересекающихся сообществ *CIS* [12]. В результате было выделено 430 пересекающихся сообществ, из которых 332 содержали только одну вершину. Эти одиночные сообщества на втором шаге были удалены из дальнейшего рассмотрения. После этого на третьем шаге в полученном графе формируем метавершины (объединяем каждое сообщество в вершину). И на четвертом шаге применяем алгоритм разбиения на неявные сообщества Louvain [13] к полученному графу метавершин. В результате данного разбиения получили 17 неявных сообществ.

Аналогично описанным действиям из Telegram были импортированы данные о взаимодействии каналов и построены еще два графа. Для всех графов применялись единые коэффициенты (1, 2, 3) в (*U,M,R*)-модели и рассматривался один и тот же промежуток времени с 8.12.2021 по 22.12.2021, за который были построены графы взаимодействия и скачаны тексты. Таким образом получено три графа:

- G_1 с начальной вершиной @infernal_money и глубиной скачивания, равной 3;
- G_2 с начальной вершиной @sportsru и глубиной скачивания, равной 5;
- G_3 с начальной вершиной *@stranavozmojnostey* и глубиной скачивания, равной 3.

Далее по «методу Галактик» построены разбиения всех графов на метасообщества (табл. 1)

Для решения задач визуализации графов больших размеров выделение метасообществ позволяет сокращать количество элементов визуализации. Количество вершин для визуализации характеризуется уменьшением для графа G_1 в 37 раз, для графа G_2 в 30 раз, для графа G_3 в 37 раз. Подразумевается, что группа вершин заменяется при визуализации одной, которая рассматривается как элемент интерфейса, позволяющий «раскрывать» и показывать всю группу вершин и их связи.

С помощью дополнительных разработанных программных средств для каждого выделенного неявного сообщества графа были объединены тексты постов Telegram-каналов, которые входят в эти сообщества, за выбранный промежуток времени (8.12.21-22.12.21). Это позволяет провести психолингвистический анализ текстов не только отдельно взятых каналов, но и сообществ в целом.

4. Психолингвистический анализ текстов полученных сообществ

Для проверки эффективности разбиения каждого из построенных «методом Галактик» графов выделенные на них сообщества были проанализированы, посчитаны психолингвистические характеристики по методике, описанной в [15]. Для каждого сообщества в отдельности и для всех текстов графа в целом был посчитан ряд характеристик текстов:

- Коэффициент лексического разнообразия 1 (ЛР1) – отношение числа уникальных лексем к числу словоупотреблений.
- Коэффициент лексического разнообразия 2 (ЛР2) – коэффициент разнообразия по псевдоосновам; отношение числа уникальных псевдооснов к числу словоупотреблений.
- Коэффициент глагольности (СГ)-отношение количества глаголов и глагольных форм (причастий и деепричастий) к общему количеству всех словоупотреблений.

Табл. 1

Графы взаимодействующих объектов сети Telegram

Исходный граф	Число вершин в графе $G(V,\tilde{E})$	Число ребер в графе $G(V,\tilde{E})$	Количество выделенных на Шаге 1 сообществ на графе $G(V,\tilde{E})$	Количество оставшихся после Шага 2 сообществ на графе $\tilde{G}$	Количество выделенных на Шаге 4 метасообществ на графе $\tilde{\tilde{G}}$
G_1	625	6137	430	98	17
G_2	590	4352	369	94	20
G_3	773	6611	511	127	21

Табл. 2

Психолингвистические характеристики текстов для трех графов

№	Характеристика	G_1	G_2	G_3
	Общий объем всех текстов (МБ)	30,4	26,9	39,4
I	Коэффициент лексического разнообразия 1 (ЛР1)	0,03	0,03	0,03
II	Коэффициент лексического разнообразия 2 (ЛР2)	0,04	0,04	0,04
III	Коэффициент глагольности (СГ)	0,16	0,16	0,16
IV	Коэффициент действия 1 (КД1)	1,25	1,24	1,18
V	Коэффициент действия 2 (КД2)	1,53	1,51	1,45
VI	Коэффициент опредмеченности действия (КОД)	0,39	0,40	0,38
VII	Коэффициент логической связности 1 (ЛС1)	2,26	2,31	2,36
VIII	Коэффициент логической связности 2 (ЛС2)	0,19	0,19	0,19
IX	Коэффициент связности лексики (СЛ)	3,74	3,55	3,59

Табл. 3

Коэффициенты лексического разнообразия графа $\widetilde{\widetilde{G_1}}$

Характеристика	Номер сообщества								
	0	1	2	3	4	5	6	7	8
Размер текстов (Кб)	381	338	5045	1718	799	491	82	7248	4958
Коэффициент лексического разнообразия 1 (ЛР1)	0,160	0,190	0,060	0,120	0,140	0,190	0,380	0,060	0,070
Коэффициент лексического разнообразия 2 (ЛР2)	0,190	0,230	0,090	0,160	0,200	0,250	0,460	0,090	0,100

