DAG에 대한 2-Hop 레이블 크기를 줄이기 위한 노드 아이디 부여 기법 설계

안진현*, 임동혁**, 김홍기*
*서울대학교 의생명지식공학 연구실
**호서대학교 컴퓨터정보공학부
e-mail: jhahncs@snu.ac.kr, dhim@hoseo.edu, hgkim@snu.ac.kr

A Design of Node ID Assignment for 2-Hop Label Size Reduction of DAG

Jinhyun Ahn*, Dong-Hyuk Im**, Hong-Gee Kim*
*Biomedical Knowledge Engineering Laboratory, Seoul National University
**Department of Computer and Information Engineering, Hoseo University

요 약

링크드오픈데이터를 통해 다양한 분야의 RDF 데이터가 공개되고 있으며 그 양이 지속적으로 증가하고 있다. RDF 데이터는 그래프 형태이기 때문에 대용량 RDF 데이터를 효율적으로 관리하기 위한 그래프 데이터베이스에 대한 연구가 중요하다. 2 개의 RDF 리소스가 그래프 상에서 연결됐는지 여부를 알아내는 기능은 RDF 요소간 연관관계를 식별하는 데에 관련이 있기 때문에 그래프 데이터베이스의 중요한 기능 중 하나이다. 대용량 그래프 데이터에 대한 그래프 도달가능성을 빠르게 처리하기 위해 2-Hop 레이블링 변형들이 제안됐다. 최근에 2-Hop 레이블 크기를 줄이기 위해 2-Hop 레이블링이 진행되기 전에 노드 아이디를 부여하는 방법이 제안됐다. 하지만 그래프의 지역 정보만을 활용하기 때문에 복잡한 형태의 그래프에 대해서는 비효율적이라는 문제점이 있다. 본 논문에서는 그래프의 전역 정보를 반영할 수 있는 Topological Sort 를 활용한 노드 아이디 부여 기법에 대한 설계를 제안한다.

1. 서론

RDF(Resource Description Framework)는 주어(subject), 술어(predicate), 그리고 목적어(object)로 구성된 트리플로 정보를 표현하는 방법론이다. 최근 다양한 분야의 RDF 데이터가 링크드오픈데이터 형태로 공개가 되고 있다. RDF 데이터는 그래프이기 때문에 대용량 RDF 데이터로부터 지식을 추출하기 위한 그래프 데이터베이스가 활발히 연구되고 있다. RDF 요소간 연결관계를 알아내는 작업은 연관관계를 식별할 수 있다는 관점에서 그래프 데이터베이스의 중요한 기능 중 하나이다. 즉, 그래프 상에서 노드간 도달가능성을 알아내는 작업이다.

그래프에서의 도달가능성과 DAG(Directed Acyclic Graph)에서의 도달가능성은 동일하다[6]. 주어진 그래프에서의 모든 SCC(Strongly Connected Component)를 하나의 노드로 변환하고 SCC 안에 있는 노드와 외부 노드간의 간선을 그대로 유지하면 DAG 를 얻을 수 있다. DAG 에서 도달가능성을 빠르게 알아낼 수 있는 다양한 방법이 제안됐다. 본 연구에서는 2-Hop 레이블링 방법에 초점을 맞춰 2-Hop 레이블의 크기를 줄이기 위한 노드 아이디 부여 방법을 제안한다.

2. 관련연구

그래프 도달가능성 판단을 위한 가장 단순한 방법은 모든 노드에 대해 TC(Transitive Closure)를 미리 계산하는 방법이다[1]. 하지만, 그래프가 조금만 커지면 TC 크기는 기하급수적으로 커진다는 단점이 있다. DFS(Depth-First-Search)나 BFS(Breadth First Search)와 같은 온라인 그래프 탐색 방법은 인덱스가 필요 없지만 질의처리 시간이 길다는 단점이 있다. 인덱스의 크기와 질의처리시간의 trade-off 관계에서 중간에 위치한 방법들이 제안됐다; 소수 레이블링[2] Tree Cover[3] interval 레이블링[4]. 본 연구에서 다루는 2-hop 레이블링[5]은 각 노드 v 를 $(L_{out}(v), L_{in}(v))$ 쌍으로 레이블링 하는 방법으로 교집합 연산으로 도달가능성을 판별한다. 2-Hop 레이블링의 최신 연구에서는 MinHash 기법에 기반하여 k개의 도달가능한 노드 집합만 유지하는 방법을 제안했다[6]. k개의 노드들로 도달가능성을 판별할 수 없는 경우 DFS 를 수행한다.

3. Topological Sort 기반 노드 아이디 부여

DAG $G(V,E)$에 대한 2-Hop 레이블링은 각 노드 $v \in V$에 $(L_{out}(v), L_{in}(v))$ 쌍 레이블을 구성하는 작업이다. 여기에서 $L_{out}(v)$는 v로부터 도달할 수 있는 노드들의 집합이고 $L_{in}(v)$은 v에 도달할 수 있는 노드들의 집합이다[5]. 정의 1은 최신의 2-Hop 레이블링 방법의 변형들 중 하나인 IP이다.

