

DOI: 10.15514/ISPRAS-2022-34(5)-8



Определение влиятельных пользователей социальной сети по двудольному графу комментариев

¹ Р.К. Пастухов, ORCID: 0000-0001-6919-0701 <pastukhov@ispras.ru>
¹ М.Д. Дробышевский, ORCID: 0000-0002-1639-9154 <drobyshevsky@ispras.ru>
^{1,2} Д.Ю. Турдаков, ORCID: 0000-0001-8745-0984 <turdakov@ispras.ru>

¹ Институт системного программирования им. В.П. Иванникова РАН,
 109004, Россия, г. Москва, ул. А. Солженицына, д. 25

² Московский государственный университет имени М.В. Ломоносова,
 119991, Россия, Москва, Ленинские горы, д. 1

Аннотация. С развитием онлайн-социальных сетей все большую актуальность приобретают задачи выявления пользователей, которые оказывают большое влияние на других участников социальных сетей. Важным источником информации служат данные о комментировании пользователями контента, создаваемого другими пользователями. В работе предлагается метод определения влиятельности, основанный на двудольном графе пользователь-комментарий-контент, а также использующий информацию о текстовых сообщениях и реакции на них других пользователей. Кроме того, предложен метод выявления сообществ пользователей в таком графе на основе общих интересов. Результаты тестирования на коллекциях данных сетей ВКонтакте и YouTube показывают корреляцию активности и влиятельности пользователя, однако самые активные комментаторы не обязательно являются самыми влиятельными. Анализ сообществ показывает положительную корреляцию между размером сообщества, числом наиболее влиятельных пользователей в нем и средней влиятельностью пользователей сообщества.

Ключевые слова: социальные сети; влиятельные пользователи; двудольный граф; поиск сообществ.

Для цитирования: Пастухов Р.К., Дробышевский М.Д., Турдаков Д.Ю. Определение влиятельных пользователей социальной сети по двудольному графу комментариев. Труды ИСП РАН, том 34, вып. 5, 2022 г., стр. 127-142. DOI: 10.15514/ISPRAS-2022-34(5)-8

Благодарности: Исследование выполнено в рамках научной программы Национального центра физики и математики (проект №9 “Искусственный интеллект и большие данные в технических, промышленных, природных и социальных системах”).

Detecting Influential Users in Social Networks Based on Bipartite Comments Graph

¹ R.K. Pastukhov, ORCID: 0000-0001-6919-0701 <pastukhov@ispras.ru>
¹ M.D. Drobyshevskiy, ORCID: 0000-0002-1639-9154 <drobyshevsky@ispras.ru>
^{1,2} D.Yu. Turdakov, ORCID: 0000-0001-8745-0984 <turdakov@ispras.ru>

¹ Ivannikov Institute for System Programming of the RAS,
 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia

² Lomonosov Moscow State University,
 GSP-1, Leninskie Gory, Moscow, 119991, Russia

Abstract. With the development of online social networks, the task of identifying users who have a great influence on other participants in social networks is becoming increasingly important. An important source of information is user comments on content created by other users. The paper proposes a method for determining influence based on a bipartite user-comment-content graph. It incorporates information about text messages and the reactions of other users to them. In addition, we propose a method for identifying user communities in such a graph based on common interests. Experiments on data collections from VKontakte and YouTube networks show the correlation between user activity and influence, however, the most active commenters are not necessarily the most influential. Community analysis shows a positive correlation between the size of a community, the number of most influential users in it, and the average influence of community users.

Keywords: social network; influential user; bipartite graph; community detection

For citation: Pastukhov R.K., Drobyshevskiy M.D., Turdakov D.Yu. Detecting Influential Users in Social Networks Based on Bipartite Comments Graph. Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 5, 2022. pp. 127-142 (in Russian). DOI: 10.15514/ISPRAS-2022-34(5)-8

Acknowledgements. This work was done under a scientific program of National center of physics and mathematics (project no. 9 “Artificial intelligence and big data in technical, industrial, natural, and social systems”).

1. Введение

Социальные сети становятся все более обширным источником информации о социуме, поведении людей, их предпочтениях, общественных трендах. Пользователи оставляют информацию о себе, устанавливают связи с другими пользователями, пишут сообщения, комментируют контент, созданный другими пользователями, ставят оценки и т.д. Для извлечения полезной информации из всех этих массивов данных последние годы разрабатываются множество подходов и алгоритмов их анализа.

В данной работе ставятся следующие задачи:

- 1) разработать метод выявления влиятельных пользователей сетевых сообществ на основании анализа активности пользователей соцсетей, включая анализ их сообщений;
- 2) разработать метод выделения сетевых сообществ пользователей на основании анализа их активности;
- 3) провести экспериментальные исследования разработанных методов на коллекциях данных, собранных из ВКонтакте и YouTube.

Разрабатываемые методы выделения сетевых сообществ на основании анализа активности пользователей соцсетей должны на основе анализа разных типов взаимодействия пользователей (комментарии, лайки, репосты и т.п.) выделять неявные сообщества (группы) наиболее активных пользователей соцсетей, оказывающих влияние на других пользователей соцсетей. Для предварительных экспериментов взяты две коллекции данных, собранных на основе нескольких групп социальной сети ВКонтакте и нескольких каналов YouTube.

В работе проведен обзор методов поиска сообществ и определения влиятельности пользователей социальных сетей. Разработан метод подсчета влиятельности пользователей в

специфике имеющихся данных, а также предложен метод поиска сообществ. Выполнено тестирование разработанных методов на двух коллекциях данных.