Характеристика	Номер сообщества								
	9	10	11	12	13	14	15	16	Все
Размер текстов (Кб)	247	2435	866	39	1899	1399	2076	1083	31104
Коэффициент лексического разнообразия 1 (ЛР1)	0,270	0,090	0,140	0,460	0,110	0,120	0,110	0,160	0,030
Коэффициент лексического разнообразия 2 (ЛР2)	0,340	0,120	0,190	0,550	0,140	0,160	0,140	0,220	0,040

- Коэффициент действия 1 (КД1) – отношение количества глаголов (деепричастия и причастия исключаются) к количеству прилагательных.
- Коэффициент действия 2 (КД2) – отношение количества глаголов и глагольных форм (деепричастий и причастий) к количеству прилагательных.
- Коэффициент опредмеченности действия (КОД) – соотношение количества глаголов (деепричастия и причастия исключаются) к количеству существительных.
- Коэффициент логической связности 1 (ЛС1) – отношение общего количества служебных слов (союзов и предлогов) к общему количеству предложений.
- Коэффициент логической связности 2 (ЛС2) – коэффициент использования служебных слов отношение общего количества служебных слов (союзов и предлогов) к общему количеству словоупотреблений.
- Коэффициент связности лексики (СЛ) – отношение числа существительных и глаголов (деепричастия и причастия исключаются) к количеству прилагательных и наречий.

Для начала рассмотрим данные показатели для массивов всех текстов каждого из трех графов после применения к ним Шага 4.

В табл. 2 приведены полученные значения перечисленных выше показателей для объединенных для каждого графа массивов текстов без разделения их по сообществам графов после применения к ним Шага 4. В первой строке таблицы приведены размеры объединенных массивов текстов.

Как видно из табл. 2, все характеристики для объединенных текстовых массивов трех графов разной направленности и тематики почти одинаковые, то есть три разных больших набора текстов сети Telegram имеют схожие психолингвистические характеристики. Поэтому можно высказать предположение, что данные общие показатели – это единые характеристики текстов в каналах мессенджера Telegram.

Рассмотрим коэффициенты лексического разнообразия ЛР1 и ЛР2. Для двух первых анализируемых графов результаты представлены в табл. 3 и 4. Значения этих коэффициентов уменьшаются с

существенным ростом объемов массивов текстов метасообществ, что наблюдается для всех анализируемых графов. Чем больше текстов в Telegram-каналах (сообществах), тем они более однообразны.

Для сообществ графов G_1 и G_2 проведена экспертная оценка тематик выделенных метасообществ.

5. Анализ сообществ графа $\widetilde{\widetilde{G_1}}$ в соответствии с экспертным анализом

На основе содержания каналов полученные сообщества были объединены в несколько крупных групп: финансы/недвижимость, политика/новости, мода/искусство, смешанные. Стоит сказать, что в определенных сообществах сложно выделить единую тематику, поэтому их стоит относить к смешанным, но есть и такие, где четко видна единая тематика всех Telegram-каналов, входящих в одно сообщество. Например, в одно из метасообществ (№1) попали следующие Telegram-каналы*: @smety* (Просто о сметах), *@stroika_glavnoe* (Стройка. Главное.), *@pesetcetera* (П.Э.С. – Проектирование. Экспертиза. Строительство.), *@erzrf* (ЕРЗ.РФ НОВОСТИ – Единый ресурс застройщиков), *@minstroyrf* (Минстрой России). Все каналы связаны единой тематикой – строительством, что также подтверждает правильность алгоритмов выделения неявных сообществ в графе.

По итогу сообщества были разделены следующим образом: финансы/недвижимость: № 0, 1, 11; политика/новости: № 2, 3, 5, 13, 14, 15; мода/искусство: № 4, 6, 12, 16; смешанные: № 7 (финансы/новости), 8 финансы/новости/политика), 9 (финансы/мода), 10 (в основном финансы).

Стоит отметить, что больше всего текстов у сообществ, где в основном представлены новостные или политические Telegram-каналы (№ 2, 7, 8, 13, 14, 15), и соответственно в них лексическое разнообразие меньше (табл. 3). Это объяснимо, так как в каналах такого типа в день больше сообщений и в основном они имеют единую форму. А меньше текстов и больше разнообразия в сообществах с каналами про моду и искусство (6, 9, 12, 16), что также объясняется более разнообразным содержанием в каналах данной сферы. Отдельно выделим 12 сообщество, которое показало наибольший коэффициент разнообразия: в данном сообществе представлены медийные Telegram-каналы. В остальных сообществах (в основном финансы и недвижимость) значения коэффициентов средние.

