빅데이터 그래프 분석의 핵심 문제들에 대해 세계 최고 성능의 알고리즘 개발

서울대학교 ▌남예현·박근수*

1. 연구배경 및 필요성

전 세계적으로 빅데이터(Big data)가 주목을 받으면서 관련 연구와 기술이 급증하고 있다. 빅데이터를 저장하는 방법들 가운데 대표적인 것이 그래프로, 소셜 네트워크(social network) 분석, 단백질-단백질 상호작용(protein-protein interaction), 화합물 검색(chemical compound search) 등에 쓰인다. 그러나 산업계와 응용분야에서 요구되는 그래프 분석 기술은 대부분 NP-hard라고 불리는 난이도 높은 문제들이다. 현존하는 그래프 관련 오픈소스 라이브러리들은 대부분 난이도가 낮은 문제를 위한 알고리즘만을 제공하고 있으며 극히 일부 라이브러리만이 NP-hard 그래프 문제에 대한 기초적인 알고리즘을 제공하고 있다.

2. 기술의 내용 및 성과의 우수성

본 연구진은 빅데이터 그래프 분석의 핵심 문제들에 대해 기존 세계 최고 성능을 깨는 알고리즘들을 연달아 발표하였다 (도장(道場)깨기를 진행하고 있음). 아래에 있는 그래프 분석의 핵심 문제에서 기존의 알고리즘 대비 성능이 수십 배 내지 수백 배 빠른 알고리즘들을 제시하였다.

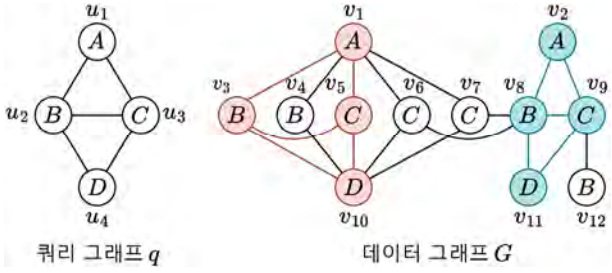

그림 1 부분그래프 매칭 예시

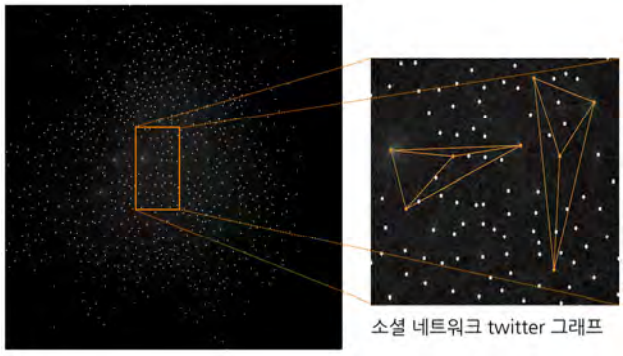

그림 2 소셜 네트워크에서의 부분그래프 매칭

2.1 부분그래프 매칭 및 부분그래프 질의 처리

부분그래프 매칭 (subgraph matching) 알고리즘 및 부분 그래프 질의 처리 (subgraph query processing) 알고리즘은 소셜 네트워크 등의 빅데이터 그래프에서 특정한 패턴을 찾아내는 기술이다. 부분그래프 매칭은 데이터 그래프에서 쿼리 그래프의 모든 임베딩(embedding)을 찾아내는 문제이다. 이때 임베딩은 쿼리 그래프의 각 정점을 데이터 그래프의 정점에 대응시키는 단사함수로, 대응되는 정점은 그 레이블(label)이 같아야 하고, 쿼리 그래프의 각 간선에 대해 양 끝 두 정점의 대응을 연결하는 간선이 데이터 그래프에도 있어야한다. 그림 1은 부분그래프 매칭 문제의 예시로, 데이터 그래프 G에서 쿼리 그래프 q를 찾으면 두 개의 임베딩 {($u1$, $v1$), ($u2$, $v3$), ($u3$, $v5$), ($u4$, $v10$)}과 {($u1$, $v2$), ($u2$, $v8$), ($u3$, $v9$), ($u4$, $v11$)} 이 나온다. uv 부분그래프 질의 처리는 다수의 그래프와 하나의 쿼리 그래프가 주어졌을 때, 쿼리 그래프의 임베딩이 존재하는 모든 데이터 그래프를 찾아내는 문제이다. 그림 2는 Twitter 그래프에서 정점 4개로 구성된 패턴을 검색한 결과를 보여준다. 본 연구진의 질의 처리 알고리즘과 매칭 알고리즘은 기존 연구 대비 수행시간을 각각 41741%, 3186% 개선하였다.

