Выделение пересекающихся сообществ в социальных графах по мажоритарному признаку соседей

Чесноков Владислав Олегович

Аспирант

Кафедра ИУ8 МГТУ имени Н.Э. Баумана, Москва, Россия

E-mail: v.o.chesnokov@yandex.ru

Выделение сообществ в графах — важная задача во многих научных областях: биологии, социологии, информатике. В социальных графах вершинами являются люди, а ребрами — связи между ними: родственные, дружественные, рабочие и др. Взрывной рост количества пользователей таких социальных сетей, как Facebook и Вконтакте, и открытый доступ к большому объему данных предоставил большой простор для различного рода исследований: определение социальных групп для таргетированной рекламы, выделение лидеров общественного мнения с целью распространения дезинформации или использования в информационном противоборстве [1], анализ кредитоспособности человека по профилю и многих других.

В работе [2] был описан алгоритм разбиения социальных графов на подмножества вершин по их признакам. Рассмотрим неориентированный невзвешенный граф $G'(V', E')$ с диаметром, равным двум, и для которого есть такая вершина u, что

$$\forall v \in V', v \neq u \; \exists \{u, v\} \in E'.$$

Определим

$$V = V' \setminus \{u\}$$

$$E = E' \setminus \{\{u, v\} | v \in V\}.$$

Пусть у каждой вершины есть набор признаков (атрибутов) из множества признаков S, т.е. задана функция $f : V' \to 2^S$. Очевидно, что не всегда существует возможность получить полную информацию о признаках вершины (из-за ошибок при передаче данных, цензуры, при неизвестном или неуказанном состоянии и т.п.). Поэтому определим функцию $f' : V' \to 2^S$ получения признаков такую, что

$$\forall v : f'(v) \subseteq f(v).$$

Задача выделения сообществ состоит в том, чтобы найти такое покрытие множества вершин V графа $G(V, E)$ при имеющейся функ-

ции f', в котором «схожие» вершины принадлежат одному множеству.

Ключевая идея предложенного алгоритма основана на гипотезе о триадной структурке социальных сетей [3]: если у вершины A есть ребра к вершинам B и C, то с большой вероятностью есть ребро между B и C. Соответственно, если большинство соседей некоторой вершины имеют общий признак, то он, скорее всего, присутствует и у данной вершины. Таким образом, на каждом шаге алгоритм обновляет признаки всех вершин по признакам соседей до тех пор, пока будет хотя бы одно изменение. Сложность данного алгоритма линейно зависит от количества ребер в графе.

Алгоритм был опробован на данных из социальных сетей Facebook и Twitter, собранных в рамках проекта SNAP [4], которые представляют собой выборку графов ближайшего окружения пользователей социальных сетей с размеченными сообществами (т.н. «ground-truth communities»). По сравнению с алгоритмом CESNA [5], на выборе из сети Facebook метрика F_1-score для разработанного алгоритма выше на 10%, а на выборке из сети Twitter имеет соспоставимую величину.

Литература

1. Вельц С. В. Моделирование информационного противоборства в социальных сетях на основе теории игр и динамических байесовских сетей. // Инженерный журнал: наука и инновации. Электронное научное техническое издание. 2013. № 11(23).
2. Чесноков В. О., Ключарёв П. Г. Выделение сообществ в социальных графах по множеству признаков с частичной информацией // Наука и образование. Электронное научно-техническое издание. 2015. № 9. 188–199.
3. Granovetter M. The Strength of Weak Ties. // American Journal of Sociology. 1973. Vol. 78, no. 6. P. 1360–1380
4. SNAP datasets. `http://snap.stanford.edu/data/index.html`
5. Mcauley J., Leskovec J. Discovering Social Circles in Ego Networks. // ACM Transactions on Knowledge Discovery from Data. 2014. February. Vol. 8, № 1. P. 4:1-4:28.