# 대용량 그래프 환경에서 스카이라인을 이용한 서브 그래프 유사도 측정 기법

## A Similarity Measurement Scheme using Skyline Queryin Large-scale Graph Environments

임 종 태, 복 경 수, 유 재 수*
충북대학교 정보통신공학과

Jongtae Lim, Kyoungsoo Bok, Jaesoo Yoo
School of Information and Communication Engineering, Chungbuk National University

요약

최근 각종 실험 장비의 발전에 따라 유사 서브 그래프 매칭에 대한 연구가 활발하게 진행되고 있다. 하지만 유사 서브 그래프 매칭은 다수의 최종 결과들이 반환되었을 경우, 사용자는 어떤 결과가 자신에게 가장 유의미한 결과인지를 판별하기 힘든 문제점이 존재한다. 본 논문에서는 대용량 그래프 환경에서 스카이라인을 이용한 서브 그래프 유사도 측정 기법을 제안한다. 제안하는 기법은 스카이라인 기법과 피드백에 기반한 랭킹을 수행하여 유사 서브 그래프 매칭에서 사용자에게 유의미한 결과를 반환한다.

## Ⅰ. 서론

최근 각종 실험 장비의 발전에 따라 첨단 분야에서의 과학데이터가 급격히 증가하고 있으며 생물정보학 분야 역시 생산되는 데이터가 급격히 증가하는 추세이다. 이러한 환경에서 대상이 되는 그래프에서 질의로 주어진 그래프와 일치하는 서브 그래프를 검색하는 서브 그래프 매칭(Subgraph matching)이 중요하게 연구되고 있다. 서브 그래프 매칭은 컴퓨터 네트워크 모니터링, 소셜 네트워크의 진화 분석, 생물학 네트워크에서 모티프(Motif) 탐지 등, 네트워크 분석 및 데이터 마이닝 분야에서 널리 활용되어 왔다.

서브 그래프 매칭은 질의에 따라 완전 일치 서브 그래프 매칭(Exact subgraph matching)[1]과 유사 서브 그래프 매칭(Approximate subgraph matching)[2]으로 구분된다. 완벽하게 일치하는 결과만을 검색하는 완전 일치 서브 그래프 매칭에 비해 유사 서브 그래프 매칭은 질의 그래프와 비슷한 결과까지 검색하기 때문에 다양한 분야의 응용 및 연구에서 활용된다. 하지만 유사 서브 그래프 매칭은 질의 그래프와 유사한 결과들을 검색하기 때문에 다수의 최종 결과들이 반환될 수 있다. 특히, 대용량 그래프 환경에서 유사 서브 그래프 매칭 결과로 다수의 최종 결과들이 반환되었을 경우 사용자는 어떤 결과가 자신에게 가장 유의미한 결과인지를 판별하기 힘든 문제점이 존재한다.

* 교신저자 : yjs@chungbuk.ac.kr
본 연구는 미래창조과학부 및 정보통신기술진흥센터의 대학 ICT연구센터육성 지원사업(IITP-2017-2013-0-00881), 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원(No. 2016R1A2B3007527), 그리고 2014년도 산업통상자원부의 재원으로 한국에너지기술평가원(KETEP)의 에너지인력양성사업(No. 20164030201330)의 연구결과로 수행되었음

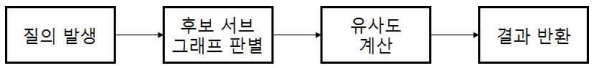
본 논문에서는 대용량 그래프 환경에서 스카이라인을 이용한 서브 그래프 유사도 측정 기법을 제안한다. 제안하는 기법은 스카이라인 기법[3]을 활용하여 다중 속성을 고려한 유사 서브 그래프 매칭에서 사용자에게 유의미한 결과를 반환한다. 또한 사용자가 반복적으로 유사 서브 그래프 매칭을 수행할 경우, 이전 결과에서 선택한 결과(피드백)들을 축적하고 이를 유사도 측정에 반영하여 보다 사용자에게 적합한 결과를 반환한다.

