**CH7 社區檢測和相似性度量**

**【Definition】**一種分析網絡結構的方法，目的是識別出網絡中具有緊密連接的節點子集（也稱為社區或社群）。社區檢測有助於揭示網絡中的模式、群組和相關子結構，並提供對網絡結構和組織的理解。

**【應用】**CH3的推薦系統（有分為產品聚類和用戶聚類）

**【社區檢測算法】**

1. 分層聚類（Hierarchical Clustering）

<eg.> GN算法（Girvan – Newman Algorithem）：基於圖論中的邊介度概念，通過移除網絡中的高介度邊來切斷社區之間的連接，直到網絡分成多個社區為止。

1. Connected Components Algorithm（連接組件）：用於檢測斷開的子圖
2. Label Propagation算法（標籤傳播）：該算法基於節點標籤的傳播，每個節點根據其鄰居節點的標籤進行更新，最終將節點分配到相應的社區。
3. Louvain算法：是一種基於最大化模塊度的迭代算法，它通過不斷重新分配節點的社區隸屬度，使得模塊度不斷增加，從而得到最優的社區分配。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名稱** | **GN算法** | **連接組件算法** | **標籤傳播算法** | **Louvain算法** |
| 適用領域 | 社區發現、圖論分析等 | 圖論分析、網絡結構分析等 | 社區發現、圖論分析等 | 社區發現、圖論分析等 |
| 計算方式 | 通過不斷移除圖中的邊，計算模塊度的變化 | 通過深度優先搜索或廣度優先搜索，尋找連通的子圖 | 根據節點的鄰居標籤，進行標籤的傳播和更新 | 通過最大化模塊度，合併和分割節點群組 |
| 範圍 | 對整個圖進行社區劃分 | 尋找連通的子圖，不具備社區劃分的能力 | 對整個圖進行節點的標籤劃分 | 對整個圖進行社區劃分 |
| 特性 | 可以得到不同級別的社區結構 | 能夠找到圖中所有連通的子圖 | 簡單且快速，但不保證得到最佳結果 | 高效且快速，可用於大型圖的社區發現 |
| 解釋 | 通過計算模塊度的變化，識別和移除橋接邊 | 通過搜索連通的子圖，識別圖中的連接組件 | 根據節點的鄰居標籤進行標籤傳播和更新 | 通過最大化模塊度，優化社區結構 |
| 使用案例 | 社交網絡分析、網絡社區發現等 | 網絡分析、連接組件識別等 | 社交網絡分析、網絡社區發現等 | 社交網絡分析、網絡社區發現等 |

* **圖組件（Graph Components）**

【Definition】圖中的相互連通的子圖。在一個圖（組件）中，兩個節點通過路徑相互連接，且不與組件外的任何其他節點連接。

【種類】

1. 強連接（Strongly Connected Components）

在有向圖中，每個節點都可以通過有向邊的路徑相互到達的最大子圖。強連接組件可以看作是圖中的獨立子圖，其中的節點彼此之間具有雙向的連接關係。

1. 弱連接（Weakly Connected Components）

在有向圖中，將所有的有向邊都視為無向邊後，得到的圖組件。弱連接組件將有向圖轉換為無向圖，並找出其中的連通子圖。

與強連接組件不同，弱連接組件不考慮有向邊的方向，而只關注節點之間是否存在路徑來連接它們。即使在有向圖中某些節點之間只有單向的有向邊，它們仍然可以屬於同一個弱連接組件，因為在無向圖中它們可以通過雙向的無向邊相互連接。

* **測量節點相似性**

【Definition】評估圖中節點之間的相似程度或相關性

1. 基於**集合**的相似性（值越接近1表示兩個節點之間的相似性越高，越接近0則表示相似性越低。）
   1. 重疊相似性（Overlapping Similarity）
   2. Jaccard相似性
2. 基於**向量**的相似性
   1. 歐幾里得距離
   2. 餘弦相似性（Cosine Similarity）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名稱** | **重疊相似性（Overlapping Similarity）** | **Jaccard相似性** | **歐幾里得距離** | **餘弦相似性（Cosine Similarity）** |
| 定義 | 衡量兩個集合之間的相似性 | 衡量兩個集合之間的相似性 | 衡量兩個向量之間的距離 | 衡量兩個向量之間的相似性 |
| 適用領域 | 集合論、資料科學、資訊檢索等 | 資料科學、資訊檢索等 | 數學、統計、機器學習、資料科學等 | 自然語言處理、資料科學等 |
| 計算方式 | 重疊元素數量除以兩個集合的總元素數量 | 交集元素數量除以兩個集合的聯集元素數量 | 各維度差的平方和開根號 | 兩個向量的內積除以兩個向量的歐幾里得長度的乘積 |
| 範圍 | 0到1之間的數值 | 0到1之間的數值 | 0到正無窮的數值 | -1到1之間的數值 |
| 特性 | 受集合大小的影響 | 受集合大小的影響 | 受向量大小的影響 | 不受向量大小的影響 |
| 解釋 | 衡量兩個集合中相同元素的比例 | 衡量兩個集合中共享元素的比例 | 衡量兩個向量之間的距離 | 衡量兩個向量的方向相似度 |
| 使用案例 | 用於文本分析中的詞語匹配、集合比較等 | 用於文本分析中的詞語匹配、集合比較等 | 用於維度相似度比較、分群等 | 用於文件相似度比較、推薦系統等 |