УДК 004.62

МЕТОДЫ МНОГОФАКТОРНОЙ КЛАСТЕРИЗАЦИИ ДАННЫХ

Телевной А.Д. (Университет ИТМО)
Научный руководитель – к.ф.-м.н., доцент Иванов С.Е.
Университет ИТМО, Санкт-Петербург

Аннотация. В данной работе рассматривается возможность существенного улучшения результатов кластеризации с помощью использования усреднения вычислений совокупности существующих метрик центральности, предложив собственную алгоритмическую схему кластеризации, включающую в себя несколько различных способов усреднения значений. Осуществлено сравнение результатов выполнения «гибридного» алгоритма с рядом наиболее применяемых алгоритмов кластеризации: Infomap, Label Propagation, FastGreedy, Louvain.

Введение. Проведение кластерного анализа графовой модели сопряжено с рядом объективных трудностей: не существует однозначно объективного критерия качества кластеризации, количество кластеров заранее неизвестно и устанавливается субъективным образом экспертом, одно из ключевых значений при кластеризации играет выбор метрики, которая также устанавливается экспертом субъективным образом. Несмотря на то, что исследования в данной предметной области проводятся уже много лет, проведенный анализ публикаций тематической области позволяет говорить о том, что исследователи все еще испытывают потребность в разработке эффективных методов анализа графов для крупномасштабных сетевых структур. В данной работе сконцентрировано внимание на возможности существенного улучшения результатов кластеризации с помощью использования на различных этапах усреднения вычислений совокупности существующих метрик центральности, предложив собственную алгоритмическую схему кластеризации.

Основная часть. В работе описана процедура выполнения «гибридного» алгоритма кластеризации данных: 1. Построение списка весов вершин с помощью последовательного применения метрик центральности связанного социального графа. 2. Нормализация максимального значения веса вершины из полученного списка методом деления значения веса каждой вершины на максимальный вес. 3. Вычисление среднего квадратичного значения веса вершин графа из списка нормализованных весов вершин, полученных на предыдущем шаге. 4. Преобразование списка вершин путем упорядочивания исходного перечня по возрастанию весов вершин. 5. Кластеризация итогового списка на кластеры фиксированного размера.

Необходимость усреднения результатов нескольких метрик объясняется желанием получения более репрезентативной картины, нежели при использовании единственной метрики.

Осуществлено сравнение результатов выполнения «гибридного» алгоритма с рядом наиболее применяемых алгоритмов кластеризации: Infomap, Label Propagation, FastGreedy, Louvain. Сравнение осуществлялось на специально собранных датасетах пользователей, являющихся участниками онлайн-сообществ социальной сети «ВКонтакте». Для сравнения результатов использовались значения модулярности и времени выполнения алгоритма.

Выводы. При количестве вершин менее 200, предложенный «гибридный» алгоритм показал наивысшее значение модулярности. Результаты алгоритма Label Propagation при количестве вершин более 1000 отмечается минимальным показателем модулярности. Наибольший показатель времени выполнения имеет алгоритм Infomap. Возможность применения нескольких метрик при «гибридном» алгоритме позволяет улучшить оценку качества кластеризации. Предложенный метод позволяет оптимизировать ресурсы при кластеризации, что является важным достоинством с учетом активно возрастающих объемов данных.