그래프 클러스터링 연구 동향

김정림*, 박상현*,+
*연세대학교 컴퓨터과학과, +교신저자
e-mail : kimgogo02@yonsei.ac.kr

Research of graph clustering

JungRim Kim*, Sanghyun Park*,+
*Dept of Computer Science, Yonsei University, +Correponding Author

1. 연구 필요성 및 문제점

방대한 양의 그래프 데이터의 등장으로 이를 효과적으로 분석하고 활용하는 것은 중요한 연구주제가 되었다. 클러스터링은 라벨링(labeling)이 되어 있지 않은 데이터 내에서 유사한 특성을 가지고 있는 그룹을 찾기 위한 방법론으로, 그래프 데이터 분석을 위한 기법으로 활용 될 수 있다. 본 논문에서는 그래프 클러스터링 관련 연구를 소개하고, 앞으로 해결해야 할 문제를 소개한다.

2. 연구내용과 방법

초기에는 그래프의 간선 정보를 이용하여 클러스터를 찾는 연구가 많이 존재하였다. Van Dogon[1]은 그래프의 간선 정보를 통해 flow matrix를 만들고, Markov chain model을 이용하여 flow matrix를 시뮬레이션 하여 클러스터를 찾는 MCL 알고리즘을 제안하였다. Newman[2]은 그래프의 betweenness를 측정하고, 이를 이용하여 그래프를 분할하여 클러스터를 찾는 Girvan-Newman 알고리즘을 제안하였다. Xiaowei[3]는 서로 다른 두 노드가 공유하고 있는 이웃노드의 수를 기반으로 노드 간의 구조적 유사도를 측정하였고, 이를 이용하여 클러스터를 찾는 SCAN 알고리즘을 제안하였다.

이후, 소셜 네트워크등과 같이 노드의 속성 값을 포함하는 그래프 데이터가 등장하였고, 이러한 그래프를 보다 정확하게 분석하기 위하여 간선 정보뿐만 아니라 그래프의 노드 정보도 함께 고려하는 방법이 연구 되었다. Ruan[4]은 노드와 간선정보를 함께 사용하기 위해 metis와 markov 클러스터링 기법을 함께 사용하였다. Boobalan[5]은 간선정보를 이용하여 클러스터 탐색 및 병합을 수행하고, 노드의 속성 값을 이용하여 클러스터를 분해하는 과정을 반복하여 클러스터를 찾는 방법을 제안하였다.

최근에는, 방대한 양의 그래프 데이터가 축적됨에 따라서 서로 다른 그래프 데이터를 통합하는 연구들이 진행되고 있고, 그 결과 그래프 데이터의 크기가 점차적으로 커지고 있는 추세이다. 그렇기 때문에, 단일 머신에서 처리하지 못하는 대용량 그래프 분석을 위한 클러스터링 알고리즘의 개발은 중요한 문제가 되었다. Weizhong[6]은 SCAN 알고리즘을 Hadoop환경의 분산시스템에 적합한 형태로 변환하는 연구를 진행하여, 단일 머신에서 처리하지 못했던 대용량 그래프 데이터를 분석하였다.

3. 결론 및 향후 연구

본 논문에서는 그래프 클러스터링 연구 동향을 살펴보았으며, 최근에는 대용량 그래프 데이터 클러스터링에 대한 관심이 커지고 있는 것을 확인할 수 있었다. 추가적으로, 기존의 대용량 그래프 클러스터링 연구는 대부분 간선 정보만을 이용하여 클러스터링을 하였으며, 이미 존재하는 알고리즘을 분산 시스템 환경에서 구현하는 연구가 대부분이었다. 그렇기 때문에, 향후 연구는 분산 시스템 환경에서 간선 정보와 노드의 속성 값을 함께 고려할 수 있는 그래프 클러스터링 기법을 연구할 계획이다.

° 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2015R1A2A1A05001845).

참고문헌

[1] Van Dongen S, "A cluster algorithm for graphs" Amsterdam: CWI, 2000

[2] Girvan M. and Newman M. E. J., "Community structure in social and biological networks", Proc. Natl. Acad. Sci. Vol. 99, 7821 - 7826, 2002

[3] X.Xu, N.Yuruk, Z. Feng, et al. "SCAN: a structural clustering algorithm for networks", ACM SIGKDD 2007, pp. 824-833, 2007

[4] Yiye Ruan, David Fuhry, Srinivasan Parthasarathy, "Efficient community detection in large networks using content and links", WWW 2013, pp. 1089 - 1098, 2013

[5] M. Parimala Boobalana, Daphne Lopeza, X.Z. Gao, "Graph clustering using k-Neighbourhood Attribute Structural similarity", Applied Soft Computing, Vol. 47, pp.216-223, 2016.

[6] Weizhong Zhao, Venkataswamy Martha, Xiaowei Xu, "PSCAN: A Parallel Structural Clustering Algorithm for Big Networks in MapReduce", IEEE AINA 2013, pp. 862 - 869, 2013.