Метод спектрального анализа сети цитирования научных журналов

Бредихин С.В., Щербакова Н.Г.
ИВМиМГ СО РАН,
Новосибирск

Акторами сети цитирования (СЦЖ) являются научные журналы (агрегированное множество статей), связи строятся на основании бинарного отношения цитирования R.

СЦЖ моделируется взвешенным орграфом

$G^{WD} = (V, E)$, журналы соответствуют вершинам $V = \{v_n\}$. Если $v_i \, R \, v_j$, то дуга $e = (v_i, v_j) \in E \subseteq V \times V$.

Требуется построить разбиение $C = \{C_1, C_2, \ldots, C_k\}$ множества V на кластеры C_i, при условии

$C_1 \cup C_2 \ldots \cup C_k = V$; $C_1 \cap C_2 \ldots \cap C_k = \varnothing$. (1)

Построение разбиения (1) выполняется на основе решения оптимизационной задачи нахождения минимального размера разреза G^{WD}:

$$Cut(C) = \sum_{h=1}^{k} E_h, \text{ где } E_h = \sum_{v_i \in C_h} \sum_{v_j \notin C_h} a_{ij}, \qquad (2)$$

a_{ij} – элемент матрицы смежности A графа G^{WD}.

Для решения (2) используются методы, основанные на вычислении собственных векторов матрицы A либо матрицы Лапласа $L=D-A$ (все собственные значения вещественные, неотрицательные).

Разработаны и реализованы алгоритмы S_1 и S_2 для разбиения неориентированных и ориентированных графов.

В алгоритме S_1 использованы идеи [Shi, Malik, 2000], [Meila, Shi, 2001] разбиения вершин неориентированного графа G^{WU} на основе оптимизации нормализованного разреза:

$$NCut(C_1, C_2) = \frac{Cut(C_1, C_2)}{D_{C_1}} + \frac{Cut(C_2, C_1)}{D_{C_2}}.$$

$$D_{C_k} = \sum_{i \in C_k} deg(i).$$

Решается уравнение вида $L\mathrm{x} = \lambda D\mathrm{x}$. (3)

D – диагональная матрица, $d_{ii} = deg(i)$, $L = D - A$.

Граф G^{WD} приводится к виду G^{WU} с помощью преобразования $A = A + A^{T}$.

Бикластеризация G^{WU} на основе второго минимального собственного вектора *v,* являющегося решением (3).

Вершина *i* отображается на элемент *i* вектора *v*.

Можно свести к решению для симметрично нормализованной матрицы Лапласа:

$$L^{sym} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2},$$

$$L^{sym} z = \lambda z, \; z = D^{1/2} x.$$

Деление *v* на две части “приблизительно” равных элементов и соотнесение их с кластерами, например, применение *к-средних* для к = 2.

Множественный нормализованный разрез:

$$MNCut(C) = \sum_{k=1}^{K} \sum_{k' \neq k} \frac{Cut(C_k, C_{k'})}{D_{C_k}}.$$

Переход к модели случайного блуждания с матрицей переходных вероятностей $P = D^{-1}A$.

В работе [Meila, Shi, 2001] показано, что достаточно вычислить $K + 1$ наибольших собственных векторов матрицы P (если λ, x решение Px = λx, то (1 – λ), x – решение уравнения (3)).

Наличие n независимых собственных векторов обеспечивается симметричностью матрицы A и обратимостью матрицы D.

Алгоритм S_1 разбиения G^{WU}

Вход: Матрица A; k – число кластеров.

1. Построение стохастической матрицы $P = D^{-1}A$.

2. Вычисление k собственных векторов $x_1, x_2, \ldots, x_k$ матрицы P, соответствующих собственным числам $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_k$.

3. Построение $n \times k$ матрицы, со столбцами $x_1, x_2, \ldots, x_k$. Строка i соответствует вершине i, графа и рассматривается как точка пространства $\mathbb{R}^k$, к которым применяется алгоритм *k-средних*.

Выход: Таблица инцидентности $\{(i, j) \mid j \in C_i\}$.

Алгоритм S_2 использует идею [Meila, Pentney, 2007] разбиения вершин орграфа G^{WD} на основе оптимизации универсального взвешенного разреза.

Сопоставим вершине i веса t_i и t_i'. Определим диагональные матрицы $T = diag(t_{ii})$ $T' = diag(t_{i\,i}')$.

Вес кластера C_k:

$$T_{C_k} = \sum_{i \in C_k} t_i \,.$$

Универсальный взвешенный разрез:

$$WCut(C) = \sum_k \sum_{k' \neq k} \frac{Cut(C_k,\ C_{k'})}{T_{C_k}}$$

$$Cut(C_k, C_{k'}) = \sum_{i \in C_k} \sum_{j \in C_{k'}} t_i' a_{ij} \,.$$

Разбиение *C* представим в виде матрицы *X* размера (*n*×*K*), в которой *k*−й столбец $\mathbf{x}_k$ является индикаторным вектором кластера C_k, т. е. $x_k(i) = 1$, если $i \in C_k$.