정의 1. 2-Hop 레이블링(IP) [6]

$G(V,E)$를 노드 집합 V와 간선 집합 E로 구성된 DAG 라고 하자. $u \in V$가 $v \in V$에 도달 가능하다면 $u \Rightarrow v$, 그렇지 않을 경우 $u \nRightarrow v$로 표기한다. V에 대한 순열매핑 $\sigma : V \rightarrow V$는 전단사(bijection) 함수로 각 노드에 부여된 아이디를 의미한다. 주어진 G 와 양의정수 k에 대해, IP는 여러 개의 순열매핑 σ_i 중 무작위로 하나를 선택하고 2-Hop 레이블 $I^k_{\sigma_i} : V \rightarrow H$를 생성한다. 여기에서 H는 $(L_{out}(v), L_{in}(v))$ 쌍들의 집합이며 다음과 같이 정의된다:

$$L_{out}(v) = min_k\{\sigma(u) | v \Rightarrow u\}$$
$$L_{in}(v) = min_k\{\sigma(u) | u \Rightarrow v\}$$

,여기에서 $min_k\{A\}$는 $|min_k\{A\}| \leq k$를 만족하는 A의 부분집합이고, 만약 $i < j$이면 $A_i < A_j$이다. 다시 말해, $min_k\{A\}$는 A에 있는 원소들 중 가장 작은 k개를 가진다.

정의 2는 주어진 DAG G에 정의 1에 따른 2-Hop 레이블링을 수행해서 얻은 레이블의 크기에 대한 정의로 $I^k_{\sigma_i}(v)$에 있는 노드 아이디들의 총합이다.

정의 2. 2-Hop 레이블 크기

$$\sum_{v \in V} (\sum_{w \in L_{out}(v)} w + \sum_{q \in L_{in}(v)} q)$$

IP에 따르면 σ_i는 Knuth shuffle 알고리즘에 의해 무작위로 생성된다[6]. 이 방법은 2-Hop 레이블 크기를 조절할 수 없다는 한계점이 있다.

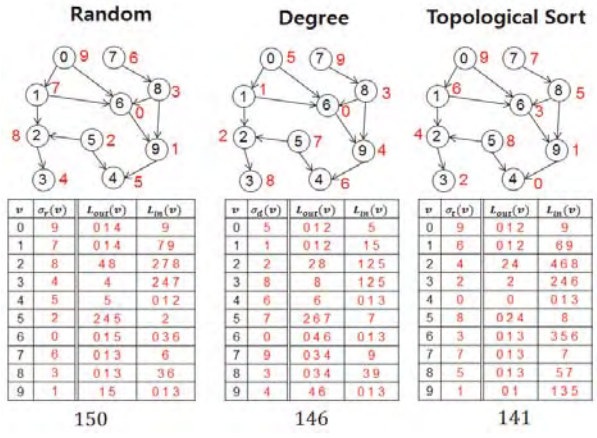

그림 1 노드 아이디 부여 방법에 따라 다른 2-Hop 레이블 크기

그림 1에 서로 다른 아이디 부여 방법에 의한 2-Hop 레이블 인덱스와 레이블 크기가 표현됐다. Random은 아이디를 임의로 부여하는 방법이고, Degree는 각 노드의 간선의 개수가 많을 수록 작은 숫자를 부여하는 방법이다[7]. Degree 방법은 어떤 노드 v가 간선이 많으면 그만큼 v로부터 도달 가능하거나 v에 도달 가능할 확률이 높다는 가정을 바탕으로 하고 있다. 노드 6의 경우 간선의 개수가 많기 때문에 가장 작은 아이디인 0을 부여했다. 하지만, Degree 방법은 노드의 간선의 개수만 고려하기 때문에 전체적인 그래프 형태를 고려하지 못한다는 한계가 있다.

Topological Sort는 본 논문에서 제안한 방법으로 다음과 같이 아이디를 부여했다. 1) G에 Topological Sort를 수행하여 V에 대한 정렬된 집합을 얻고 그것의 역정렬된 집합 V^{TS}를 취한다. 2) 순열매핑을 다음과 같이 정의한다. $\sigma_t(v) = i$. 여기에서 v는 V^{TS}의 i번째 원소이다.

노드 3의 경우 Degree 방법은 8을 Topological Sort 방법은 2를 부여했다. 노드 3의 경우 간선의 개수는 한 개이지만 노드 0,1,2로부터 도달 가능하기 때문에 다른 노드에 포함될 확률이 높다고 할 수 있다. 실제로 그림 1의 예제의 경우 레이블 크기가 줄었다.

4. 결론

본 논문에서는 2-Hop 레이블 크기를 줄이는 방법에 있어서, 기존 Degree 방법이 지역 정보만 활용한다는 한계점을 극복하기 위해 전역 정보를 반영한 Topological Sort 방법을 제안했다. 향후 연구에서는 본 방법을 구현하여 실제 링크드오픈데이터의 데이터로 검증을 할 예정이다. 또한, Topological Sort가 모든 DAG 형태에 적합한 것이 아니므로 Topological Sort와 Degree를 모두 고려한 방법을 고안할 예정이다.

Acknowledgment

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (2013-0-00109, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)

참고문헌

[1] Simon, K. “An improved algorithm for transitive closure on acyclic digraphs” Theoretical Computer Science 58(1), 325-346 (1988)

[2] Wu, G., Zhang, K., Liu, C., Li, J. “Adapting prime number labeling scheme for directed acyclic graphs” DASFAA, 2006

[3] Agrawal, R., Borgida, A., Jagadish, H.V. “Efficient management of transitive relationships in large data and knowledge bases”, SIGMOD, 1989

[4] Seufert, S., Anand, A., Bedathur, S., Weikum, G. “Ferrari: Flexible and efficient reachability range assignment for graph indexing” ICDE, 2013

[5] Cohen, E., Halperin, E., Kaplan, H., Zwick, U. “Reachability and distance queries via 2-hop labels” SIAM Journal on Computing 32(5), 1338-1355 (2003)

[6] Wei, H., Yu, J.X., Lu, C., Jin, R. “Reachability querying: An independent permutation labeling approach” VLDB, 2014

[7] Jinhyun Ahn, "Optimization Techniques for 2-hop Labeling of Dynamic Directed Acyclic Graphs", ISWC-DC, 2016