Рассмотрим следующую группу рассчитанных коэффициентов (табл. 5): коэффициент глагольности, коэффициенты действия и коэффициент опредмеченности действия. Все эти коэффициенты коррелируют между собой, так как при их расчете значение числителя зависит от количества глаголов в тексте и они характеризуют активность и практическую направленность текстов на действия. В данном случае явно выделилось одно сообщество (№ 1, все каналы про стройку) – в текстах данного сообщества наблюдается мало глаголов относительно других частей речи. В финансах и недвижимости выделяются средние значения коэффициентов. В каналах, посвященных моде и искусству, преобладают повышенные коэффициенты, сильно отличается только одно сообщество – № 6. Но если внимательнее изучить каналы данного сообщества, то можно увидеть, что они действительно отличаются от других сообществ данной тематики: здесь преобладают каналы об украшениях и искусстве.

В политических и новостных сообществах заметна интересная деталь: в целом у данных каналов наблюдаются пониженные значения коэф-

Табл. 4

Коэффициенты лексического разнообразия графа $\widetilde{\widetilde{G_2}}$

Характеристика	Номер сообщества									
	0	1	2	3	4	5	6	7	8	9
Размер текстов (Кб)	1130	323	235	122	1475	4140	228	195	282	1240
Коэффициент лексического разнообразия 1 (ЛР1)	0,09	0,28	0,16	0,31	0,12	0,08	0,27	0,23	0,22	0,12
Коэффициент лексического разнообразия 2 (ЛР2)	0,15	0,33	0,25	0,38	0,16	0,11	0,34	0,31	0,27	0,17

Характеристика	Номер сообщества										
	10	11	12	13	14	15	16	17	18	19	Все
Размер текстов (Кб)	9715	252	265	1722	270	209	3670	432	842	840	27587
Коэффициент лексического разнообразия 1 (ЛР1)	0,05	0,21	0,26	0,10	0,19	0,24	0,07	0,17	0,16	0,14	0,03
Коэффициент лексического разнообразия 2 (ЛР2)	0,07	0,29	0,33	0,15	0,26	0,31	0,10	0,23	0,22	0,19	0,04

Табл. 5

Коэффициенты глагольности и действий графа $\widetilde{\widetilde{G_1}}$

Характеристика	Номер сообщества								
	0	1	2	3	4	5	6	7	8
Коэффициент глагольности (СГ)	0,146	0,130	0,160	0,160	0,170	0,170	0,140	0,160	0,160
Коэффициент действия 1 (КД1)	1,310	0,860	1,180	1,090	1,210	1,630	0,940	1,360	1,340
Коэффициент действия 2 (КД2)	1,600	1,180	1,460	1,390	1,640	1,880	1,230	1,620	1,610
Коэффициент опредмеченности действия (КОД)	0,360	0,240	0,370	0,360	0,350	0,470	0,330	0,410	0,420

Характеристика	Номер сообщества								
	9	10	11	12	13	14	15	16	Все
Коэффициент глагольности (СГ)	0,150	0,160	0,159	0,165	0,150	0,163	0,151	0,162	0,158
Коэффициент действия 1 (КД1)	1,040	1,160	1,330	1,480	1,180	1,310	1,090	1,370	1,250
Коэффициент действия 2 (КД2)	1,290	1,420	1,640	1,670	1,450	1,590	1,360	1,640	1,530
Коэффициент опредмеченности действия (КОД)	0,350	0,370	0,370	0,470	0,350	0,420	0,360	0,450	0,390

фициентов, характеризующих активность и направленность на действия (можно объяснить тем фактом, что каналам данных направлений больше присуща описательная форма повествования), но из этого ряда выделяются два сообщества – № 5, 14. В пятом сообществе наблюдаются самые высокие коэффициенты из всех, а в четырнадцатом – повышенные относительно среднего по всем и по новостным/политическим, в частности. Подробнее изучив состав данных сообществ, можно выделить особенность: в их состав входят Telegram-каналы с альтернативными, оппозиционными взглядами. Стоит отметить, что подобные повышенные коэффициенты характерны для объектов, склонных к активным действиям, что в целом соотносится с направленностью указанных сообществ.

Последним блоком посчитанных характеристик являются коэффициенты логической связности и коэффициент связности лексики (табл. 6). В данных коэффициентах повышением выделяются сообщества по финансам и недвижимости. Из них явно выделяется нулевое сообщество. Это объяснимо тем, что в нем содержится много крупных и официальных Telegram-каналов (например, *@tinkoff_invest_official*). И в целом эта особенность присуща и другим тематикам. Те сообщества, где есть крупные или официальные каналы, имеют более высокие коэффициенты связности. Помимо финансов это наблюдается в политических и новостных каналах, куда также часто входят крупные или официальные региональные каналы (например, официальные каналы губернаторов областей). Как видим, четко выделяется особенность: высокие коэффициенты логической связности характеры официальным или более крупным Telegram-каналам. У остальных сообществ данные характеристики меньше (в нашем примере выделяются сообщества о моде и искусстве, где в основном представлены средние и маленькие каналы).

