대용량 그래프 데이터를 효율적으로 시각화하는 방법에 대한 최신 연구 조사

곽우석**, 나인주****, 김현지*, 이경준*, 서인*, 한욱신***
*포항공과대학교 창의 IT 융합공학과
**포항공과대학교 컴퓨터공학과
***포항공과대학교 창의 IT 융합공학과/컴퓨터공학과
****한국과학기술원 전산학부
e-mail : uws8505@postech.ac.kr, ijna@dblab.postech.ac.kr, hjkim@dblab.postech.ac.kr, kjlee@dblab.postech.ac.kr, iseo@dblab.postech.ac.kr, wshan@postech.ac.kr

The State of the Art in Visualizing Large Graph Data

Useok Kwak**, In-Ju Na****, Hyeonji Kim*, Kyeong-Jun Lee*, In Seo*, Wook-Shin Han***
*Dept. of Creative IT Engineering, POSTECH
**Dept. of Computer Science and Engineering, POSTECH
***Dept. of Creative IT Engineering/Dept. of Computer Science and Engineering, POSTECH
****Dept. of Computer Engineering, KAIST

요 약

소셜 네트워크, 웹 시멘틱, 협력 네트워크 등과 같이 다양한 응용에서 대용량 그래프 데이터를 이용한다. 최근 이러한 데이터를 분석하기 위해 대용량 그래프 데이터를 효율적으로 시각화 하는 연구가 제안되었다. 이에 본 연구에서는 대용량 그래프 데이터를 효율적으로 시각화하는 방법에 대한 최신 연구 동향을 조사한다.

1. 서론

소셜 네트워크 서비스, 웹 시멘틱, 협력 네트워크, 생물학 등과 같이 많은 응용들이 그래프 데이터 모델을 사용한다[6]. 데이터 과학자들은 데이터의 시각화를 통해 데이터를 분석하고, 이를 통해 유용한 정보를 얻는다[6]. 그래프 데이터도 1960 년대부터 시각화에 대한 연구가 시작 되었으나[6], 최근 데이터의 크기가 급격히 증가함에 따라 대용량 그래프 데이터를 효율적으로 시각화 하기 위한 연구가 활발히 진행되고 있다[1-3,5-6,8]. 대용량 그래프 데이터는 전체 데이터를 시각화 할 경우 유의미한 정보를 추출하기 어려우며, 시각화 하는 과정에서 매우 많은 시간이 소요될 수 있다[3,7]. 또한, 기존 그래프 마이닝의 결과를 시각화하는 시스템에서 마이닝의 결과인 특정 노드 집합으로부터 이웃한 노드들을 탐색할 때, 탐색된 노드의 수가 급격히 커지는 문제가 발생한다[1,5]. 본 연구에서는 대용량 그래프 데이터를 효율적으로 시각화하는 최신 연구 동향에 대해 조사한다.

2. 상호적인 데이터 마이닝 및 시각화 기술

OPAvion[1]과 Perseus[5]는 그래프를 마이닝하고, 결과를 시각화하여 보여주는 응용이다. 두 응용은 페이지 랭크(PageRank), 연결 요소(connected components), 연결 정도 분포 (degree distribution), 이상 탐지(anomaly detection) 등의 마이닝 연산을 제공하며 결과를 시각화하여 보여준다. 시각화한 특정 노드로부터 이웃한 노드들을 탐색하는 기능을 제공하는데 특정 홉부터 노드의 수가 급격히 늘어나 시각화하기 어려울 수 있다. 이를 해결하기 위해 OPAvion[1]은 신뢰 전파 알고리즘(belief propagation), Perseus[5]는 페이지 랭크 알고리즘을 이용하여 사용자가 찾고자 하는 노드를 추정하여 점진적으로 노드들을 시각화한다.

3. 샘플링을 통한 데이터 시각화

대용량 그래프 데이터는 데이터 전체를 시각화 할 경우 시간이 매우 오래 걸리고, 시각화한 데이터로부터 정보를 찾아내기가 어렵다. 이러한 문제를 해결하기 위해 데이터 샘플링을 하고 샘플링 한 데이터를 시각화 하는 방법이 제안되었다[4, 8].

데이터를 샘플링 할 때는 표본 데이터가 원본 데이터의 특성을 반영해야 하며 샘플링 과정이 효율적으로 수행되어야 한다. 단순한 두 가지 방법은 임의 표본 추출법(random sampling)[4]과 층화 표본 추출법(stratified sampling)인데, 두 방법 모두 원본 데이터의 특성을 잘 반영하지 못하는 경우가 발생한다[8]. 반면 visualization aware sampling(VAS)[8]는 원본 데이터의 특성 중 회귀(regression), 밀도 추정(density estimation)

"이 논문은 2017 년도 정부(미래창조과학부)의 재원으로 한국연구재단-차세대정보・컴퓨팅기술개발사업의 지원을 받아 수행된 연구임(No. NRF-2012M3C4A7033342)."

그리고 클러스터링(clustering)을 잘 반영할 수 있다.

4. 간소화(simplification)을 통한 데이터 시각화

Centrality 를 이용한 간소화

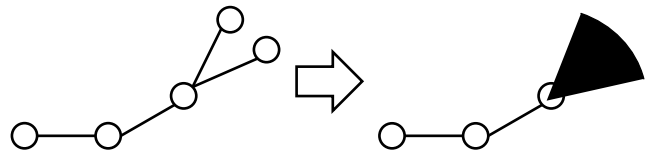
중개중심성은 그래프에서 각 노드의 중심성을 나타내는 척도로 최단 경로를 기반으로 한다[7]. 각 노드의 중개중심성 값은 그래프에서 해당 노드를 통과하는 최단 경로의 개수이다[7]. 소셜 네트워크, 협력 네트워크에서는 중심성이 큰 노드들이 허브 역할을 한다[6].

