알고리즘에 따른 그래프 분할 기법 분석

Analysis of Graph Partitioning Methods Based on Algorithms

송 상 호[1], 이 현 병[1], 최 도 진[2], 임 종 태[1],
복 경 수[2], 유 재 수[1)]
충북대학교[1], 창원대학교[2], 원광대학교[3]

Sangho Song[1], Hyeon-Byeong Lee[1],
Dojin Choi[2], Jongtae Lim[1], Kyoungsoo Bok[3],
Jaesoo Yoo[1]
Chungbuk National Univ[1],
Changwon National Univ[2], Wonkwang Univ[3]

요약

그래프는 추천 시스템, 기계 학습, 소셜 네트워크 분석 등과 같은 다양한 응용 분야에서 엔터티의 관계를 모델링하는 데이터 표현이다. 최근에는 정점과 간선이 많은 그래프가 나타나고 있다. 실제 그래프의 양은 많고 불규칙한 구조로 확장성 문제가 발생한다. 따라서 다양한 그래프 분할 기술을 통해 대규모 그래프의 확장성 문제를 개선하고 처리 시간 복잡성을 줄일 수 있다. 논문에서는 그래프 분할의 다양한 기법과 접근 방법에 대하여 분석한다. 기존 분할 기법과 달리 GNN을 위한 그래프 분할 기법은 L-홉 이웃에 자주 접근할 수 있는 분할 기법이 필요하다.

I. 서론

그래프는 정점과 간선의 형태로 엔터티 관계를 모델링하는 데이터 표현이다. 일반적으로 정점은 그래프의 개체를 나타내고, 간선은 그래프의 개체 간의 관계를 나타낸다. 그래프는 검색 엔진에서 웹 페이지의 관련성을 표현하거나 생물학에서 단백질 간의 상호 작용, 소셜 네트워크에서 사용자와 그룹의 사용 작용을 그래프로 표시할 수 있다. Facebook, Google 등과 같은 소셜 네트워크 서비스를 운영하는 기업에서는 거대해지는 그래프를 분석하기 위해 다양한 솔루션을 제안했다. 그래프의 크기가 커짐에 따라 분산 환경에서 그래프 분석을 수행할 수 있다. 그러나 그래프 컴퓨팅은 액세스 불규칙성, 지역성 문제, 서로 다른 클러스터에 있는 그래프의 불균형 분포로 인한 문제가 있다.

그래프 분할은 분할을 최소화하고 로드 밸런스를 최대화하여 그래프를 별개의 서브 그래프로 자르는 기법이다. 그래프 분할은 대규모 그래프의 중요한 전처리 단계에서 사양된다. 그래프 분할은 데이터 마이닝, 기계 학습 및 패턴 분석에서 사용된다. 이와 관련된 연구가 지난 10년 동안 진행되었다. 그래프 분할 기법은 정점 분할, 간선 분할, 하이브리드 분할 세 가지로 분류할 수 있다. 본 논문에서는 그래프 분할 기법들을 분석한다. 이를 통해 기존 그래프 분할 기법과 GNN을 위한 그래프 분할 기법의 차이를 파악한다.

1) 교신저자 : yjs@cbnu.ac.kr
이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단(No. 2022R1A2B5B02002456, 기여율 34%). 과학기술정보통신부 및 정보통신기획평가원의 지역지능화혁신인재양성(Grand ICT연구센터)(IITP-2024-2020-0-01462, 기여율 33%), 농촌진흥청 연구사업 (세부과제번호: RS-2021-RD010195, 기여율 33%)의 지원에 의해 이루어졌음

II. 그래프 분할 기법 분석

그래프 G는 G = (V, E) 로 정의된다. 여기서 V={v1, v2, v3 ... , vn}과 E = {e1, e2, ..., em}은 각각 정점과 간선의 집합이다. 그래프는 가중치 그래프와 비가중치 그래프로 분류할 수 있다. 가중치 그래프는 간선에 가중치를 가져올 수 있고, 비가중치 그래프는 각 간선의 가중치가 1인 가중치 그래프로 해석할 수 있다. 그래프 분할은 정점 분할, 간선 분할, 하이브리드 분할로 분류할 수 있다. 분할 기법의 성능 평가 지표로 지역성, 처리 시간, 확장성, 통신 비용 등을 사용한다.

1. 정점 분할

정점 분할(Vertex Partitioning)은 그림 1a과 같이 edge-cut이라고도 한다. 이는 부하 분산과 관련된 edge cut을 최소화하면서 서로 다른 파티션에 정점을 할당하여 그래프를 서브 그래프로 나눈다. 정점 분할의 목표는 서로 다른 집합에 속하는 간선 가중치의 합이 최소화되는 k개의 분리된 집합으로 정점을 나누는 것이다.

