Phân Tích Thuật Toán Louvain

Duy Lâm

December 2024

1 Giới Thiệu Thuật Toán Louvain

Thuật toán Louvain là một phương pháp phổ biến và hiệu quả để phát hiện cộng đồng trong đồ thị. Thuật toán hoạt động dựa trên việc tối ưu hóa chỉ số **Modularity**, nhằm đo lường chất lượng của các cộng đồng trong đồ thị. Quá trình thực hiện bao gồm hai pha chính:

- **Pha 1: Gán lại các nút vào cộng đồng.** Mỗi nút được gán vào cộng đồng mà việc di chuyển sẽ làm tăng chỉ số Modularity lớn nhất.
- **Pha 2: Gom các cộng đồng thành siêu nút.** Các cộng đồng được gom lại thành một siêu nút và thuật toán lặp lại trên đồ thị mới.

Thuật toán dừng khi chỉ số Modularity không thể cải thiện thêm, đảm bảo tính hiệu quả và tốc độ trên đồ thị lớn.

2 Tóm Tắt Các Chỉ Số Đánh Giá Việc Phát Hiện Cộng Đồng

Dưới đây là các chỉ số phổ biến dùng để đánh giá chất lượng phát hiện cộng đồng:

- **Modularity (Q):**

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{1}$$

  Trong đó:

  - A_{ij}: Ma trận trọng số của đồ thị.

- k_i, k_j: Tổng trọng số của các cạnh liên quan đến các nút i và j.
- m: Tổng trọng số tất cả các cạnh.
- $\delta(c_i, c_j)$: Hàm Kronecker (1 nếu $c_i = c_j$, ngược lại 0).

Giá trị $Q > 0.3$ được coi là tốt, và $Q > 0.5$ là rất tốt.

- **Conductance (Φ):**

$$\Phi(S) = \frac{\text{cut}(S, \bar{S})}{\min(\text{vol}(S), \text{vol}(\bar{S}))} \tag{2}$$

Trong đó:

- $\text{cut}(S, \bar{S})$: Tổng trọng số các cạnh giữa S và phần bù $\bar{S}$.
- $\text{vol}(S)$: Tổng trọng số của tất cả các cạnh liên quan đến các nút trong S.

Giá trị nhỏ hơn của Φ phản ánh phân cụm tốt hơn.

- **Normalized Cut (NC):**

$$NC(S) = \frac{\text{cut}(S, \bar{S})}{\text{vol}(S)} + \frac{\text{cut}(S, \bar{S})}{\text{vol}(\bar{S})} \tag{3}$$

Giá trị nhỏ hơn của NC cho thấy biên giữa các cộng đồng rõ ràng hơn.

3 Phân Tích Biểu Đồ Phát Hiện Cộng Đồng

Dựa trên biểu đồ đã cung cấp, chúng ta thực hiện so sánh giữa ba thuật toán phát hiện cộng đồng: Girvan-Newman, Label Propagation và Louvain.

3.1 Số Lượng Cộng Đồng

- **Girvan-Newman:** Phát hiện được 5 cộng đồng, phản ánh sự phân chia chi tiết nhưng có thể kém hiệu quả.
- **Label Propagation:** Phát hiện được 3 cộng đồng, cho thấy tốc độ nhanh nhưng chất lượng trung bình.
- **Louvain:** Phát hiện được 4 cộng đồng, cân bằng giữa chất lượng và số lượng.

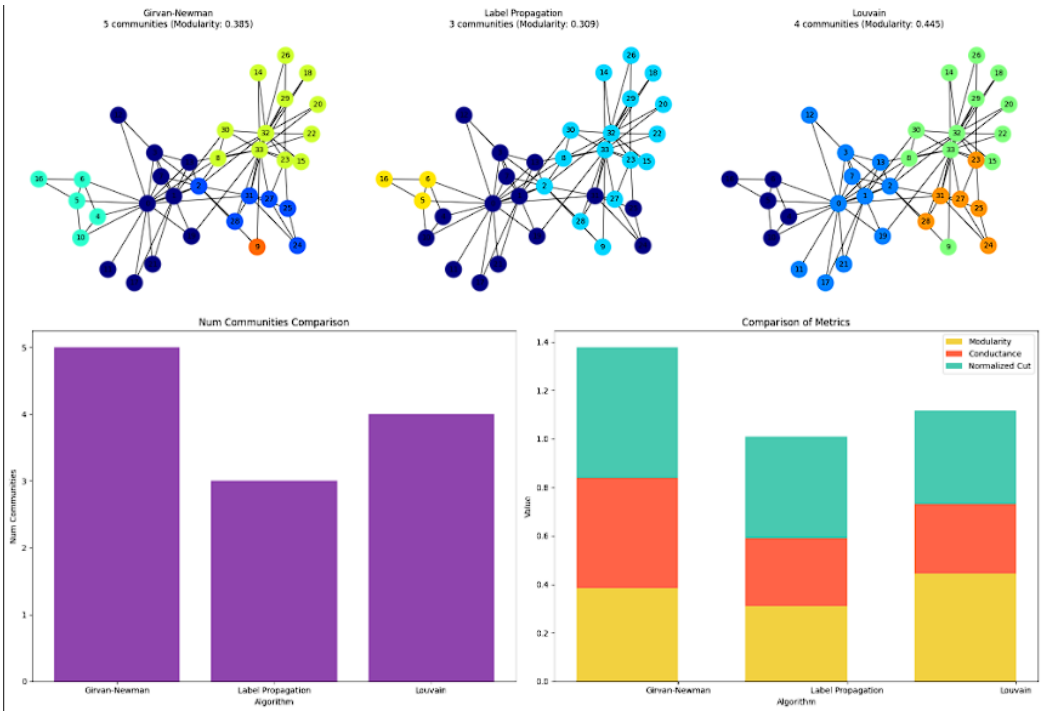
3.2 Chỉ Số Modularity

$$Q_{\text{Girvan-Newman}} = 0.385, \quad Q_{\text{Label Propagation}} = 0.309, \quad Q_{\text{Louvain}} = 0.445 \quad (4)$$

Louvain có chỉ số Modularity cao nhất, chứng minh khả năng phân cụm rõ ràng hơn.

3.3 Biểu Đồ Phân Tích Kết Quả

Dưới đây là biểu đồ minh họa kết quả phát hiện cộng đồng:



Hình 1: Biểu đồ so sánh ba thuật toán phát hiện cộng đồng.

3.4 Nhận Xét

- **Girvan-Newman:** Chậm và phù hợp với đồ thị nhỏ, nhưng tạo ra nhiều cộng đồng chi tiết.
- **Label Propagation:** Nhanh nhưng ít chính xác.
- **Louvain:** Hiệu quả nhất với chỉ số Modularity cao, phù hợp cho đồ thị lớn.