Таким образом, выделенные метасообщества для графа G_1 отличаются в большинстве случаев психолингвистическими характеристиками текстов и эти отличия коррелируют с экспертной оценкой тематической направленности конкретных метасообществ. Данные факты подтверждают релевантность модели разбиения графов сети взаимодействующих каналов мессенджера Telegram на неявные сообщества.

Табл. 6

Коэффициенты связности лексики графа $\widetilde{\widetilde{G_1}}$

Характеристика	Номер сообщества								
	0	1	2	3	4	5	6	7	8
Коэффициент логической связности 1 (ЛС1)	2,370	2,020	2,440	2,180	1,340	2,040	1,780	2,230	2,260
Коэффициент логической связности 2 (ЛС2)	0,220	0,190	0,180	0,180	0,190	0,190	0,190	0,190	0,190
Коэффициент связности лексики (СЛ)	4,220	3,850	3,780	3,570	3,890	4,160	3,070	3,840	3,690

Характеристика	Номер сообщества								
	9	10	11	12	13	14	15	16	Все
Коэффициент логической связности 1 (ЛС1)	2,110	2,490	2,470	1,790	2,390	2,390	2,320	2,190	2,260
Коэффициент логической связности 2 (ЛС2)	0,200	0,200	0,190	0,180	0,180	0,200	0,180	0,200	0,190
Коэффициент связности лексики (СЛ)	3,320	3,630	4,122	3,490	3,810	3,720	3,460	3,480	3,740

Табл. 7

Коэффициенты глагольности и действий графа $\widetilde{\widetilde{G_2}}$

Характеристика	Номер сообщества									
	0	1	2	3	4	5	6	7	8	9
Коэффициент глагольности (СГ)	0,17	0,15	0,18	0,13	0,15	0,16	0,15	0,16	0,15	0,15
Коэффициент действия 1 (КД1)	1,93	1,24	1,83	0,89	1,01	1,29	1,24	1,76	1,41	0,93
Коэффициент действия 2 (КД2)	2,15	1,48	2,03	1,08	1,26	1,54	1,49	2,00	1,71	1,18
Коэффициент опредмеченности действия (КОД)	0,56	0,41	0,54	0,28	0,34	0,42	0,39	0,44	0,40	0,33

Характеристика	Номер сообщества										
	10	11	12	13	14	15	16	17	18	19	Все
Коэффициент глагольности (СГ)	0,16	0,16	0,16	0,15	0,16	0,16	0,16	0,16	0,16	0,15	0,16
Коэффициент действия 1 (КД1)	1,23	1,60	1,37	1,14	1,67	1,53	1,17	1,62	1,10	2,01	1,24
Коэффициент действия 2 (КД2)	1,52	1,81	1,61	1,41	1,91	1,73	1,48	1,82	1,32	2,31	1,51
Коэффициент опредмеченности действия (КОД)	0,39	0,49	0,44	0,37	0,48	0,46	0,40	0,52	0,44	0,42	0,40

Табл. 8

Коэффициенты связности лексики графа $\widetilde{\widetilde{G_2}}$

Характеристика	Номер сообщества									
	0	1	2	3	4	5	6	7	8	9
Коэффициент логической связности 1 (ЛС1)	1,94	2,56	2,02	2,07	2,28	2,46	2,37	2,00	1,83	2,65
Коэффициент логической связности 2 (ЛС2)	0,19	0,20	0,21	0,20	0,18	0,19	0,20	0,19	0,18	0,19
Коэффициент связности лексики (СЛ)	4,07	3,41	3,86	3,54	3,39	3,53	3,61	4,47	3,72	3,12

Характеристика	Номер сообщества										
	10	11	12	13	14	15	16	17	18	19	Все
Коэффициент логической связности 1 (ЛС1)	2,37	2,26	2,17	2,35	1,95	1,83	2,54	2,01	2,06	1,39	2,31
Коэффициент логической связности 2 (ЛС2)	0,19	0,20	0,20	0,20	0,20	0,21	0,18	0,21	0,18	0,20	0,19
Коэффициент связности лексики (СЛ)	3,66	3,64	3,43	3,41	3,91	3,65	3,39	3,52	2,90	4,98	3,55

6. Анализ сообществ графа $\widetilde{\widetilde{G_2}}$ в соответствии с экспертным анализом

Рассмотрим характеристики текстов сообществ графа $\widetilde{\widetilde{G_2}}$, полученного при старте импорта данных от канала *@sportsru*. Как и указано в табл. 1, после выполнения Шага 4 было получено 20 сообществ.