* 종신회원

2.2 슈퍼그래프 검색

슈퍼그래프 검색 (supergraph search) 알고리즘은 다수의 데이터 그래프들과 하나의 쿼리 그래프에 대하여, 쿼리 그래프에 부분그래프로 포함된 데이터 그래프들을 찾아내는 기술이다. 본 연구진의 알고리즘은 기존 연구 대비 수행시간 개선율 9205%를 달성하였다.

2.3 연속적 부분그래프 매칭

연속적 부분그래프 매칭 (continuous subgraph matching) 알고리즘은 데이터 그래프가 변화할 때마다 쿼리 그래프와 동형이면서 새로 생기거나 없어지는 데이터 그래프의 부분그래프를 찾아내는 분석 기술로서 사이버 보안, 사기 탐지, 소셜 네트워크 서비스 등에서 이용된다. 본 연구진의 알고리즘은 기존 연구 대비 수행시간 개선율 8823%를 달성하였다.

2.4 그래프 동형 및 그래프 동형 질의 처리

그래프 동형 (graph isomorphism) 알고리즘은 두 개의 그래프가 동형인지 판별하는 알고리즘으로 본 연구진의 알고리즘은 지난 30여 년 동안 최고 성능의 알고리즘이었던 nauty/Traces에 비해 획기적인 성능 향상을 얻었다. 본 연구진의 알고리즘은 기존 연구 대비 수행시간 개선율 12529%를 달성하였다. 그래프 동형 질의 처리 (graph isomorphism query processing) 알고리즘은 다수의 데이터 그래프와 하나의 쿼리 그래프가 주어졌을 때, 쿼리 그래프와 동형인 데이터 그래프를 전부 찾아내는 기술로, 본 연구진의 알고리즘이 기존 연구 대비 수행시간 개선율 583%를 달성하였다.

2.5 다각적 top-k 부분그래프 질의 처리

다각적 top-k 부분그래프 질의 처리 (diversified top-k subgraph query processing) 알고리즘은 데이터 그래프와 쿼리 그래프, 그리고 정수 k가 주어졌을 때, 데이터 그래프에서 쿼리 그래프와 동형인 부분 그래프들 가운데 서로 다른 정점을 가장 많이 포함하는 k개의 부분 그래프를 찾는 기술이다. 본 연구진의 알고리즘이 기존 연구 대비 수행시간 개선율 705%를 달성하였다.

이에 관한 논문 총 6편이 컴퓨터 분야의 최우수학술대회인 SIGMOD 2019와 SIGMOD 2021 (‘가’의 결과), VLDB 2020 (‘나’의 결과), VLDB 2021 (‘다’의 결과), 그리고 ICDE 2021과 ICDE 2022 (‘라’의 결과)에 발표되었다.

본 연구진의 연구 결과는 2022년 국가연구개발우수성과 100선 및 정보전자 분야 최우수성과에 선정되었다.

3. 과학기술적 파급효과

앞서 소개한 그래프 분석 기술들에 대하여 그동안 전 세계적으로 많은 연구가 진행되었고 다양한 알고리즘들이 개발되어 왔다. 본 연구진은 이 그래프 분석 기술들에 대해 이전 알고리즘 대비 수십 배 내지 수백 배 빠르게 해당 문제들을 해결함으로써 획기적인 성능 개선을 얻었다. 이러한 연구 결과에 대해 University of Edinburgh, New York University, Arizona State University, Hong Kong University of Science and Technology, Chinese University of Hong Kong, Peking University, Fudan University, Osaka University, University of Verona, University of Salemo, Eindhoven University of Technology, University of Sydney 등 30여 개 해외 유수 대학의 연구진들로부터 알고리즘에 대한 문의와 코드 공유 요청을 받았다. 또한 일본의 통신기업 NTT와 신약개발 관련 스타트업 ㈜AIgenDrug으로부터도 코드 공유 요청을 받았다. 이에 본 연구진이 개발한 알고리즘의 코드를 오픈 소프트웨어 형태로 GitHub에 공개하였다.