A = T'A; D = T'D.

$$WCut(C) = \sum_{k=1}^{K} y_k^T B\, y_k$$

$$B = T^{-1/2}(D - A)T^{-1/2},$$

$$y_k = T^{1/2}x_k / \sqrt{T_{C_k}}.$$

Показано, что задача сводится к нахождению собственных векторов эрмитовой части матрицы *B* - $H(B) = ½(B + B^T)$.

Алгоритм S_2 разбиения G^{WD}

(общий вид)

Вход: Матрица A; k – число кластеров.

1. Построение матриц $A=T'A$; $D=$ diag$\{out_deg_i\}$.

2. Построение эрмитовой матрицы

$H(B) := 1/2T^{-1/2}(2D - A - A^{T})T^{-1/2}$.

3. Построение $n \times K$ матрицы Y со столбцами, состоящими из собственных векторов $H(B)$, соответствующих собственным числам $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_k$.

4. Кластеризация строк матрицы $X = T^{-1/2}Y$ (или нормализованных строк матрицы Y, норма 1) как точек пространства $\mathbb{R}^k$.

Выход: Таблица инцидентности $\{(i, j) \mid j \in C_i\}$.

Детали алгоритма S_2.

1. Построение матриц:

$P = D^{-1} A$ – матрицы случайного блуждания;

P_u – матрица унифицированных вероятностей перехода;

$P_\lambda = \lambda P + (1-\lambda)P_u$, ($\lambda = 0{,}85$), для которой вычисляется вектор π стационарного распределения.

Неотрицательная неразложимая и апериодическая матрица P_λ является примитивной, поэтому стационарный вектор единственный.

Полагаем $A = P_\lambda$, $T = T' = \Pi = \mathrm{diag}(\pi_{ii}), \pi_{ii} = \pi_i$.

Детали алгоритма S_2 (продолжение)

2. Построение симметричной матрицы $H(B)$:

$$H(B) = I - \left(\Pi^{1/2} P_\lambda \Pi^{-1/2} + \Pi^{-1/2} P_\lambda^T \Pi^{1/2}\right)/2 .$$

3. Вычисление собственных векторов $x_1, x_2, \ldots, x_k$ матрицы $H(B)$ и построение $n \times K$ матрицы Y с ортонормированными столбцами, состоящими из собственных векторов $H(B)$, соответствующих собственным числам $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_k$.

4. Кластеризация строк матрицы $X = T^{-1/2} Y$ (или нормализованных строк матрицы Y, норма 1) как точек пространства $\mathbb{R}^k$.

Выход: Таблица инцидентности $\{(i, j) \mid j \in C_i\}$.

Сложность алгоритмов S_1 и S_2

$G = (V, E)$; $|V| = n$

S_1 : $O(dnm) + O(nkdi)$

S_2 : $O(dnm) + O(nkdi) + O(n^2 i)$

1. Вычисление m собственных векторов (LAPACK):$O(dnm)$, d – среднее число ненулевых элементов строки матрицы.

2. Алгоритм k-means (алгоритм Лойда): $O(nkdi)$ k – число кластеров, d – размерность векторов, i – число итераций.

3. Вычисление главного собственного вектора (степенной метод): $O(n^2 i)$.

Вычислительный эксперимент

Апробация алгоритмов S_1 и S_2 выполнена на «сырых» данных о цитировании научных журналов, извлеченных из распределенной библиографической базы данных *RePEc*.

Отношение $v_i \, R \, v_j$, в данном случае трактуется так: «журнал v_j цитирует журнал v_i». Моделью СЦЖ является главная слабо связная компонента G^{WD} с параметрами: $|V| = 1729$, $|E| = 135702$.

Результаты

- Разработаны и реализованы 2 алгоритма разбиения вершин G^{WD} и G^{WU} на основе нормализованного разреза.
- Проведено сравнение результатов работы алгоритмов путем вычисления индексов согласованности.
- Результаты представлены в виде таблиц.
- Выполнен анализ тематики журналов, отнесенных к одному кластеру, выявлены разделы экономики, связанные общими научными интересами.

Индексы согласованности разбиений на основе алгоритмов S_1 и S_2

	k = 2	k =9	k = 111
NMI	0,083465	0,307437	0,486886
RAND	0,355641	0,703563	0,817692
ARI	– 0,006179	0,004772	0,007779

Резюме

Выявлена зависимость результатов работы алгоритмов от факторов:

- ориентации ребер (ориентированный, неориентированный);
- силы связи вершин (взвешенный, невзвешенный);
- способа приведения орграфа к неориентированному виду ($A+A^T$, $A\,A^T$, $A^T\,A$);
- метода сравнения результатов (индексы согласованности).

Литература

1. Shi J., Malik J. Normalized cut and image segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000. V. 22, iss. 8. P. 888-905.

2. Meila M., Shi J. A random walks view of spectral segmentation // Proc. International Workshop on AI and Statistics (AISTAT) 2001.

3. Meila M., Pentney W. Clustering by weighted cuts in directed graphs // Proc. of the 2007 SIAM International Conference on Data Mining in directed cuts. 2007. P. 135-144. Apr. 26-28, 2007. Minneapolis, Minnesota.

4. Бредихин С.В., Ляпунов В.М., Щербакова Н.Г. Спектральный анализ сети цитирования научных журналов // Проблемы информатики. 2018. № 2. С. 27-40.