УДК 004 информационные технологии

Алгоритмы поиска подграфов социального графа, включающих заданное множество пользователей

Авторы:
Д. С. Филиппов, Университет ИТМО, г. Санкт-Петербург
А. А. Фильченков, Университет ИТМО, г. Санкт-Петербург

В последнее время очень активно развивается анализ графов социальных сетей – это помогает находить новую информацию и использовать ее в различных целях: начиная военными, заканчивая маркетинговыми. В данной работе мы рассматриваем проблему поиска сообществ в социальных сетях: по нескольким выделенным вершинам в графе социальной сети с атрибутами у вершин требуется найти сообщество, которое их объединяет. Сообщество должно быть плотно связанным, а также содержать большинство вершин из запроса (но необязательно все). Задача имеет актуальность во многих областях: поиск террористических группировок (по нескольким людям найти всю группировку), поиск заболевших людей (по нескольким зараженным и социальному графу выявить потенциальную группу новых зараженных), маркетинг для сообществ в социальных сетях (таргетировать только в определенные сообщества людей), и т.п.

Целью нашей работы является поиск алгоритма, оптимально решающего поставленную задачу. В процессе исследования было доказано, что поставленная задача является NP-полной, однако, это не означает, что ее нельзя решить с хорошим приближением. Требуется провести анализ метрик размеров и плотности сообществ, полученных в результате работы различных алгоритмов, а впоследствие предложить новый алгоритм, решаюший поставленную задачу лучше существующих решений.

В качестве базовых решений были рассмотрены недавние решения (до 2015 года), решающие похожую задачу. Наиболее интересными оказались решения Barbieri et al. [1] и Zheng et al. [2]. В первой статье авторы ищут минимальный по размеру k-core с максимальным k, содержащий все вершины из запроса, а авторы второго алгоритма предлагают комбинированное решение, основанное на нескольких идеях. Однако, ни одна из этих статей не рассматривает возможность шума в запросе, когда некоторые из выделенных вершин не относятся к сообществу и плохо коррелируют с остальными вершинами. Однако, на наш взгляд именно задача с возможным шумом в запросе является более актуальной, поэтому наш алгоритм будет покрывать и такой случай тоже.

На данный момент было проведено исследование существующих решений, а также реализованы и проведен анализ двух базовых алгоритмов, представленных ранее. В данный момент идет разрабатывание нового решения, основанного на представленных алгоритмах, однако рассматривающего также случай шума в запросах, а также новые идеи, которые не применялись в рассмотренных алгоритмах. Уже есть первый прототип алгоритма, теоретическое исследование которого показывает, что у него есть хорошие шансы показать неплохие результаты на реальных данных. В ближайшее время этот алгоритм будет реализован, после чего будет проведен ряд исследований, подтверждающий или опровергающий его оптимальность

Практические результаты, проведенные на двух рассмотренных существующих алгоритмах (исследования проводились на датасетах DBLP и Facebook), позволили добиться тех же результатов, что и в статьях авторов, а также позволили немного изменить алгоритмы, чтобы они могли поддерживать шум в запросах. Однако, так как изначально эти алгоритмы не нацелены на такой случай, результаты получились не очень впечатляющими, что мы надеемся исправить в нашем алгоритме.

Ссылки:

[1] N. Barbieri и др. Efficient and effective community search. Data Mining and Knowledge Discovery, volume 29, issue 5, 2015
[2] D. Zheng и др. Querying Intimate-Core Groups in Weighted Graphs. IEEE 11th International Conference on Semantic Computing (ICSC), 2017