По итогам экспертных оценок выделены следующие большие тематики, объединяющие сообщества в несколько основных групп: спорт (отдельно выделен футбол), политика/экономика, музыка и смешанные. Контингент читателей исходного канала действительно часто интересуется и смешивает политику со спортом, а группа «музыка» аналогична группе «мода/искусство» для $\widetilde{\widetilde{G_1}}$.

Таким образом, сообщества были разделены так: спорт (№ 2, 11); футбол (№ 0, 7, 8, 14, 15, 17, 19); политика/экономика (№ 4, 9, 13, 16, 18); смешанные – спорт/политика/экономика (№ 3, 5, 10). Также во многих сообществах видно и более узкое распределение, так, в сообществе № 0 почти все каналы связаны с конкретным футбольным клубом (ФК Спартак), а в № 14 – с ФК Локомотив. Среди спорта от всех остальных отделился киберспорт (№ 2) и каналы про ставки на спорт (№ 19).

Стоит отметить, что аналогично графу $\widetilde{\widetilde{G_1}}$, больше всего текстов у сообществ, где в основном представлены политические или экономические Telegram-каналы и смешанные сообщества, во все из которых входят в том числе политические каналы (№ 0, 4, 5, 9, 10, 13, 16, 18), и соответственно в них лексическое разнообразие меньше (табл. 4). Средние показатели разнообразия характерны для сообществ про спорт в целом и футбол, что объяс-

нимо более разнообразными каналами и взглядами в данной сфере. Тут также стоит отметить два сообщества про футбол (17 и 19), которые имеют более низкие показатели разнообразия (ближе к политическим каналам). Это можно объяснить тем, что в 17 сообществе представлены каналы по аналитике футбола, тексты которых имеют примерно одинаковую форму, а в 19 большинство каналов связаны со ставками, где также используется меньше словесных форм. А меньше текстов и самое большее разнообразие в сообществах с каналами про музыку.

Рассмотрим следующую группу рассчитанных коэффициентов, характеризующих активность и практическую направленность текстов на действия: коэффициент глагольности, коэффициенты действия и коэффициент опредмеченности действия (табл. 7). В данном случае явное повышение коэффициентов наблюдается в каналах про спорт/футбол. Этот факт в целом объясняется спецификой тематики: при описании спорта используется больше глаголов и повышенные коэффициенты характерны для объектов, склонных к активным действиям. А в политических и экономических сообществах аналогично графу $\widetilde{\widetilde{G_1}}$ наблюдаются пониженные значения коэффициентов (им больше присуща описательная форма повествования).

Последним блоком посчитанных характеристик являются коэффициенты логической связности и коэффициент связности лексики (табл. 8). Если рассматривать все три коэффициента сразу, то явно выраженной логики разбиения нет. Это отличие от графа $\widetilde{\widetilde{G_1}}$ можно объяснить тем, что среди каналов графа $\widetilde{\widetilde{G_2}}$ мало крупных и официальных. Большинство каналов про спорт представлены малыми и средними авторскими каналами. Но стоит отметить, что каждая из выделенных тематик (политика/экономика, спорт, футбол, музыка) имеет схожие коэффициенты внутри себя, что подтверждает разность текстов у каналов различной направленности. Например, в политических/экономических сообществах (№ 4, 9, 13, 16, исключение – 18) наблюдаются одинаковые коэффициенты, причем ЛС1 повышен (максимум из всех сообществ), а ЛС2 и СЛ понижены (минимум из всех сообществ). Также все футбольные сообщества (№ 0, 7, 8, 14, 15, 17, 19) имеют пониженный ЛС1 и повышенный СЛ (максимум из всех сообществ). Схожа с графом $\widetilde{\widetilde{G_1}}$ картина с коэффициентом ЛС2: сообщества с официальными каналами (№14, 15, 17) имеют более высокие значения характеристики.

7. Сравнение характеристик для $\widetilde{\widetilde{G_1}}$ и $\widetilde{\widetilde{G_2}}$

Сравним характеристики политических/новостных сообществ графов $\widetilde{\widetilde{G_1}}$ (табл. 9) с политическими/экономическими сообществами графа $\widetilde{\widetilde{G_2}}$ (табл. 10). Будем рассматривать средние значения коэффициентов по указанным сообществам. Как видно из таблиц, коэффициенты лексического разнообразия почти одинаковые. Глагольные коэффициенты тоже в целом схожи, в частности средние значения СГ и КОД почти одинаковые, но и КД1 и КД2 различаются не сильно, учитывая разброс данных коэффициентов в остальных сообществах (КД1 от 0,886 до 2,012; КД1 от 1,079 до 2,3). Аналогично глагольным коэффициентам схожи и коэффициенты ЛС1, ЛС2 и СЛ.