## Ⅱ. 제안하는 서브 그래프 유사도 측정 기법

### 1. 특징

제안하는 유사도 측정 기법은 여러 가지 유사도 계수가 고려된 유사 서브 그래프 질의에서 사용자가 가중치를 특별히 입력하지 않아도 스카이라인 기법을 이용하여 결과 중에 사용자에게 추천할 만한 결과를 판별한다. 또한, 유사 서브그래프 매칭이 반복 수행될 경우 피드백을 반영하여 사용자에게 결과를 추천해주는 것을 목적으로 한다.

그림 1은 유사 서브그래프 매칭의 처리 흐름을 나타낸다. 질의는 질의의 기준이 될 질의 그래프와 검색의 대상이 될 대용량 그래프가 입력된다. 질의가 입력되면 유사 서브 그래프 매칭 기법은 대용량의 대상 그래프로부터 가지치기 기법을 수행하여 실제 비교 대상이 되는 후보 서브 그래프를 판별한다. 마지막으로 유사 서브 그래프 매칭 기법은 질의 그래프와 후보 서브 그래프를 다양한 유사도 계수를 통해 비교를 수행하여 각 후보 그래프가 질의 그래프와 얼마나 유사한지를 측정한다.

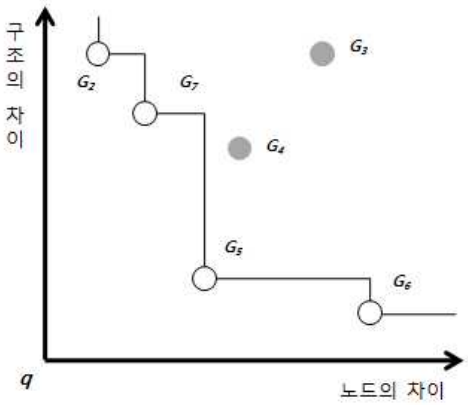



▶▶ 그림 1. 유사 서브 그래프 매칭 처리 절차

## 2. 스카이라인을 이용한 유사도 측정

기존 유사 서브 그래프 매칭 기법에서 사용자는 전체 처리 절차 중에 유사도를 계산하여 결과를 반환하기 위해 필수적으로 가중치를 입력해야 한다. 사용자 중에는 가중치를 적절하게 설정할 수 있는 사용자가 존재하는 반면, 그렇지 않은 사용자도 존재한다. 이러한 사용자가 적절한 결과를 반환받기 위해서는 가중치를 설정하기 위한 실험적인 비용이 추가로 발생한다.

본 논문에서는 스카이라인 질의 처리 기법을 이용하여 가중치를 입력하지 않아도 유의미한 결과를 반환할 수 있는 유사도 측정 기법을 제안한다. 그림 2는 제안하는 스카이라인 기법을 활용한 유사도 측정 기법을 나타낸다. 그림 2에서는 표현하기 쉽게 2차원의 스카이라인 처리로 표현했다. 질의 그래프 q에 대하여 구조의 차이와 노드의 차이를 고려하여 그래프를 구성할 경우 대상 그래프들이 그림 2와 같이 위치한다. 스카이라인 질의 처리 기법은 다음과 같은 대상 그래프 중에서 $G_2$, $G_7$, $G_5$, $G_6$와 같이 지배되지 않는 결과만을 반환하며, $G_4$, $G_3$와 같이 지배되는 결과는 제외한다. 그리고 반환되는 결과의 유사도 측정 기법은 크게 두 가지 방법을 사용한다. 첫 번째는 그림 2와 같은 스카이라인 그래프를 기준으로 지배하는 영역이 넓은 결과를 상위 랭크로 반환하는 방법이다. 이 경우 결과는 〈$G_5$, $G_7$, $G_6$, $G_2$〉 순으로 반환된다. 두 번째는 통계적 기법 활용이 가능할 경우, 다른 대상 그래프들을 많이 지배하는 결과를 상위 랭크로 반환하는 방법이다. 이 경우 결과는 〈$G_5$, $G_7$, $G_2$, $G_6$〉 또는 〈$G_5$, $G_2$, $G_7$, $G_6$〉 순으로 반환된다. 만약 사용자가 반환받으려고 하는 결과 개수가 스카이라인 결과 개수보다 많은 경우, 이미 반환된 스카이라인 결과 이외의 대상 그래프들로 스카이라인 처리를 반복 수행하여 충분한 결과를 반환한다.