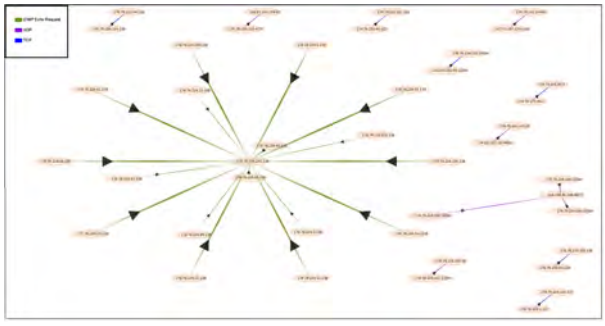

그림 3 네트워크 상에서 DDoS 공격 그래프

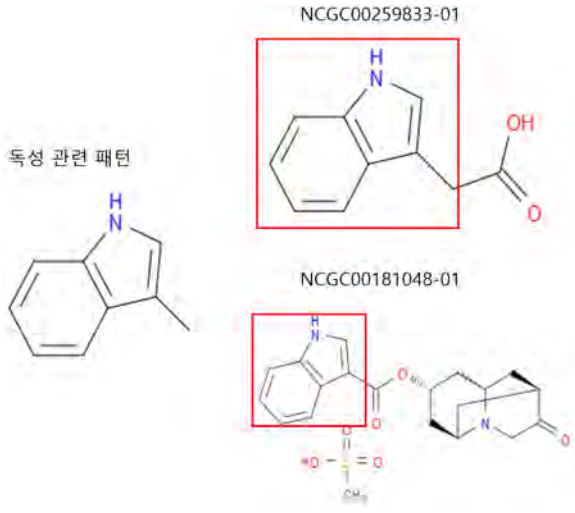

그림 4 독성 관련 화합물 검색

4. 경제사회적 파급효과

현재 빠른 속도로 빅데이터가 생성되고 있고 이를 활용한 서비스가 크게 발전하고 있다. 그래프 빅데이터 생성과 분석도 활성화되어 대규모 그래프 문제에 대한 효율적인 알고리즘에 대한 수요가 증가하고 있다. 본 연구진의 기술은 빅데이터 그래프에서 특정한 패턴 검색, 사이버 공격 탐지, 신약개발 등에 사용될 수 있다. 그림 3은 네트워크 상에서 DDoS 공격을 피해자를 중심으로 그래프로 표현한 것이고, 본 연구진이 개발한 알고리즘을 통해 DDoS 공격을 탐지할 수 있다.

그림 4는 화합물 데이터베이스에서 본 연구진의 알고리즘을 사용해서 독성을 띠는 패턴을 포함하고 있는 화합물을 검색한 결과를 보여준다. 본 연구진이 개발한 그래프 분석 기술을 사용하면 신약개발의 중요 과정인 독성 분석이 용이해지고 이로 인해 신약개발에도 도움이 될 것으로 기대된다. 한 예로, 신약개발 관련 스타트업인 ㈜AIgenDrug에서 화합물 데이터베이스 분석을 위해 본 연구진이 개발한 알고리즘을 요청함에 따라 소스코드를 제공하였다.

약 력

남 예 현

2020 서울대학교 컴퓨터공학부 학사
2020~현재 서울대학교 컴퓨터공학부 석박사통합과정
관심분야 : 그래프 알고리즘
Email : yhnam@theory.snu.ac.kr

박 근 수

1983 서울대학교 컴퓨터공학과 학사
1985 서울대학교 컴퓨터공학과 석사
1992 미국 Columbia 대학교 전산학 박사
1991.11~1993.8 영국 런던 대학교 King's college 조교수
1993.8~현재 서울대학교 컴퓨터공학부 교수
관심분야 : 컴퓨터이론, 그래프 알고리즘, 생물정보학, 암호학
Email : kpark@theory.snu.ac.kr