

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

Сергей Шилин

Международная конференция SCVRT2015
24-27 ноября 2015 г.



Содержание

1. Введение

- Социальный граф
- Сообщество
- Проблема

2. Выделения сообществ

- Критерии качества
- Алгоритмы: igraph
- Тестирование алгоритмов

3. Вывод

- Актуальность задачи

Введение: Социальный граф

- 1 Одна большая общая компонента связности

Введение: Социальный граф

- 1 Одна большая общая компонента связности
- 2 Распределение на степенях вершин

Введение: Социальный граф

- 1 Одна большая общая компонента связности
- 2 Распределение на степенях вершин
- 3 Среднее расстояние

Введение: Социальный граф

- 1 Одна большая общая компонента связности
- 2 Распределение на степенях вершин
- 3 Среднее расстояние
- 4 Коэффициент кластеризации

Введение: Социальный граф

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

Введение: Сообщество

- Формально : структура графа образована сообществами, если он отличается от случайного графа.

Введение: Сообщество

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

Введение: Проблема

- Проблема выделения сообществ есть задача анализа графов.

Введение: Проблема

- 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.

Введение: Проблема

- 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.

Введение: Проблема

- 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.
- **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.

Пример: кластеризация веб-доменов



Рис. 1 : Кластеризация графа для 1285 доменов

Критерии качества

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

- Модулярность

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

Критерии качества

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

- Модулярность

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

- Редакторское расстояние для разбиений (split-join distance)

Критерии качества

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

- Модулярность

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

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

Алгоритмы: igraph

- **Betweenness** коэффициент «центральности по посредничеству» (Betweenness)

Алгоритмы: igraph

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

Алгоритмы: igraph

- **Betweenness** коэффициент «центральности по посредничеству» (Betweenness)
- **Fastgreedy** жадная оптимизация функции модулярности
- **Multilevel** многоуровневая оптимизация функции модулярности с эвристикой

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

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

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

Алгоритмы: igraph

- **Betweenness** коэффициент «центральности по посредничеству» (Betweenness)
- **Fastgreedy** жадная оптимизация функции модулярности
- **Multilevel** многоуровневая оптимизация функции модулярности с эвристикой
- **LabelPropagation** 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 векторов матрицы модулярности, которая получается из матрицы смежности

Тестирование алгоритмов

- Моделирование данных
 - 1 Генерация графа
 - 2 Зашумление графа

Тестирование алгоритмов

- Моделирование данных
 - ① Генерация графа
 - ② Зашумление графа
- Реальные данные

Актуальность задачи

- Распознавание структуры, скрытой в реальных социальных сетях, является ключевой задачей, решение которой необходимо для понимания организации сложных сетей.

Актуальность задачи

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