대용량 네트워크 압축 기반 클러스터링 알고리즘 개발

Development of Clustering Algorithm based on Massive Network Compression

서 동 민*,**, 유 석 종*, 이 민 호*,**
한국과학기술정보연구원*,
과학기술연합대학원대학교**

Dongmin Seo*,**, Seok Jong Yu*, Min-Ho Lee*,**
Korea Institute of Science and Technology Information*, Univ. of Science & Technology**

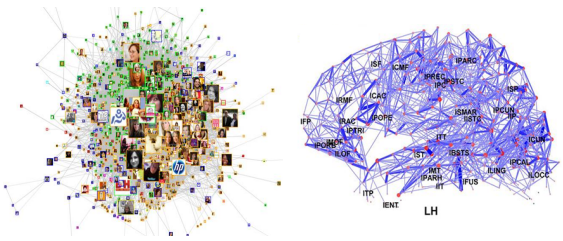
요약

빅데이터란 대용량 데이터 활용 및 분석을 통해 가치 있는 정보를 추출하고, 이를 바탕으로 대응 방안 도출 또는 변화를 예측하는 기술을 의미한다. 그리고 빅데이터 분석에 활용되는 데이터인 페이스북과 같은 소셜 데이터, 유전자 발현과 같은 바이오 데이터, 항공망과 같은 지리정보 데이터들은 대용량 네트워크로 구성되어 있다. 네트워크 클러스터링은 서로 유사한 특성을 갖는 네트워크 내의 데이터들을 동일한 클러스터로 묶는 기법으로 네트워크 데이터를 분석하고 그 특성을 파악하는데 폭넓게 사용된다. 최근 빅데이터가 다양한 분야에서 활용되면서 방대한 양의 네트워크 데이터가 생성되고 있고, 이에 따라서 대용량 네트워크 데이터를 효율적으로 처리하는 클러스터링 기법의 중요성이 증가하고 있다. MCL(Markov Clustering) 알고리즘은 플로우 기반 무감독(unsupervised) 클러스터링 알고리즘으로 확장성이 우수해 다양한 분야에서 활용되고 있다. 하지만, MCL은 대용량 네트워크에 대해서는 많은 클러스터링 연산을 요구하며 너무 많은 클러스터를 생성하는 문제를 갖는다. 본 논문에서는 네트워크 압축을 기반으로 한 클러스터링 알고리즘을 제안함으로써 MCL보다 클러스터링 속도와 정확도를 향상시켰다. 또한, 희소행렬을 효율적으로 저장하는 CSC(Compressed Sparse Column) 자료구조와 MapReduce 기법을 제안한 클러스터링 알고리즘에 적용함으로써 대용량 네트워크에 대한 클러스터링 속도를 향상시켰다.

Ⅰ. 서론

최근 빅데이터(Big-data)가 IT 뉴스의 핵심 키워드로 부상했다[1]. 빅데이터란 대용량 데이터 활용·분석을 통해 가치 있는 정보를 추출하고, 이를 바탕으로 대응 방안 도출 또는 변화를 예측하기 위한 정보화 기술을 말한다[1]. 특히, 빅데이터는 지속적으로 변화하면서 산업별, 시장별 구분에 따라 다르게 적용될 수 있기 때문에 방대한 볼륨과 빠른 속도로 축적되고 있는 다양한 형태의 데이터에 대한 고급 분석, 즉 빅데이터 분석을 통해 다양한 가치를 창출할 수 있게 되었다[1].

그림 1의 (a)는 페이스북에서 인맥 구조를, 그림 1의 (b)는 뇌를 구성하는 중요 화합물, 유전자/단백질 구조를 가시화 한 것인데, 모두 네트워크 구조를 보이고 있다. 네트워크는 하나 이상의 데이터를 데이터 간 관계를 기반으로 연결시킨 자료구조를 의미하고, 네트워크 구조를 분석하면 밝혀지지 않은 데이터 간 특성 및 패턴을 발견할 수 있기 때문에 대부분의 빅데이터 분석은 네트워크 분석을 기반으로 한다. 특히, 네트워크에 대한 클러스터링(clustering)은 위상학적(topological) 또는 의미적(semantic)으로 유사한 특성을 갖는 네트워크 내의 데이터들을 동일한 클러스터로 묶는 기법으로 네트워크 데이터를 분석하고 그 특성과 패턴을 분석하는데 폭넓게 사용된다. 최근 빅데이터가 다양한 분야에서 활용되면서 방대한 양의 네트워크 데이터가 생성되고 있고, 이에 따라서 대용량 네트워크 데이터를 효율적으로 처리하는 클러스터링 기법의 중요성이 증가하고 있다.



(a) Facebook network (b) Brain network

▶▶ 그림 1. 대용량 네트워크 데이터 예

(출처: http://www.wired.com/2012/04/facebook-disease-friends/, https://commons.wikimedia.org/wiki/File:Network_representation_ of_brain_connectivity.JPG)

본 논문에서는 대용량 네트워크 분석 지원을 위한 네트워크 압축을 기반으로 한 클러스터링 알고리즘을 개발했다. 또한, 희소행렬을 효율적으로 저장하는 CSC (Compressed Sparse Column) 자료구조와 MapReduce 기법을 개발한 클러스터링 알고리즘에 적용함으로써 대

* 본 연구는 한국과학기술정보연구원의「초고성능컴퓨팅기반 건강한 고령사회 대응 빅데이터 분석기술개발 (K-16-L03-C02-S02)」사업으로부터 지원받아 수행된 연구임.