2. Обзор литературы

В задаче необходимо с одной стороны определить сообщества пользователей социальной сети, с другой стороны, выделить наиболее активных и влиятельных ее членов. Приведем далее обзор методов определения структуры сообществ в социальных сетях, а затем рассмотрим подходы и алгоритмы поиска влиятельных вершин.

2.1 Методы определения сообществ пользователей в социальных сетях

Пользователи онлайн социальных сетей не всегда явно указывают сообщества, к которым они принадлежат. Кроме того, некоторые сообщества могут возникать неявно, например, коллеги, одноклассники, жильцы дома и т.п. Данные социального графа являются богатым источником информации для поиска явных и неявных сообществ, их анализа и дальнейшего использования. С ростом популярности онлайн социальных сетей последние годы активно появляются и развиваются методы поиска скрытой структуры сообществ. Большинство методов основываются только на социальных связях, которые отражены в виде связи дружбы, подписки и т.д. и в социальном графе выражаются ребрами, соединяющими соответствующих пользователей.

Можно выделить три основных класса алгоритмов, показывающих наиболее высокое качество определения сообществ. Это подходы на основе локальной оптимизации, вероятностные методы и методы распространения меток. Рассмотрим подробнее представителей этих классов, а также некоторые другие подходы.

2.1.1 Методы на основе локальной оптимизации

Подход заключается в оптимизации некоторой локальной функции качества для каждого сообщества независимо. Например, задается ядро будущего сообщества и постепенно расширяется путем добавления новых смежных вершин. В методе GCE [1] задается следующая функция качества:

$$F_S = \frac{k_{in}^S}{(k_{in}^S + k_{out}^S)^\alpha}.$$

При этом S – индуцированный подграф, являющееся растущим ядром сообщества. k_{in}^S – удвоенное число ребер, которые имеют начало и конец в S , k_{out}^S – число ребер с одной вершиной в S . α – параметр метода.

Другой популярный метод – OSLOM [2]. Метод способен обнаруживать сообщества в сетях с учетом направлений ребер, весов ребер, перекрывающихся сообществ, иерархии и динамики сообщества. Он основан на локальной оптимизации функции пригодности, выражающей статистическую значимость кластеров по отношению к случайным флуктуациям, которая оценивается с помощью инструментов экстремальной и порядковой статистики. OSLOM можно использовать отдельно или в качестве процедуры уточнения сообществ, обнаруженных другими методами.

2.1.2 Вероятностные модели

В подходах этого типа задается параметрическая модель, отражающая распределение на множестве графов с фиксированным числом вершин. Обычно в модель включается матрица связи вершин и сообществ. Работа алгоритма заключается в максимизации правдоподобия исходного графа в рамках модели. В итоге получается оптимальная матрица инцидентности вершин и сообществ. Примерами таких алгоритмов являются MOSES [3], BCD [4], BIGCLAM [5], CESNA [6].

Отдельно можно выделить методы максимизации модулярности. Изначально модулярность была определена Ньюманом и Гирваном [7] для измерения качества разбиения графа на набор кластеров как доля ребер внутри кластеров за вычетом ожидаемой такой доли в случайном графе с теми же вершинами и их степенями:

$$Q = \frac{1}{2E} \sum_{c \in C} \sum_{i,j \in c} \left(A_{ij} - \frac{k_i k_j}{2E} \right).$$

Здесь A_{ij} обозначает матрицу смежности графа, C – набор сообществ (покрытий), k_i – степень вершины i . Численное значение характеризует степень конкретной кластеризации графа. Она высока, когда кластеры плотные и слабо связаны друг с другом, и низка, когда кластеры формируются случайным образом.

В литературе было предложено множество эвристических методов для нахождения разбиений вершин графа с высокой модулярностью за разумное время. К таким подходам относятся жадные алгоритмы, спектральные методы, экстремальная оптимизация, имитация отжига, метод выборки и математическое программирование – подробнее можно найти обзор в работе [8]. В той же работе описан настраиваемый алгоритм обнаружения сообществ, который итеративно улучшает показатель качества сообщества Q путем разделения и слияния заданной структуры сетевого сообщества. Другой метод [9] представляет собой усовершенствованную версию популярного метода Louvain [10] для работы с направленными графами на основе расширения понятия модулярности для направленных графов.

2.1.3 Методы распространения меток

Идея распространения меток по ребрам графа популярна для графов с очень большим числом вершин и ребер благодаря низкой вычислительной сложности, хорошей масштабируемости и простоте реализации алгоритмов. Основная схема методов данного класса – итеративный процесс рассылки и получения вершинами графа меток сообществ вдоль ребер. На каждом шаге вершина накапливает входящие метки и отправляет свою метку соседям. Обычно отправка меток происходит синхронно и независимо, но существуют также и асинхронные схемы. Типичными представителями класса являются методы SLPA [11], BMLPA [12] и COPRA [13], которые отличаются только способом инициализации меток и стратегиями их обмена.

2.1.4 Методы, основанные на эго-сообществах

Существует ряд методов, основанных на понятии эго-сети пользователя – объединении вершин и ребер первой окрестности некоторой вершины. При этом рассматривается набор сообществ, которым принадлежат вершины эго-сети. Алгоритмы предлагают различные способы агрегации эго-сообществ в более большие. В методе Node Perception [14] строится мета-сеть эго-сообществ, в которой затем производится кластеризация. Алгоритмы DEMON [15] и EgoClustering [16] являются его модификациями. Метод EgoLP [17] использует эго-сообщества для агрегации и фильтрации меток сообществ, распространяющихся между узлами. Благодаря хорошей масштабируемости распределенная реализация EgoLP способна обрабатывать графы с миллиардами вершин.