▶▶ 그림 2. 스카이라인을 이용한 유사도 측정

## 3. 피드백 이용한 유사도 보정

스카이라인 질의 처리를 이용한 유사도 측정 기법의 경우 사용자가 가중치를 입력하지 않고도 유의미한 유사 서브 그래프를 반환하는 방법이다. 하지만 사용자가 직접 입력한 가중치를 이용한 방법과 비교를 했을 때, 이 결과는 사용자에게 적합한 결과로 정렬이 되었는지 확실하지 않다. 따라서 이를 보완하기 위하여 제안하는 기법에서는 사용자가 선택한 결과인 피드백을 축적하여 이를 유사도 보정에 활용한다. 그림 2와 같은 환경에서 결과가 〈$G_5$, $G_7$, $G_6$, $G_2$〉로 반환되었을 때, 이 중에서 사용자가 가장 유의미한 유사 서브 그래프로 $G_2$를 선택했다면 이 사용자는 구조의 차이보다는 노드의 차이를 더 중요하게 고려하는 사용자라고 추측할 수 있다. 따라서 이렇게 반복적으로 입력되는 피드백들을 가중치로 반영하여 유사도를 하나의 값으로 계산한다.

제안하는 기법에서는 이렇게 계산된 유사도를 통해 사용자의 피드백을 고려한 유사도 보정을 수행한다. 피드백을 고려한 최종 유사도는 식 (1)와 식 (2)을 통해 계산된다. 식에서 threshold(thrld)는 사용자가 유사 서브 그래프 매칭을 반복 수행하였을 때, 몇 번의 피드백이 축적되어야 이를 사용자에 적합한 결과를 도출할 수 있는 충분한 피드백이 축적된 것으로 판단할 것인지를 결정하는 수치이다. #feedback(#fdb)은 피드백이 축적된 개수이다. 식 (1)을 통해 피드백이 축적되지 않은 초기에는 스카이라인을 이용한 유사도(가중치 모두 동일) 기반의 결과를 사용자에게 반환하고, 추후 피드백이 축적됨에 따라서 피드백을 통해 보정된 유사도 기반의 결과 비율을 점진적으로 증가시킨다. 마지막으로 충분한 피드백이 축적된 이후에는 식 (2)와 같이 피드백을 통해 보정된 유사도 기반의 결과만을 반환한다.

$$sim(q,G) = \frac{thrld - \#fdb}{thrld} sim_{skyline}(q,G) + \frac{\#fdb}{thrld} sim_{feedback}(q,G) \quad (1)$$

$$(\text{if } \#feedback \le threshold)$$

$$sim(q,G) = \frac{\#fdb}{thrld} sim_{feedback}(q,G) \quad (2)$$

$$(\text{if } \#feedback > threshold)$$

## Ⅲ. 결론

본 논문에서는 대용량 그래프 환경에서 스카이라인을 이용한 서브 그래프 유사도 측정 기법을 제안했다. 제안하는 기법은 스카이라인 기법과 피드백에 기반한 유사도 측정을 수행하여 유사 서브 그래프 매칭에서 사용자에게 적합한 결과를 반환한다. 향후 연구로는 유사 서브 그래프 매칭이 활용되는 다양한 성능평가를 수행하여, 제안하는 기법의 우수성을 입증할 예정이다.

## ■ 참 고 문 헌 ■

[1] T. Nilothpal, and Z. J. Mohammed, “A distributed approach for graph mining in massive networks”, Data Mining and Knowledge Discovery Vol.30, No.5, pp.1024-1052, 2016.

[2] Y. Tian, R. C. McEachin, C. Santos, D. J. States, and J. M. Patel “SAGA: a subgraph matching tool for biological graphs,” Bioinformatics, Vol.23, No.2, pp.232-239, 2007.

[3] S.Borzsonyi, D. Kossmann, and K. Stocker, “The Skyline Operator,” Proc. of IEEE International Conference on Data Engineering, pp.421-430, 2001.