Girvan[7]은 대용량 그래프를 효율적으로 시각화하기 위해 중개중심성값을 이용한다. 중개중심성 값이 높은 에지를 제거하여 네트워크의 구조를 더 간단한 구조의 네트워크 구조를 얻는다.

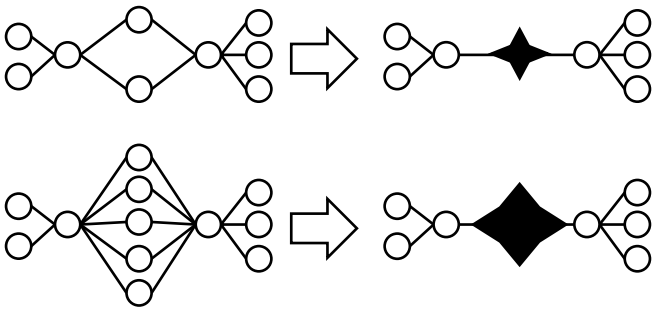
Centrality sensitivity 는 중심성 값의 분포를 나타내고, 그래프의 변화가 어떻게 전파될 것인가에 대한 중요한 척도이다. Correa[2]는 centrality sensitivity 를 이용하여 그래프를 시각화하는 방법을 사용한다.

모티프 간소화

모티프 간소화는 특정 조건을 만족하는 서브그래프(motif)를 특정 모양의 도형으로 대체하여 시각화함으로써 데이터의 양을 감소시키고 가시성을 높이는 방법이다[3].



(그림 1) fan motif 의 시각화 예



(그림 2) D-connector motif 의 시각화 예

Dunne[3]은 fan motif, D-connector motif, D-clique motif 를 사용할 것을 제안한다. 먼저, fan motif 은 잎 노드(leaf node)들과 연결된 node 와 잎 노드들을 의미한다. 그림 1 과 같이 fan motif 를 하나의 도형으로 대체하여 시각화한다. D-connector motif 는 앵커 노드(anchor node)를 연결하는 노드들로, 그림 2 와 같이 D-connector motif 를 하나의 도형으로 대체한다. 마지막으로, D-clique motif 는 각 쌍이 적어도 하나의 링크로 연결된 서브그래프를 말한다. D-clique motif 또한 하나의 도형으로 변환하여 시각화한다. Dunne[3]은 각 motif 를 하나의 도형으로 대체할 뿐만 아니라 대체하는 도형의 크기를 다양화 함으로써 가시성을 향상시킨다. Fan motif 는 잎 노드의 개수가 증가할수록, D-connector motif 는 앵커 노드의 수가 증가할수록, D-clique motif 는 clique 를 구성하는 노드의 수가 증가할수록 도형의 크기를 증가시킨다. 그림 2 는 connector 의 개수가 많을수록 도형의 크기가 커지는 예를 보여준다.

5. Conclusion

본 연구에서는 대용량의 그래프 데이터를 효율적으로 시각화하기 위한 방법에 대해 조사하였다. 상호적인 데이터 마이닝 및 시각화 방법과, 데이터 샘플링을 이용한 시각화 방법, 간소화를 이용한 시각화 방법에 대해 설명하였다. 이러한 연구들은 정적인 데이터는 효율적으로 처리할 수 있다.

소셜 네트워크 서비스, 웹 시멘틱과 같이 그래프 데이터들을 이용하는 응용 중에 데이터가 빈번히 업데이트되는 응용들이 있다. 동적 데이터에 대한 시각화에 대한 기존 연구는 대용량의 그래프 데이터를 효과적으로 지원하지 못한다. 대용량의 동적 그래프 데이터를 효율적으로 시각화하는 방법에 대한 연구가 필요하다.

참고문헌

[1] Akoglu, L., Chau, D. H., Kang, U., Koutra, D., and Faloutsos, C. “Opavion: Mining and visualization in large graphs," In *Proc. 2012 ACM SIGMOD Int'l Conf. on Management of Data*, pp. 717-720, Arizona, USA, May 20-24, 2012.

[2] Correa, C., Crnovrsanin, T., and Ma, L. “Visual reasoning about social networks using centrality sensitivity,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 1, pp 106-120, 2012.

[3] Dunne, C., and Shneiderman, B. "Motif simplification: improving network visualization readability with fan, connector, and clique glyphs." In *Proc. 2013 ACM SIGCHI Int'l Conf. on Human Factors in Computing Systems*, pp. 3247-3256, Paris, France, April 27-May 2, 2013.

[4] Leskovec, J., and Faloutsos, C. “Sampling from large graphs,” In *Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge discovery and data mining.* pp. 631-636, Philadelphia, USA, 2006.

[5] Koutra, D., Jin, D., Ning, Y., and Faloutsos, C. “Perseus: an interactive large-scale graph mining and visualization tool., In *Proc. of the VLDB Endowment*, vol. 8. no. 12, pp. 1924-1927, 2015.

[6] Ma, K, and Muelder, M. “Large-scale graph visualization and analytics,” *Computer*, vol. 46, no. 7, pp. 39-46, Aug. 2013.

[7] M. Girvan and M. Newman, “Community Structure in Social and Biological Networks,” *Proc. Nat'l Academy of Sciences of USA*, vol. 99, no. 12, pp. 7821-7826, 2002.

[8] Park, Y., Cafarella, M., and Mozafari, B. “Visualization-aware sampling for very large databases,” In *Proc. 32h IEEE Int'l Conf. on Data Engineering (ICDE)*, pp. 755-766, Helsinki, Finland, May. 16-20, 2016.