2.2 Методы поиска влиятельных пользователей в социальных сетях

Постоянно растущее количество информации, обращающееся в социальных сетях, заставляет участников этих сетей конкурировать за внимание и влияние, полагаясь на других людей для распространения своих сообщений. Популярность и статус участников социальных сетей измеряются уровнем внимания, которое они получают в виде подписчиков, которые создают ссылки на свои учетные записи для автоматического получения контента, который они создают. Другим аспектом является влияние, которым обладают эти люди, которое

определяется фактическим распространением их контента через сеть. Это влияние определяется многими факторами, такими как новизна и резонанс их сообщений с сообщениями их подписчиков, а также качество и частота контента, который они генерируют.

Для выявления наиболее влиятельных пользователей сети в литературе предложены разные подходы в зависимости от размеров сети, имеющейся информации о структуре сети, дополнительном знании и содержании сообщений пользователей и других факторов. Далее рассмотрим следующие категории подходов: методы поиска хабов, методы использующие известную структуру графов, методы, учитывающие временную динамику, и наконец методы, использующие текстовые сообщения пользователей.

2.2.1 Методы поиска хабов

Известно, что по сравнению с обычными пользователями влиятельные пользователи [18] играют более важную роль в социальных графах с точки зрения связности, влияния и распространения информации. Они задают тенденцию в оценке или отношении к тому или иному факту или событию. Вследствие этого анализ структуры связности и обмена информацией только во “влиятельной” части сети, который содержит всех влиятельных пользователей и их попарные отношения, дает ряд перспективных возможностей.

Влиятельность вершины можно определить, как меру ее центральности в графе. Чаще всего используется степень, то есть количество связей с другими вершинами. Предполагается, что люди с большим количеством подписчиков или друзей, так называемые хабы, в социальной сети имеют широкий охват аудитории, а значит, могут формировать мнение других пользователей.

Большинство современных социальных сетей – Фейсбук, Твиттер, Инстаграм, Вконтакте – не позволяют напрямую скачать полную структуру социального графа для проведения анализа. Поэтому разработан ряд методов поиска хабов в больших сетях на основе знания окрестности текущего узла. В алгоритме взвешенного случайного блуждания [19] необходимо знать степени соседей посещаемого узла на каждом шаге алгоритма, чтобы обнаружить узлы с наивысшей степенью. В нескольких работах [20, 21] исследуется алгоритм максимальной наблюдаемой степени (MOD – Maximal observed degree). Стратегия обхода выбирает узел с наивысшей наблюдаемой степенью на каждом шаге. Метод DE-Crawler [22] представляет комбинацию двух основных стратегий: MOD и случайные блуждания. Первая отвечает за уплотнение, когда алгоритм стремится найти как можно больше узлов в текущей плотной области (сообществе). Фаза блужданий применяется для расширения, когда необходимо переместиться в другую плотную область. Авторы показывают, что предложенный метод лучше всего подходит для решения задачи обнаружения как можно большего количества узлов.

Показано, что метод случайного блуждания также эффективен для быстрого нахождения узлов большой степени в очень больших графах [23]. Экспериментально показано что метод с высокой вероятностью находит топ списка узлов с хорошей точностью. В другой работе предлагается рандомизированный двухэтапный алгоритм [23] определения топ-к узлов с высокой точностью, используя малый (например, 1000) бюджет запросов. Например, для социальной сети Twitter, имеющей более миллиарда пользователей, они определяют топ-100 пользователей с наибольшим количеством подписчиков с точностью более 90%. Позднее предложена более усовершенствованная версия – трехшаговый алгоритм [24].

2.2.2 Поиск влиятельных вершин в графах с полной информацией

При наличии полной информации о социальном графе значительно расширяется возможный инструментарий для анализа поведения пользователей. Исследования распространения информации в Twitter показали [25], что большинство пользователей действуют как пассивные потребители информации и не передают контент в сеть. Следовательно, чтобы

люди стали влиятельными, они должны не только привлечь внимание и, но и преодолеть пассивность пользователей. В работе [25] предлагается алгоритм, определяющий влияние и пассивность пользователей на основе их активности по передаче информации. Оценка, проведенная с набором данных о 2,5 миллионах пользователей, показывает, что вычисленная мера влиятельности является хорошим предсказателем кликов по URL-адресам, превосходит некоторые другие меры, которые явно не учитывают пассивность пользователей. Кроме того, авторы показывают, что высокая популярность не обязательно означает высокое влияние, и наоборот.

Предложенная модель делает следующие предположения:

- 1) оценка влияния пользователя зависит от количества людей, на которых он влияет, а также от их пассивности;
- 2) оценка влияния пользователя зависит от того, насколько вовлечены люди, на которых он влияет. Вовлеченность измеряется количеством внимания, которое пользователь уделяет данному пользователю по сравнению со всеми остальными;
- 3) оценка пассивности пользователя зависит от влиятельности тех, с кем он контактирует, но не находится под влиянием;
- 4) оценка пассивности пользователя зависит от того, насколько он отвергает влияние другого пользователя по сравнению со всеми остальными.

Алгоритм влияния-пассивности итеративно одновременно вычисляет показатели пассивности и влияния следующим образом. Задан граф G с вершинами N , ребрами E и их весами W . Веса w_{ij} на ребре $e = (i, j)$ представляет собой отношение влияния i на j к общему влиянию, которое i пытается оказать на j . Алгоритм выводит функцию $I: N \rightarrow [0, 1]$, которая представляет относительное влияние узла на сеть, и функцию $P: N \rightarrow [0, 1]$, которая представляет относительную пассивность узла.