Таким образом, учитывая, что данные различные политические сообщества были сформированы на графах исходно разной направленности, можно предположить, что посчитанные средние психолингвистические характеристики являются отличительными для политических каналов в сети взаимодействия мессенджера Telegram.

Табл. 9

Характеристики графа $\widetilde{\widetilde{G_1}}$

Характеристика	Номер сообщества						
	2	3	5	13	14	15	среднее
Размер текстов (Кб)	5045	1718	491	1899	1399	2076	2105
Коэффициент лексического разнообразия 1 (ЛР1)	0,06	0,12	0,19	0,11	0,12	0,11	0,118
Коэффициент лексического разнообразия 1 (ЛР1)	0,09	0,16	0,25	0,14	0,16	0,14	0,157
Коэффициент глагольности (СГ)	0,16	0,16	0,17	0,15	0,16	0,15	0,159
Коэффициент действия 1 (КД1)	1,18	1,09	1,63	1,18	1,31	1,09	1,247
Коэффициент действия 2 (КД2)	1,46	1,39	1,88	1,45	1,59	1,36	1,522
Коэффициент опредмеченности действия (КОД)	0,37	0,36	0,47	0,35	0,42	0,36	0,388
Коэффициент логической связности 1 (ЛС1)	2,44	2,18	2,04	2,39	2,39	2,32	2,293
Коэффициент логической связности 2 (ЛС2)	0,18	0,18	0,19	0,18	0,20	0,18	0,185
Коэффициент связности лексики (СЛ)	3,78	3,57	4,16	3,81	3,72	3,46	3,750

Табл. 10

Характеристики графа $\widetilde{\widetilde{G}}_2$

Характеристика	Номер сообщества					
	4	**9**	**13**	**16**	**18**	среднее
Размер текстов (Кб)	1475	1240	1722	3670	842	1790
Коэффициент лексического разнообразия 1 (ЛР1)	0,115	0,122	0,104	0,070	0,158	0,114
Коэффициент лексического разнообразия 2 (ЛР2)	0,156	0,169	0,146	0,098	0,215	0,157
Коэффициент глагольности (СГ)	0,146	0,147	0,153	0,162	0,156	0,153
Коэффициент действия 1 (КД1)	1,014	0,927	1,143	1,167	1,097	1,070
Коэффициент действия 2 (КД2)	1,261	1,175	1,409	1,479	1,324	1,330
Коэффициент опредмеченности действия (КОД)	0,338	0,334	0,372	0,395	0,443	0,376
Коэффициент логической связности 1 (ЛС1)	2,279	2,652	2,350	2,542	2,061	2,377
Коэффициент логической связности 2 (ЛС2)	0,180	0,185	0,199	0,184	0,180	0,186
Коэффициент связности лексики (СЛ)	3,391	3,118	3,414	3,388	2,902	3,243

Отметим следующий результат. В выделенных политических и новостных сообществах выявлена закономерность: у большинства данных каналов наблюдаются пониженные значения коэффициентов глагольности по сравнению с каналами иных тематик. При этом для каналов с альтернативными и оппозиционными взглядами наоборот данные характеристики выше среднего по всем каналам и по новостным/политическим в частности. Подобные повышенные коэффициенты действительно характерны для объектов, склонных к активным действиям.

Приведенные результаты по сравнению значений психолингвистических характеристик текстов метасообществ в сопоставлении с экспертной оценкой тематической направленностью конкретных метасообществ показывают, что в большинстве случаев выделенные метасообщества отличаются между собой в соответствии с тематической направленностью. Это свидетельствует о корректности выделения неявных сообществ предлагаемым «методом Галактик» и возможностями анализа информационного взаимодействия между объектами сети на основе данного метода.

Обсуждение. В статье [1] основоположники данной тематики указали на несколько актуальных для всех алгоритмов выделения неявных сообществ аспектов. Одним из них является проблема так называемого «resolution limit» [16], предела разрешения, который не позволяет находить небольшие сообщества в больших сетях. Представленный в данной работе «метод Галактик» как раз решает данную проблему, позволяя за счет перехода от анализа всего исходного графа – всей «Вселенной» к отдельным подграфам – «Галактикам» выделять и маленькие сообщества.

Другой актуальный вопрос, поднятый в [1] состоит в том, что обычно для оценки качества разбиения используют либо сравнение с заранее заданным разбиением на искусственных сетях, либо сравнение за счет дополнительных данных о вершинах в реальных сетях. Первый подход дает сбой для разреженных графов, характерных для социальных сетей. Приведенный в данной статье психолингвистический анализ текстов и их характеристик относится ко второму подходу для оценки качества получаемого разбиения.