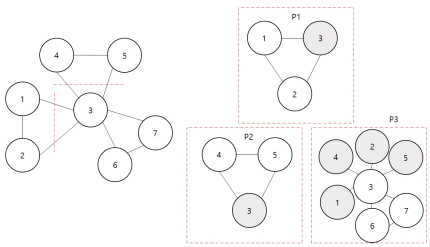
2. 간선 분할

간선 분할(edge Partitioning)은 그림 1b와 같이 vertex-cut이라고도 한다. 최대 로드 밸런스와 최소 정점

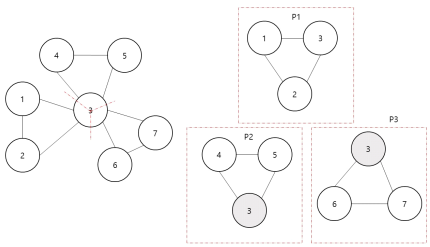
분할을 고려하면서 서로 다른 파티션 세트에 간선을 할당하여 그래프를 여러 개의 서브 그래프로 나눈다. 간선 분할을 사용하면 분할된 정점을 보유하는 노드는 정점의 복제본을 보존해야 한다. 이러한 복제된 정점은 파티션 간의 연산을 연결하는 역할을 할 수 있다. 복제 계수와 간선의 수는 통신 비용과 부하 비용에 따라 결정된다.

3. 하이브리드 분할

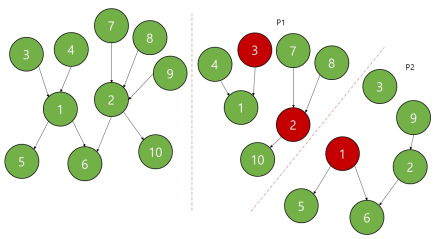
간선 분할은 노드에 균등하게 간선을 할당하고 정점만 복사하여 각 파티션 내에서 서브 그래프를 구성한다. 따라서 간선 분할에서는 주로 전체 복제된 정점을 최소화하는데 중점을 둔다. 그러나 하이브리드 분할은 모든 정점의 복제를 줄이는 대신 정점의 낮은 차수와 높은 차수를 구분하는 것을 목적으로 한다. 더 나은 정점 분할 혹은 간선 분할을 사용한다. 하이브리드 분할은 정점 분할과 간선 분할을 혼합하여 사용한다. 실제 그래프의 대부분은 낮은 차수를 갖고 적은 수의 정점이 높은 차수를 갖는다. 부하 분산을 위해 높은 차수의 정점을 파티션에 복제하여 파티션 간의 통신을 줄인다. 그림 1c에서 볼 수 있듯이 차수 임계값은 3으로 하여 분할할 경우 정점 1, 2는 높은 차수이고 다른 모든 정점은 낮은 차수를 갖는다.



(a) 정점 분할



(b) 간선 분할

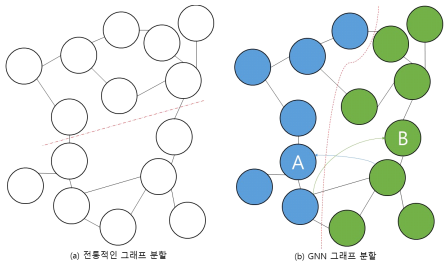


(c) 하이브리드 분할

▶▶ 그림 1. 전통적인 그래프 분할 기법

4. GNN을 위한 파티셔닝

기존 그래프 알고리즘에 비해 GNN은 정점의 L-홉 이웃 정점에 자주 접근해야 하며, GNN 정점은 고차원 특징 표현이 있다. 결과적으로 통신 부하 불균형 문제는 GNN에서 훨씬 더 두드러진다. 따라서 GNN에 대한 그래프 분할은 통신 부하의 균형을 고려하여 레이블이 지정된 정점의 이웃을 서로 다른 파티션에 고르게 분산해야 한다. 그림 2와 같이 GNN을 위한 그래프 파티셔닝은 전통적인 기법과 다르게 정점이 L-홉 이웃에 접근하도록 분할한다.



▶▶ 그림 2. GNN을 위한 그래프 분할 기법

Ⅳ. 결론

본 논문에서는 전통적인 그래프 분할 기법과 GNN을 위한 그래프 분할 기법에 대해 분석하였다. 그래프 데이터의 급격한 증가에 대처하기 위해 효율적인 그래프 분할 기법이 필요하다. 또한 그래프 알고리즘에 따라 효율적인 그래프 분할 기법이 존재한다. 전통적인 그래프 분할 기법은 정점 분할, 간선 분할, 하이브리드 분할 기법이 존재한다. GNN은 정점의 L-홉 이웃에 자주 접근해야 하므로 전통적인 그래프 분할 기법과 다른 분할 기법이 필요하다. 향후에는 분할 기법에 따른 지역성, 처리 시간, 확장성 등의 실험 평가를 진행할 예정이다.

▌참 고 문 헌▐

[1] Joseph E. Gonzalez, Yucheng Low, Haijie Gu, Danny Bickson, and Carlos Guestrin. 2012. PowerGraph: distributed graph-parallel computation on natural graphs. In Proceedings of the 10th USENIX conference on Operating Systems Design and Implementation (OSDI'12). USENIX Association, USA, 17-30.

[2] Q. Wang, X. Ai, Y. Zhang, J. Chen and G. Yu, "HyTGraph: GPU-Accelerated Graph Processing with Hybrid Transfer Management," in 2023 IEEE 39th International Conference on Data Engineering (ICDE), Anaheim, CA, USA, 2023 pp. 558-571.

[3] Yuan, H., Liu, Y., Zhang, Y., Ai, X., Wang, Q., Chen, C., Yu, G. Comprehensive Evaluation of GNN Training Systems: A Data Management Perspective. arXiv preprint arXiv:2311.13279, 2023