Для каждого ребра (i, j) определяется скорость принятия как

$$u_{ij} = \frac{w_{ij}}{\sum_{k:(k,j) \in E} w_{kj}}.$$

Это значение представляет собой количество влияния, которое пользователь j принял от пользователя i , нормализованное по общему влиянию, принятому j от всех пользователей в сети. Уровень принятия можно рассматривать как вовлеченность или лояльность пользователя j пользователю i . С другой стороны, для каждого $e = (j, i) \in E$ определяется скорость отказов, как

$$v_{ij} = \frac{1 - w_{ji}}{\sum_{k:(j,k) \in E} (1 - w_{jk})}.$$

Поскольку значение $(1 - w_{ji})$ представляет собой количество влияния, которое пользователь i отвергает от j , то значение v_{ij} представляет влияние, которое пользователь i отвергает от пользователя j , нормализованное по общему влиянию, отклоненному от j всеми пользователями в сети. Алгоритм основан на следующих операциях:

$$I_i \leftarrow \sum_{j:(i,j) \in E} u_{ij} P_j$$

и

$$P_i \leftarrow \sum_{j:(j,i) \in E} v_{ji} I_j.$$

В итоге исследование показывает, что корреляция между популярностью и влиянием слабее, чем можно было бы ожидать. Это отражение того факта, что для распространения информации в сети люди должны пересылать ее другим членам, поэтому им приходится активно участвовать, а не просто пассивно читать ее.

Авторы другой работы [26] предложили модель двудольного графа с направленным взвешенным пользовательским контентом. Модель измеряет влияние, комбинируя как структурные свойства сети, так и контент, опубликованный пользователями. Итеративный алгоритм предназначен для вычисления двух показателей: влияния пользователей и охвата досок (board) в сети Pinterest. Эксперименты показывают, что модель может эффективно обнаруживать наиболее влиятельных пользователей и популярный контент.

Предложенный алгоритм основывается на идеях, подобных предыдущему алгоритму, и тоже является итерационным.

С помощью алгоритма авторы успешно определили десять самых влиятельных пользователей в наборе данных. Кроме того, хотя, как правило, влиятельные пользователи имели очень большое количество подписчиков (по сравнению со средним числом подписчиков для пользователя), не было никакой конкретной корреляции между ранжированием влиятельности и количества подписчиков. Этот результат согласуется с выводом из предыдущей работы и многими другими.

2.2.3 Поиск влиятельных сообществ пользователей

Другой подход используется в работе [27]. Рассматривается задача поиска сообществ пользователей как обнаружение скрытой структуры сообщества в больших социальных сетях на основе их общих интересов. Исходная мотивация состоит в наблюдении, что пользователи обычно уделяют больше внимания тем пользователям, у которых схожие интересы, что позволяет разделить пользователей на разные сообщества в соответствии с их общими интересами. Авторы предлагают два алгоритма (эвристический и HWeBA) для эффективного обнаружения влиятельных сообществ, которые они называют ядрами, использующих общие интересы в крупных социальных сетях.

Предварительный статистический анализ показал, что сумма весов ребер влиятельных пользователей намного больше, чем у других пользователей в социальной сети системы Epinions. Другими словами, каждый член ядра сообщества имеет более высокую сумму весов ребер (например, степень общих интересов) в/из ядра, чем вершина вне ядра.

В статье [28] рассматривается трактовка сообществ с разных точек зрения в зависимости от свойства изучаемой сети. Кроме классических структурных сообществ на основе связи дружбы и подписок выделяются сообщества, основанные на деятельности, на основе темы контента и на основе взаимодействия. Авторы анализируют набор данных Twitter, используя три различных веса структурной сети, предназначенных для выделения этих трех типов сообществ.

Поиск сообществ, основанных на деятельности, мотивируется вопросом “У каких пользователей в сети есть схожие профили активности?”. Сообщество рассматривается как пользователи, которые используют (или не используют) сервис в одно и то же время. Авторы говорят о группах пользователей, которые обладают так называемой “способностью прогнозирования на основе активности” друг друга, т.е., учитывая пользователя u и подписчика f , которые являются членами сообщества, обнаруженного с этим взвешиванием, существует снижение неопределенности в отношении профиля активности подписчика f (твиттинг или молчание), учитывая историю твитов пользователя u , игнорируя информацию, представленную в истории твитов подписчика f . Алгоритм рассматривает каждого пользователя Твиттера как единицу обработки информации, но полностью игнорирует содержание их твитов. Затем взвешиваются направленные ребра (сообщаемые отношения между подписчиками и подписчиками) между пользователями с помощью так называемой энтропии передачи, рассчитанной между историей твитов подписчика f и пользователя u .

Кратко типы сообществ можно охарактеризовать следующими вопросами:

- на основе структуры: кто ваши заявленные друзья; за кем вы следуете?
- на основе активности: кто имеет схожие профили активности?

- тематические: о чем вы говорите?
- на основе взаимодействия: с кем вы общаетесь?

Авторы показывают, что использование разных определений сообщества раскрывает разные и интересные отношения между пользователями. Кроме того, утверждается, что различные взгляды на сеть не выявляются при использовании только структурной информации о сети или какой-либо одной схемы взвешивания.

