УДК 519.171.1

Разработка и реализация моделей и алгоритмов выделения сообществ в графах взаимодействующих объектов

С.А. Шилин[1]

[1]Московский физико-технический институт (государственный университет)

Актуальность

1. Распознавание структуры, скрытой в реальных социальных сетях, является ключевой задачей, решение которой необходимо для понимания организации сложных сетей.
2. Кластеризация элементов сложных сетей позволяет анализировать их на более высоком уровне, что в разы проще.

Социальный граф

1. Одна большая общая компонента связности
2. Распределение на степенях вершин
3. Среднее расстояние
4. Коэффициент кластеризации
5. Структура сообществ

Сообщество

1. Формально структура графа образована сообществами, если он отличается от случайного графа.
2. С содержательной точки зрения - это группа вершин сети, участники которой связаны друг с другом значительно теснее, чем с остальными вершинами сети.

Проблема

1. Проблема выделения сообществ есть задача анализа графов. Существует множество алгоритмов с использованием методов из различных дисциплин. Не все алгоритмы надёжны и могут быть применены на практике. Не понятно, как хранить и анализировать графы очень больших размеров.
2. Задача - Для конкретной задачи разработать алгоритм выделения сообществ, который покажет хорошие результаты для конкретной задачи в сравнении с существующими методами.

Пример: кластеризация веб-доменов

Есть данные о посещениях пользователями различных доменов, есть граф, где в качестве узлов выступают домены, а в качестве рёбер - аффинити между доменами.

Аффинити между доменами x и y - это выборочная оценка того, насколько события «посещение юзером u домена x» и «посещение юзером u домена y» близки к независимости.

Критерии качества

После того как отработал алгоритм выделения сообществ, необходимо оценить качество получившегося результата.

- Модулярность

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \delta(C_i, C_j), \tag{1}$$

A - матрица смежности; d - степень вершины; C - метка вершины (номер сообщества, к которому относится вершина); 5 = 1, если *Ci* = *Cj*, иначе 0 - дельта-функция, m - кол-во ребер в графе.

- Редакторское расстояние для разбиений (split-join distance)
- Нормализованная взаимная информация

Алгоритмы: igraph

- Betweenness коэффициент «центральности по посредничеству» (Betweenness)
- Fastgreedy жадная оптимизация функции модулярности

- Multilevel многоуровневая оптимизация функции модулярности с эвристикой
- LabelPropogation присвоение меток к каждой вершине
- Walktrap короткие случайные блуждания не приводят к выходу из текущего сообщества
- Infomap случайное блуждание, основанное на понятии информационных потоков в сетях, кодирования и сжатия информации
- Eigenvector собственных векторах матрицы модулярности, которая получается из матрицы смежности

Результаты работы. Тестирование алгоритмов

1. Моделирование данных
2. Генерация графа
3. Зашумление графа
4. Реальные данные

Литература

1. *Корелин, В.Н.* Применение модифицированного алгоритма LSH для кластеризации внешнего окружения веб-пространства университетов [Электронный ресурс] = Clastering of the external web environment of universities using a modified LSD algorithm / В.Н.Корелин, И.С.Блеканов, С.Л. Сергеев. - Электрон. текстовые дан. (1 файл: 318 КБ). // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета = St. Petersburg state polytechnical university journal. Computer science. Telecommunications and control systems. Сер.: Информатика. Телекоммуникации. Управление: научное издание. - Санкт-Петербург, 2015. - № 5 (229). - Загл. с титул. экрана. - Электронная версия печатной публикации. - Свободный доступ из сети Интернет (чтение, печать, копирование). - Текстовый файл. - Adobe Acrobat Reader 7.0. - <URL:http://elib.spbstu.ru/dl/2/j16-60.pdf>