용량 네트워크에 대한 클러스터링 속도와 정확도를 향상시켰다.

Ⅱ. 관련연구

그림 2는 Stijn van Dongen이 개발한 MCL(Markov Clustering algorithm)[2]으로 MCL은 플로우 기반 무감독 클러스터링 알고리즘으로 확장성이 우수해 다양한 분야에서 활용되고 있다. 하지만, 그림 2에서와 같이 Expand 연산 수행시 원본 네트워크 M을 기반으로 하기 때문에 작은 사이즈의 클러스터가 많이 만들어지는 문제가 발생한다. 또한, [3]에서 지적된 것과 같이 대용량 네트워크에 대해서는 많은 클러스터링 연산을 요구하는 문제를 갖는다.

Algorithm 1 MCL
A := A+I //Add self-loofs to the graph
M := AD^{-1} //Initialize M as the canonical transition matrix
repeat
  M := Mexp := Expand(M)
  M := Minf := Inflate(M,r)
  M := Prune(M)
until *M* converges
Interpret *M* as a clustering

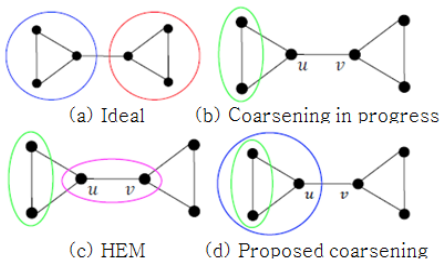
▶▶ 그림 2. MCL 알고리즘 (출처: [3])

Ⅲ. 제안하는 클러스터링 알고리즘

그림 3은 제안하는 클러스터링 알고리즘을 보여준다. 그림 3에서 Coarsening(N0)는 주어진 네트워크를 압축하는 함수로, 그림 4에서와 같이 기존 대표 네트워크 압축 기법 중 하나인 HEM(Heavy Edge Matching)[4]에 대해 기존 압축에 참여한 노드에 대한 재압축을 허용함으로써 네트워크 압축률 및 정확도를 개선하였다. 그리고 제안하는 클러스터링 알고리즘에서는 희소행렬(sparse matrix)에 대한 저장 공간 활용을 개선하기 위해 [5]에서 소개된 CSC(Compressed Sparse Column) 자료구조를 통해 행렬을 관리했다. 또한, 그림 3에서 Inflate와 Prune 연산 방법은 MCL과 동일하나 병렬 수행될 수 있는 행렬 연산이기에 MapReduce를 통해 병렬 수행함으로써 클러스터링 연산 속도를 개선하였다.

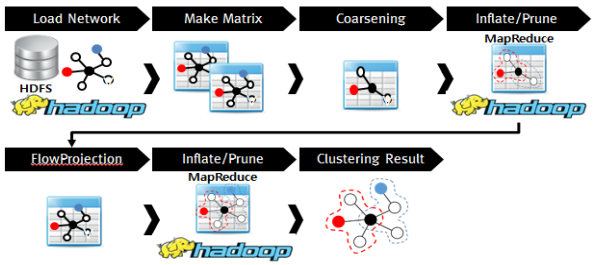
Algorithm 2 Proposed clustering algorithm
Input : Network *N*, regularization factor *r*
N_0 := *N*
M_0 := N_0 // Initialize matrix
M_C := *Coarsening*(N_0) // Matrix of a coarsened network
M_f := M_C // Coarsened matrix
foreach *i = to 3* **do**
  M_f := *Prune*(*Inflate*(M_f×M_C),*r*) // In MapReduce
end
  M_f := *FlowProjection*(M_f)
repeat
  M := Prune(*Inflate*(M_f×M_0),*r*) // In MapReduce
until *M* converges
Interpret *M* as a clustering

▶▶ 그림 3. 제안하는 클러스터링 알고리즘



▶▶ 그림 4. 제안하는 네트워크 압축 기법

그림 5는 제안하는 MapReduce 기반 클러스터링 진행과정을 보여준다.



▶▶ 그림 5. 제안하는 MapReduce 기반 클러스터링 진행과정

Ⅳ. 결론 및 향후연구

본 논문에서는 HEM을 개선시킨 네트워크 압축 기법을 제안하고 클러스터링 과정에 수행되는 일부 연산을 병렬처리 함으로써, 기존 클러스터링 알고리즘 대비 정확도와 속도를 향상시킨 클러스터링을 개발하였다. 향후에는 다양한 데이터 셋을 가지고 제안하는 알고리즘에 대한 성능평가를 수행할 것이며, 최근 Hadoop/MapReduce를 기반으로 한 네트워크 분석 기법들이 연구되고 있기에 본 논문에서 개발한 알고리즘과의 성능평가를 통해 우수성을 입증할 계획이다.

■ 참 고 문 헌 ■

[1] 서동민, 최윤수, 전선희, 이민호, "바이오 패스웨이 다차원 분석 시스템 개발", 한국콘텐츠학회논문지, 제14권, 제11호, pp.467-475, 2015.
[2] S. V. Dongen, Graph Clustering by Flow Simulation, PhD thesis, University of Utrecht, 2000.
[3] V. Satuluri and S. Parthasarathy, "Scalable graph clustering using stochastic flows: applications to community discovery", Pro. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp.737-745, 2009.
[4] G. Karypis and V. Kumar, "A fast and high quality multilevel scheme for partitioning irregular graphs", SIAM Journal on Scientific Computing, pp.359-392, 1998.
[5] I. DUFF, R. GRIMES, and J. LEWIS, "Sparse matrix test problems", ACM Trans. Math. Soft., pp.1-14, 1989.