2.2.4 Использование текстовой информации

Анализ текстовой информации, создаваемой пользователями социальных сетей, чаще всего используется в задачах классификации. Например, в работе [22] рассматривалась задача предсказания пола пользователей Youtube по текстам их комментариев. Социальная сеть моделировалась двудольным графом, одна часть вершин обозначала пользователей, другая видеоролики, связи отражали факт просмотра пользователем видеоролика. Задача предсказания пола пользователей решалась алгоритмом распространения меток (известных значений пола) вместе с использованием признаков, полученных из текстов комментариев пользователей.

Из текстов извлекались следующие три группы признаков:

- 1) посимвольные: средняя длина комментария, соотношение заглавных букв к общему количеству букв, отношение пунктуации к общему количеству символов;
- 2) на основе токенов: средняя длина комментария в словах, отношение уникальных слов к общему количеству токенов, строчные униграммы с общим количеством по всем комментариям (использовалось 10 000 наиболее часто встречающихся униграмм, частоты вычислялись по отдельному набору комментариев), использование местоимений, определителей, служебных слов;
- 3) на основе предложений: средняя длина комментария в предложениях, средняя длина предложения в слова.

Итоговое качество предсказания пола пользователей в экспериментах было около 90%.

3. Исходные данные

При разработке решения задачи были рассмотрены две коллекции данных из социальной сети ВКонтакте и из YouTube. На основе структуры имеющихся данных был разработан метод определения влиятельности пользователей и адаптирован алгоритм поиска сообществ с учетом заданной специфики. Далее опишем подробнее структуру данных, описание алгоритма и результаты экспериментов.

3.1 Описание коллекций данных

В обоих датасетах представлен набор элементов контента со своими характеристиками и для каждого из них лента комментариев пользователей платформы. В первом случае это посты в группах ВКонтакте, во втором – страницы видеозаписей на YouTube. Структура данных естественным образом моделируется с помощью двудольного направленного графа, где вершинами являются элементы контента и пользователи, а ребрами факт комментирования пользователем элемента контента. Граф является двудольным, поскольку связей между пользователями в данных нет, элементы контента также не связаны друг с другом. Далее приведем детали по каждому датасету.

3.1.1 Датасет ВКонтакте

Коллекция представляет собой информацию о 43 751 посте из 17 групп, родственных по географическому фактору. Число уникальных пользователей, оставивших комментарии, 64 441. Число ребер 485 262. То есть каждый пост в среднем комментирует около 11 пользователей, один пользователь в среднем пишет комментарий в 7.5 постах.

Для каждого поста имеется следующая информация: дата публикации, тип поста, текстовое содержание, информация о прикрепленных фото и/или видео, число лайков, число репостов, число просмотров и некоторые другие сведения. Также есть лента комментариев пользователей под этим постом. По комментарию известны дата, идентификатор автора, текстовое содержание, количество ответов на данный комментарий, информация о прикрепленных фото и/или видео.

3.1.2 Датасет YouTube

В коллекции собраны комментарии к 2628 видео из 12 каналов на политические темы. Число уникальных пользователей, оставлявших комментарии, 1 246 060. Число ребер 2 536 874, то есть пересечение по пользователям довольно низкое – в среднем пользователь участвует в 2 видео. В среднем под видео 965 комментариев.

Для каждого видео имеется следующая информация: заголовок, описание, дата публикации, текстовое содержание, список тегов, число просмотров, число лайков, число дислайков, число комментариев, длительность видео. Для самих комментариев доступны: дата публикации, дата обновления, текстовое содержание, имя и идентификатор автора, число лайков и ответы на комментарий.

4. Предложенные алгоритмы

4.1 Алгоритм определения влиятельности пользователей

Учитывая двудольную природу модели данных, был предложен итеративный алгоритм вычисления влиятельности пользователя, подобный описанным в обзорах [25] и [26]. Однако в рассматриваемой задаче присутствуют некоторые отличия, из-за которых напрямую применить описанные методы нельзя. Например, все ребра направлены только от пользователей к постам/видео, но не наоборот. В связи с этим была предложена следующая методика.

Для удобства далее будем называть постом любой элемент контента, как пост так и видео. Основные принципы формирования влиятельности были предложены такие.

- пользователь влиятельный, если 1) на его комментарии много отвечают и 2) он комментирует важные посты.
- пост важный, если его 1) много комментируют и 2) комментируют влиятельные пользователи.

Далее следует формализация этих принципов. Пусть $G = (U, P, E)$ – двудольный граф, где U – пользователи, P – посты, E – ребра. Важность поста p обозначим $Reach(p)$, влиятельность пользователя u как $Inf(u)$. Схема взаимного влияния постов и пользователей дается следующими рекуррентными формулами (на шаге $k+1$):

$$Inf^{k+1}(u) = \sum_{p \in P} Reach^k(p) \cdot rep(u, p),$$

$$Reach^{k+1}(p) = \sum_{u \in U} Inf^k(u) \cdot eng(u, p).$$

Здесь $rep(u, p)$ обозначает количество ответов на комментарий пользователя u под постом p , а $eng(u, p)$ – вовлеченность пользователя u в обсуждение под постом p , измеряемую как общую длину всех его текстовых сообщений, написанных под этим постом.

В начале первого шага алгоритма влиятельность всех пользователей равна $Inf^0(u) = 1$, важность поста равна некоторой мере его популярности $Reach^0(p) = popularity(p)$. В качестве популярности можно взять число лайков, просмотров или число репостов. Далее алгоритм на каждом шаге обновляет значения влиятельности для всех пользователей и

важности для всех постов в соответствии с приведенными выше уравнениями. В конце каждого шага происходит нормировка значений:

$$Inf(u_i) = Inf(u_i) \frac{|U|}{\sum_k Inf(u_k)},$$

$$Reach(p_i) = Reach(p_i) \frac{|P|}{\sum_k Reach(p_k)}.$$

Для сходимости на рассмотренных датасетах алгоритму достаточно около 30 итераций, после чего ранжированный список наиболее популярных пользователей/постов перестает изменяться.