Заключение

В рамках данной работы представлен «метод Галактик» – новый подход выделения неявных пересекающихся сообществ на графах взаимодействующих объектов.

Впервые предлагается совместное использование методов выделения неявных сообществ и компьютерной лингвистики анализа психолингвистических факторов текстов для оценки качества выделения сообществ.

Приведенные результаты по сравнению значений психолингвистических характеристик текстов сообществ в сопоставлении с экспертной оценкой тематического содержания массивов текстов конкретных сообществ показывают, что выделенные сообщества отличаются в большинстве случаев психолингвистическими характеристиками текстов, что коррелирует с тематической направленностью конкретных сообществ. Таким образом, показана корректность выделения неявных сообществ предлагаемым методом и возможности анализа информационного взаимодействия на основе этого выделения.

Преобразование графа к метасообществам в процессе «метода Галактик» уменьшает количество структурных единиц графа, необходимых для анализа и позволяет сокращать количество элементов в задаче визуализации графов больших размеров.

Работы в сфере анализа социальных сетей и мессенджеров в настоящее время имеют высокую степень актуальности. Авторы видят боль-

шие перспективы в сочетании алгоритмических методов выделения сообществ и психолингвистическом анализе с последующим построением цифровых профилей. В том числе, в дальнейшем представляет интерес сравнение психолингвистических характеристик текстов для различных социальных сетей и выявление присущих им закономерностей.

Литература

1. *Fortunato S., Newman M.E.J.* 20 years of network community detection. Nat. Phys. 2022. 18. P. 848–850.
2. *Лещёв Д.А., Сучков Д.В., Хайкова С.П., Чеповский А.А.* Алгоритмы выделения групп общения // Вопросы кибербезопасности. 2019. Т. 32. № 4. С. 61-71
3. *Соколова Т.В., Чеповский А.А.* Анализ профилей сообществ социальных сетей // Системы высокой доступности. 2018. Т. 14. № 3. С. 82-86.
4. *Коломейченко М.И., Поляков И.В., Чеповский А.А., Чеповский А.М.* Выделение сообществ в графе взаимодействующих объектов // Фундаментальная и прикладная математика. 2016. Том 21. №3. С. 131-139.
5. *Roth M., Ben-David A., Deutscher D.*. Suggesting Friends Using the Implicit Social Graph — KDD'10. July. 25–28. 2010. Washington. DC. USA. 2010.
6. *Girvan M., Newman M.* Community structure in social and biological networks. Proceedings of the National Academy of Sciences. 2002. Vol. 99. No 12. P. 7821-7826.
7. *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. No 10. P. 10008.
8. *Rosvall M.* The map equation / M. Rosvall, D. Axelsson, C. T. Bergstrom // The European Physical Journal Special Topics. 2009.
9. *Chepovskiy A.A., Leshchev D.A., Khaykova S.P.* Core Method for Community Detection, in: Complex Networks & Their Applications IX. Volume 1: Proceedings of the Ninth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2020. Springer, 2021. P. 38-50.
10. *Попов В.А., Чеповский А.А.* Модели импорта данных из Твиттера // Вестник НГУ. Серия: Информационные технологии. 2021. Т.19. №2. С. 76–91.
11. *Попов В.А., Чеповский А.А.* Модели импорта данных из мессенджера Telegram // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2022. Т.20. №2. С. 60-71.
12. *Kelley S.* The existence and discovery of overlapping communities in large-scale networks. Ph.D. thesis, Rensselaer Polytechnic Institute. Troy. NY. 2009.
13. *Que X., Checconi F., Petrini F., Gunnels J.* Scalable Community Detection with the Louvain Algorithm // 29th IEEE International Parallel & Distributed Processing Symposium. 2015. May. P. 25-29.
14. *Коломейченко М.И., Поляков И.В., Чеповский А.А., Чеповский А.М.* Методы визуального анализа графов. М.: Национальный открытый университет «ИНТУИТ». 2016. 167 с.
15. *Аванесян Н.Л., Соловьев Ф.Н., Чеповский А.А.* Характеристики текстов сообществ социальных сетей // Вестник НГУ. Серия: Информационные технологии. 2021 Т.19, №1. С. 5–14.
16. *Fortunato S. & Barthélemy M.* Resolution limit in community detection. Proc. Natl. Acad. Sci. USA. 2007. 104. P. 36-41.

Попов Владимир Александрович. Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия. Аспирант. Количество печатных работ: 2. Область научных интересов: информационные технологии, программирование. E-mail: vapopov_1@edu.hse.ru

Чеповский Александр Андреевич. Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия. Кандидат физико-математических наук, доцент. Количество печатных работ: 35. Область научных интересов: компьютерные науки, теория графов, теория колец. E-mail: aachepovsky@hse.ru (ответственный за переписку).

