[빅데이터 HW1]

Overlapping Community Detection Using Neighborhood-Inflated Seed Expansion

2014314856 오예린

이 논문은 disjoint(하나의 노드는 하나의 클러스터에만 들어가야 함)하고 exhaustive(노드들은 반드시 클러스터에 들어가야 한다는 것) 했던 기존의 community detection 방법을 Neighborhood-Inflated Seed expansion을 통해 개선한 새로운 Overlapping Community (외부보다 내부에서 더 응집력 있는 connection을 갖는 노드의 집합) Detection을 제시한다. 이 알고리즘은 실제 네트워크에서 노드가 여러 community에 속하는 자연스러운 현상을 가능하게 만들며, $C_1 \cup .. \cup C_K \subseteq V$ 인 클러스터를 찾는 것을 목표로 함으로써 어느 클러스터에도 속하지 않는 outlier 또한 존재하도록 한다.

논문에서 소개하는 알고리즘은 여타의 최신 overlapping community detection 알고리즘보다 실행 시간, 커뮤니티의 응집력, 정확도 측면에서 뛰어난 성능을 보인다. 이 알고리즘은 seed-and-grow 방식을 취하는데 이는, 커뮤니티의 구심점인 노드를 잘 찾아서 확장하는 것을 말한다. 하지만 좋은 seed를 고르는 문제는 일일이 seed를 탐색해야하는 greedy한 방법으로 계산 량이 많기에, seed를 고르는 데 있어 "Graclus centers"와 “Spread hubs" 전략을 취한다.

"Gracluse centers" seeding: kernel k-means와 그래프 클러스터링의 objectives가 동일함을 이용한다. distance kernel을 이용하여 그래프의 응집력 있는 노드 집합들 사이에 good seed를 위치시킬 수 있다. (클러스터의 중심 노드를 seed로, 중심 노드의 이웃 노드들을 seed region으로 삼는다)

"Spread hubs": 높은 degree를 갖는 노드들의 independent한 set을 뽑는다.

seed set을 확장할 때에는 personalized PageRank clustering을 이용한다. 기존 PageRank에서의 random walk이 1-s의 확률로 아무 노드로나 뛸 수 있게 했다면, personalized PageRank clustering은 1-s의 확률로 이미 정해놓은 노드 set으로 뛸 수 있게 한다. 이 논문에서는 이미 정해놓은 노드 set이 seed 노드가 되며, 또한 seed region(seed의 이웃 노드들)을 PPR의 인풋으로 함께 넘겨 성능을 향상시켰다.

정리하면, NISE는 filtering, seeding, seed expansion, propagation의 4단계로 이루어져 있다.

Filtering Phase: 노이즈나, overlapping community에 속하지 않는 노드들을 제거하는 과

정이다. single edge를 가지는 biconnected component를 제거해서 가장 큰 connected component를 찾는다. filtering의 결과 biconnected core graph(단절점을 갖지 않는 connected graph)가 나온다. 단절점은 그 노드와, 연결된 모든 edge를 제거했을 때, 그래프가 2개 이상의 connected component로 나뉘는 것을 말한다.

Seeding Phase: Graclus Centers와 Spread Hubs의 방법으로 seed를 찾는 과정이다.
1. Graclus Centers:
① 일단, disjoint한 clustering을 하는 graph partitioning 기법을 적용해 작은 conductance를 갖는 sets를 구한다.
② 각 set에서 가장 중심에 있는 노드를 골라 seed set을 만든다.
③ 클러스터 안의 노드와 그 중심 사이의 거리를 kernel을 이용해 구한다.

2. Spread Hubs:
① 초반에 모든 노드들은 unmarked 되어있다.(k개의 seed가 골라질 때까지)
② unmarked된 노드들 중에서 가장 degree가 높은 것이 seed로 선택된다.(선택된 vertex와 그 neighbors까지 mark한다.)
③ degree가 같은 vertex들이 있을 경우에는 근처에 있는 seed들이 뽑히지 않도록 independent set을 이용한다.

Seed Expansion Phase: personalized PageRank vector를 이용하여 seed를 중심으로 cluster를 확장한다. 1-α의 확률로 seed 노드로 뛰게 하여 seed 가까이에 있는 노드들이 방문되도록 한다. 먼저 PPR vector를 계산하고, PPR score가 높은 순에서 낮은 순으로 노드를 확인하여 가장 작은 conductance set이 생성되도록 한다. 이렇게 낮은 conductance community를 만들기 위해선 seed 노드뿐 아니라, 그 이웃 노드까지도 restart node로 삼는 neighborhood inflation이 중요하다.

Propagation Phase: 알고리즘의 가장 마지막 단계이다. filtering phase를 거쳐 biconnected core를 만든 후, 이 biconnected core graph의 seed를 찾고, 확장했다면 이후에 filtering 단계에서 떼었던 영역까지 community를 확장한다. filtering 단계에서 떼어졌던 whisker가 bridge를 통해 연결된다.

Experimental Results: NISE와 다른 최신 overlapping community detection methods (Bigclam, Demon. Oslom)를 10개의 real-world network에 적용해 비교했다.
그 결과, NISE는 다른 방법들에 비해 더 큰 사이즈의 클러스터를 찾아냈으며(10배에서 100배까지) 클러스터의 사이즈에 대한 분산도 크게 나타났다. 또한 10개의 영역 모두에서 잘 돌아갔으며, 속도 또한 빨랐다. 즉, NISE는 다른 최신 overlapping community detection methods보다 실행 시간, 커뮤니티의 응집력, 정확도의 측면에서 월등함을 보였으며, 이는 논문에서 제시한 새로운 seeding strategies인 "graclus centers"와 “spread hubs"가 기존의 방법들보다 우월해, seed set의 확장을 성공시키는 중요한 역할을 했기 때문이다.