4.2 Алгоритм выявления сообществ пользователей

Поскольку явных социальных связей между пользователями в исходных данных нет, было решено использовать данные их активности в комментариях. Краугольная идея подхода состоит в том, что если два пользователя комментируют один и тот же пост, то их интересы близки. Чем больше общих постов они комментируют, тем сильнее между ними связь. Вычислив подобные связи между любыми двумя пользователями датасета, можно выявить разбиение пользователей на группы по интересам. Для этого предлагается построить граф связей между пользователями, и провести в нем поиск сообществ методом максимизации модулярности.

Дадим более формальное описание алгоритма. Рассматривается граф H с вершинами, соответствующими пользователям U . На первом этапе для каждой пары пользователей $u_i, u_j \in U$ вычисляется количество постов, в которых они оба оставляли комментарии. Это количество присваивается весу ребра в графе w_{ij} . Далее на взвешенном графе H запускается алгоритм максимизации модулярности Louvain [10].

5. Результаты экспериментов

Применение разработанного алгоритма к датасету ВКонтакте дает результаты, представленные в табл. 1 и 2. В первой показаны 10 постов с наибольшей важностью $Reach(p)$. Входящая степень – число уникальных пользователей-комментаторов. Видно, что нет явной корреляции между посчитанной важностью поста и числом лайков, несмотря на начальную инициализацию значений важности. Эксперименты показали, что другие варианты начальной инициализации (число просмотров, все поровну 1) дают такой же или близкий итоговый результат.

Табл. 1: Top-10 постов по важности в датасете ВКонтакте

Table 1. Top-10 important posts in Vkontakte

Пост	Важность	Вход.степень	Число лайков
106351883 1490574	174.34	39	478
33863926 366207	165.98	12	105
33863926 302228	159.97	26	623
106351883 1198977	156.19	33	48
106351883 1354344	155.87	14	25
106351883 1451193	152.16	60	213
176135883 39907	150.87	12	116
39490001 864801	150.16	5	29
150269648 306785	149.06	44	141
150269648 309372	147.62	32	304

Во второй таблице даны 10 наиболее влиятельных пользователей на основании значения $Inf(u)$. Исходящая степень – число прокомментированных пользователем постов. Хотя здесь можно наблюдать корреляцию активности комментирования и влиятельности, что довольно естественно, видно, что ранжирование по ним не совпадает. То есть, пользователь, который

пишет больше постов, не обязательно является более влиятельным. Играет роль также и важность постов, которые он комментирует. Этот результат находится в соответствии с выводами из работ по определению влиятельности пользователей в социальных сетях Twitter и Pinterest [25, 26].

Пользователи в табл. 2 с идентификатором, начинающимся со знака минус, соответствуют владельцам групп, размещавшим посты, то есть они комментировали от имени соответствующей группы.

Табл. 2. Топ-10 пользователей по влиятельности в датасете Вконтакте
Table 2. Top-10 influential users in Vkontakte

Пользователь	Влиятельность	Исх.степень
-69217442	7952.7	2787
225117304	1102.9	3119
580849174	606.0	1816
161096326	483.1	1122
144232132	320.1	876
616137950	290.3	562
295325982	268.0	1010
-195744122	256.1	589
510446853	250.3	865
572780482	244.9	754

Далее строился граф связей по интересам между пользователями. Для 64441 вершин граф содержит 6 584 434 ребер связей. Метод поиска сообществ находит около 30 сообществ (в зависимости от запуска). Самые большие сообщества имеют размеры 18 059, 12 946, 12 635, 9920, 4752, 1172 вершин. Если проверить, в каких сообществах оказываются топ-100 влиятельных пользователей, то получится, что 70 из них в наибольшем, в следующем по размеру 16, затем 10 и наконец 4 в сообществе с 9920 вершинами. Кроме того, можно посчитать среднюю влиятельность пользователей внутри сообщества. Результаты приведены в табл. 3. Самое большое сообщество оказывается самым влиятельным в среднем. Маленькие сообщества почти всегда содержат маловлиятельных пользователей. Из таблицы наблюдается положительная корреляция между размером сообщества, числом наиболее влиятельных пользователей и его средней влиятельностью. Однако, сообщества размеров 1054 и 618 выбиваются из этой зависимости. Можно предположить, что они образуют отдельные группы со своими лидерами мнений, и потому могут быть интересны для более детального анализа.

Табл. 3. Распределение влиятельных пользователей по сообществам в датасете Вконтакте
Table 3. Distribution of influential users over communities.

Размер сообщества	Средняя влиятельность	Число влиятельных из топ-100
18 059	3.141	70
12 946	1.505	16
12 635	1.217	10
9920	0.714	4
4752	0.729	0
1172	0.243	0
1088	0.429	0
1054	2.220	0

618	2.799	0
549	0.886	0
394	0.619	0
386	1.212	0
126	0.133	0
97	0.069	0
72	0.158	0
36	0.261	0
17	0.014	0
15	0.000	0
14	0.004	0
12	0.009	0
11	0.016	0
8	0.003	0
7	0.000	0
6	0.000	0
5	0.000	0
3	0.000	0
2	0.000	0
1	0.000	0

Приведем также результаты поиска важных видео и влиятельных пользователей для датасета YouTube (табл 4, 5). В целом наблюдается аналогичная картина, что и для соответствующих результатов на датасете ВКонтакте.