Use of the "Galaxies method" to reveal overlapping communities on the Telegram channels interaction graph

A.V. Popov, A.A. Chepovskiy
Higher School of Economics, Moscow, Russia.

Abstract. In this paper, the authors present the "Galaxies method" to reveal implicit communities on the graph of interacting objects obtained by importing a network of channels from the Telegram messenger. This method is based on successive identification of overlapping communities on the initial weighted graph, further construction of a new graph, in which the vertices are the communities revealed at the first step, called by the authors "metavertices". The weighted edges of the new graph between the "metavertices" are built based on the weights between each pair of vertices in the original graph. Further, non-overlapping communities are identified on the new graph. The result is a partition of the original graph into overlapping "meta-communities". To assess the quality of partitioning using the presented method, the authors carried out a psycholinguistic analysis of the obtained metacommunities, identified patterns depending on the thematic orientation of channels within the metacommunity. As a result of processing and analysis of the obtained metacommunities, the quality of the partition was confirmed. The combination of the algorithmic method of revealing communities and psycholinguistic analysis has a practical application for the analysis of information impact in social networks.
Keywords: *Telegram, analysis of social networks, data import from social networks, model of information impact, graph of interacting objects, identification of communities, psycholinguistic analysis of texts*.
DOI: 10.14357/20790279220405

References

1. *Fortunato S., Newman M.E.J.* 20 years of network community detection. Nat. Phys. 2022. 18. P. 848–850.
2. *Leshchev D.A., Suchkov D.V., Khaykova S.P., Chepovskiy A.A.* Algoritms to revel communication groups // Voprosi kiberbezopasnosti. 2019. V. 32. No. 4. P. 61-71
3. *Sokolova T.V., Chepovskiy A.A.* Analysis on communities profiles in social networks // Highly Available System. 2018. V. 14. No. 3. P. 82-86.
4. *Kolomeychenko M.I., Polyakov I.V., Chepovskiy A.A., Chepovskiy A.M.* Detection of communities in graph of interactive objects // Fundamentalnaya i Prikladnaya Matematika. 2016. V. 21. No. P. 131-139.
5. *Roth M., Ben-David A., Deutscher D.* Suggesting Friends Using the Implicit Social Graph — KDD'10. July. 25–28. 2010. Washington. DC. USA. 2010.
6. *Girvan M., Newman M.* Community structure in social and biological networks // Proceedings of the National Academy of Sciences. 2002. Vol. 99. No. 12. P. 7821-7826.
7. *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. No. 10. P 10008.
8. *Rosvall M.* The map equation / M. Rosvall, D. Axelsson, C. T. Bergstrom // The European Physical Journal Special Topics. 2009.
9. *Chepovskiy A.A., Leshchev D.A., Khaykova S.P.* Core Method for Community Detection / Complex Networks & Their Applications IX. Volume 1: Proceedings of the Ninth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2020. Springer. 2021. P. 38-50.
10. *Popov V.A., Chepovskiy A.A.* Twitter Data Import Models // Vestnik NSU. Series: Information Technologies. 2021. V.19. No. 2. P. 76–91.
11. *Popov V.A., Chepovskiy A.A.* Telegram Messenger Data Import Models // Vestnik NSU. Series: Information Technologies. 2022. V.20. No. 2. P. 60-71.
12. *Kelley S.* The existence and discovery of overlapping communities in large-scale networks. Ph.D. thesis, Rensselaer Polytechnic Institute. Troy. NY. 2009.
13. *Que X., Checconi F., Petrini F., Gunnels J.* Scalable Community Detection with the Louvain Algorithm // 29th IEEE International Parallel & Distributed Processing Symposium. 2015. May. P. 25-29.
14. *Kolomeychenko M.I., Polyakov I.V., Chepovskiy A.A., Chepovskiy A.M.* Metodi vizualnogo analiza graphov. M.: National Open University «INTUIT». 2016. 167 c
15. *Avanesyan N.L., Solovev F.N., Chepovskiy A.A.* Characteristics of Texts of Social Networks Communities // Vestnik NSU. Series: Information Technologies. 2021. V.19. No. 1. P. 5–14.
16. *Fortunato S. & Barthélemy M.* Resolution limit in community detection. Proc. Natl. Acad. Sci. USA. 2007. 104. P. 36-41.

Popov A.V. Graduate student, National Research University Higher School of Economics, Moscow, Russia. vapopov_1@edu.hse.ru

Alexander A.C. Ph.D. (mathematics), Associate Professor, National Research University Higher School of Economics, Moscow, Russia. aachepovsky@hse.ru.