Анализ сообществ здесь не проводился, поскольку из-за большого общего числа уникальных пользователей и в то же время большого числа различных комментаторов на одно видео, граф связей по интересам между ними содержит существенное число ребер, что требует существенных вычислительных ресурсов для применения метода Louvain. В таком случае в качестве альтернативы могут быть рассмотрены методы поиска сообществ, имеющие меньшую сложность алгоритма, либо же можно сократить число ребер графа путем введения минимального порога на число общих комментариев для пользователей.

Табл. 4. Топ-10 видео по важности в датасете YouTube
Table 4. Top-10 important videos in YouTube.

Пост	Важность	Вход.степень	Число лайков
95ReakCrKX0	108084	17966	323038
AR6ovvs6lHg	82268	63466	704654
91yZNamfN1I	73368	21811	293005
vps43rXgaZc	71278	103523	1247927
ttPXXyUyx6Q	69170	17568	255072
QDHYQ9Nd-GU	64301	37334	510176
PWt27h_scaY	61658	13867	164038
602k5_r9ZCs	59179	7389	78470
SgV0-0puqWM	58458	19046	216249
skFNJ3tB67M	57957	12069	134189

Табл. 5. Топ-10 пользователей по влиятельности в датасете YouTube
Table 5. Top-10 influential users in YouTube.

Пользователь	Влиятельность	Исх. степень
UCMCgOm8GZkHp8zJ6l7_hluA	765623	22
UCsAw3WynQJm7tMy093y37A	35067	15
UC-TsVR4jjKT7ZSXoiYQZNIQ	12815	5
UChG-0MFkgZxAwcxdxi0m_U_JA	12498	2
UC-2rXPfDuLZn2qKmXaR5jWw	11970	1
UCEbCKFjRBtSAa83ROejtyeA	11112	4
UC53JOxGaJlh6vD0ZHUjWhsA	9987	1
UC5RRYPRf06pHSgUOiCy97JA	9391	1
UC7LzUL-OCMf7ehTZn8lIXEw	9109	6
UCegPwkRt1G-Y87TEuvyg64w	8664	1

6. Заключение

В рамках НИР проведен обзор методов поиска пользователей социальных сетей. Описаны несколько популярных классов подходов, основанных на информации о графовой структуре сети: методы на основе локальной оптимизации, вероятностные модели, методы распространения меток и методы, основанные на эго-сообществах. Затем дан обзор методов выявления влиятельных пользователей в социальных сетях. Сначала описаны подходы для случая изначально скрытого графа, затем более подробно рассмотрены несколько алгоритмов для случая полной информации о графе. На основе этих методов разработан алгоритм, предложенный для решения поставленной в работе задачи. Далее следует описание неклассических методов поиска сообществ – основанные на деятельности, на основе темы контента и на основе взаимодействия. В конце обзора приводится пример использования текстовой информации сообщений пользователей.

На втором этапе проведены экспериментальные исследования разработанного метода на двух коллекциях данных: из сети ВКонтакте и из YouTube. В обоих датасетах представлены данные о некотором контенте, размещаемом на платформе (посты для ВКонтакте и видео для YouTube) и комментарии пользователей к ним. Модель данных представляет собой двудольный граф, где двумя множествами вершин являются пользователи и контент, а ребра соответствуют комментированию. Итеративный алгоритм вычисляет значение влиятельности для каждого пользователя и величину важности элемента контента. В соответствии с основной идеей алгоритма, пользователь влиятельный, если 1) на его комментарии много отвечают и 2) он комментирует важные элементы контента. Важность элемента контента, в свою очередь, определяется тем, что его 1) много комментируют и 2) комментируют влиятельные пользователи. В итоге самыми влиятельными обычно оказываются наиболее активные пользователями, однако корреляция между влиятельностью и числом комментариев не полная – активный комментатор всего подряд не обязательно будет влиятельным.

После этого были выделены сообщества пользователей, основанные на их общих интересах: чем больше общих элементов контента комментируют два пользователя, тем теснее между ними связь. Алгоритм находит квазиоптимальное разбиение вершин графа на группы, максимизирующее модулярность. По итогам экспериментов наблюдается положительная корреляция между размером сообщества, числом наиболее влиятельных пользователей в нем и средней влиятельностью пользователей сообщества. Некоторые сообщества выбиваются из этой зависимости и потому могут быть интересны для экспертного анализа – возможно, они отражают формирование активных групп по интересам.

Список литературы / References

[1] Conrad Lee, Fergal Reid et al. Detecting highly overlapping community structure by greedy clique expansion. arXiv preprint arXiv:1002.1827, 2010, 10 p.

[2] Andrea Lancichinetti, Filippo Radicchi et al. Finding statistically significant communities in networks. *PLoS one*, vol. 6, issue 4, 2011, e18961.

[3] Aaron McDaid and Neil Hurley. Detecting highly overlapping communities with model-based overlapping seed expansion. In Proc. of the International Conference on Advances in Social Networks Analysis and Mining, 2010, pp. 1120-1119.

[4] Morten Mørup and Mikkel N Schmidt. Bayesian community detection. *Neural computation*, vol. 24, issue 9, 2012, pp. 2434-2456.

[5] Jaewon Yang and Jure Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In Proc. of the Sixth ACM International Conference on Web Search and Data Mining, 2013, pages 587–596.

[6] Jaewon Yang, Julian McAuley, and Jure Leskovec. Community detection in networks with node attributes. In Proc. of the IEEE 13th international conference on data mining, 2013, pp. 1151-1156.

[7] Mark E.J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, vol. 69, issue 2, 2004, article id 026113.

[8] Mingming Chen, Konstantin Kuzmin, and Boleslaw K Szymanski. Community detection via maximization of modularity and its variants. *IEEE Transactions on Computational Social Systems*, vol. 1, issue 1, 2014, pp. 46-65.

[9] Nicolas Dugué and Anthony Perez. Directed Louvain: maximizing modularity in directed networks. Research Report hal-01231784, Université d’Orléans. 2015, 15 p.

[10] Vincent D. Blondel, Jean-Loup Guillaume et al. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008, article id P10008, 15 p.

[11] Jierui Xie, Boleslaw K. Szymanski, and Xiaoming Liu. SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In Proc. of the IEEE 11th International Conference on Data Mining Workshops, 2011, pp. 344-349.

[12] Zhi-Hao Wu, You-Fang Lin et al. Balanced multi-label propagation for overlapping community detection in social networks. *Journal of Computer Science and Technology*, vol. 27, issue 3, 2012, pp. 468-479.

[13] Steve Gregory. Finding overlapping communities in networks by label propagation. *New journal of Physics*, vol. 12, issue 10, 2019, article id 103018.

[14] Sucheta Soundarajan and John E Hopcroft. Use of local group information to identify communities in networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 9, issue 3, 2015, pp. 1-27.

[15] Michele Coscia, Giulio Rossetti et al. Demon: a local-first discovery method for overlapping communities. In Proc. of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 615-623.

[16] Bradley S. Rees and Keith B. Gallagher. Egocustering: overlapping community detection via merged friendship-groups. *Lecture Notes in Social Networks*, vol. 6, 2013, pp. 1-20.

[17] Nazar Buzun, Anton Korshunov et al. Egolp: Fast and distributed community detection in billion-node social networks. In Proc. of the IEEE International Conference on Data Mining Workshops, 2014, pp. 533-540.

[18] Mohammed Ali Al-Garadi, Kasturi Dewi Varathan et al. Analysis of online social network connections for identification of influential users: Survey and open research issues. *ACM Computing Surveys (CSUR)*, vol. 51, issue 1, 2018, pp. 1-37.

[19] Colin Cooper, Tomasz Radzik, and Yiannis Siantos. A fast algorithm to find all high degree vertices in power law graphs. In Proc. of the 21st International Conference on World Wide Web, 2012, pp. 1007-1016.

[20] Konstantin Avrachenkov, Prithwish Basu et al. Online myopic network covering. UMass Technical Report UM-CS-2012-034, arXiv preprint arXiv:1212.5035, 2012, 18 p.

[21] Konstantin Avrachenkov, Prithwish Basu et al. Pay few, influence most: Online myopic network covering. In Proc. of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2014, pp. 813-818.

[22] Katja Filippova. User Demographics and Language in an Implicit Social Network. In Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 1478–1488.

- [23] Konstantin Avrachenkov, Nelly Litvak et al. Quick detection of nodes with large degrees. *Internet Mathematics*, 10(1-2):1–19, 2014.
- [24] 22 Danil Shaikhelislamov, Mikhail Drobyshevskiy et al. Three-step algorithms for detection of high degree nodes in online social networks. In *Proc. of the Ivannikov Memorial Workshop (IVMEM)*, 2020, pp. 43-48.
- [25] Daniel M Romero, Wojciech Galuba et al. Influence and passivity in social media. *Lecture Notes in Computer Science*, vol. 6913, 2011, pp. 18-33.
- [26] Zhiguo Zhu, Jingqin Su, and Liping Kong. Measuring influence in online social network based on the user-content bipartite graph. *Computers in Human Behavior*, vol. 52, 2015, pp. 184-189.
- [27] Weishu Hu, Zhiguo Gong et al. Identifying influential user communities on the social network. *Enterprise Information Systems*, vol. 9, issue 7, 2015, pp. 709-724.
- [28] David Darmon, Elisa Omodei, and Joshua Garland. Followers are not enough: A multifaceted approach to community detection in online social networks. *PloS one*, vol. 10, issue 8, 2015, article id e0134860.

Информация об авторах / Information about authors

Роман Константинович ПАСТУХОВ – научный сотрудник. Его научные интересы включают распределенные вычисления, анализ графов, сложные сети, модели случайных графов, анализ социальных сетей, большие данные.

Roman Konstantinovich PASTUKHOV is a researcher. His scope of scientific interests include distributed computing, graph analysis, complex networks, random graph models, social network analysis, big data.

Михаил Дмитриевич ДРОБЫШЕВСКИЙ – к.ф.-м.н., научный сотрудник. Его научные интересы включают приложения машинного обучения, сложные сети, модели случайных графов, вложения графов, объяснимый искусственный интеллект.

Mikhail Dmitrievich DROBYSHEVSKIY is a Ph.D., researcher. His scope of scientific interests includes machine learning applications, complex networks, random graph models, graph embedding, explainable AI.

Денис Юрьевич ТУРДАКОВ – к.ф.-м.н., заведующий отделом “Информационные системы” ИСП РАН, доцент МГУ. Сфера научных интересов: машинное обучение, интеллектуальный анализ данных, извлечение информации, обработка естественного языка, сложные сети, анализ социальных сетей, большие данные.

Denis Yurievich TURDAKOV – Ph.D., head of the Information Systems Department at ISP RAS, associated professor at MSU. Research interests: machine learning, data mining, information extraction, natural language processing, complex networks, social network analysis, big data.