✅

DeepWalk

Domain	Graph
tag	Graph learning, Multi-label classification, Pattern recognition
Conference / Journal	KDD
Publish year	2014
정리 날짜	@2024년 1월 2일
AI summary	DeepWalk은 자연어 처리에서 사용되는 기법을 그래프 표현 학습에 적용하는 방법으로, 랜덤 워크를 통해 그래프 데이터를 변환하고, Skip-Gram을 사용하여 학습합니다. 이를 통해 그래프의 vertices 간의 social representation을 학습하며, Multi-label classification task에서 좋은 성능을 보입니다.
AI key info	Graph Dense Representation, Random Walk, Skip-Gram, Label, Collective classification problem, DeepWalk, Random Walk, Skip-Gram, Latent representation, Social representation of vertices

Summary

자연어 처리에서 활용되는 기법을 Graph Dense Representation 학습에 적용하고,

이를 위해 Random Walk라는 변환 과정을 이용함

변환한 데이터를 자연어 처리에서 사용되는 word2vec인 Skip-Gram을 활용함

Lable과 그래프 구조를 구분하여 학습함

- 짧은 random walk를 사용하여 그래프의 vertices간의 social representation(latent)을 학습

Background

기존의 classification ML은 element를 label에 매칭하는 hypothesis를 학습하는 것을 목표로 함

- DeepWalk는 example의 dependency 정보를 활용함?

Collective classification problem

- 네트워크 내 노드를 분류할 때, 한 노드의 분류가 다른 노드들의 분류에 영향을 받는 상황
  - 전통적으로는 undirected Markov network(노드를 변수로 치환) 활용해 iteration
  - 이는 lable을 feature의 하나로 사용함
- Deepwalk는 lable과 독립적으로 그래프 구조를 학습하는 unsupervised learning 사용
  - cascading error 회피 가능: 한 노드의 실패가 연속적으로 다른 노드 실패

Lable node가 부족할 때

- 일반화가 어려움

Methodology

DeepWalk

- 그래프에서 sequence를 생성해 Skip-Gram 방식으로 학습
  - 그래프에서 random walk 방식으로 sequence를 생성함
  - 생성한 sequence로 word2vec방식 중 하나인 Skip-Gram 사용하여 학습

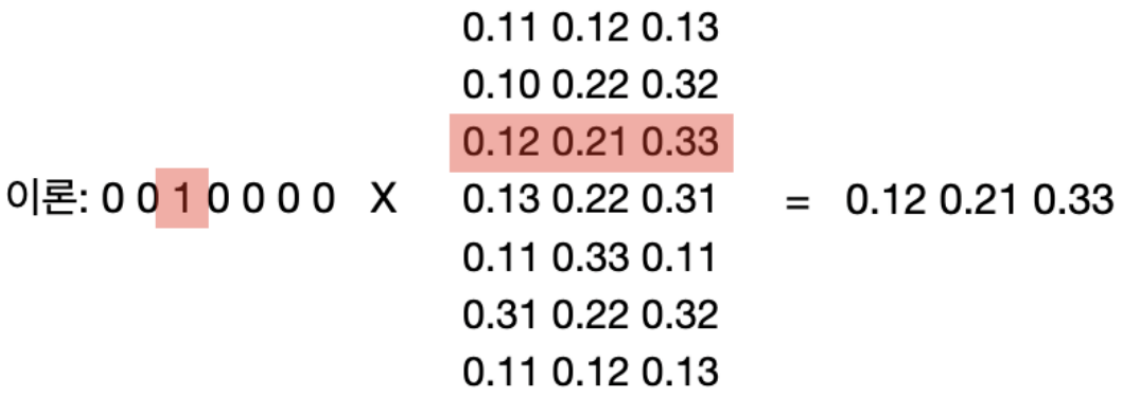
Random Walk

- vertex Vi의 이웃을 랜덤하게 골라 탐색 반복
  - 병렬화 가능

    - Short random walk이기에 그래프 일부분 업데이트시 전체가 아닌 일부만 계산하면 됨
- 이를 반복하여 vertex로 이루어진 시퀀스 만듬

Skip-Gram

- 시퀀스에서 대상 단어의 주변 단어를 예측
    - 주변 단어와 연관성이 높다고 판단
- ex. ['나는', '딥러닝', '공부를', '카페에서', '책을', '보며', '하고있다']
    - '공부를'과 '나는', '딥러닝', '카페에서', '책을'과 연관성이 높다고 판단
- 각 sequence의 위치를 one-hot vector로 표시하여, 가중치 행렬 생성
    - '딥러닝' → [0 1 0 0 0 0 0]



Conclusion

여러 ML 알고리즘과 융합해서 사용 가능

- 발전된 Language model에 적용할 수 있음

Scailabilty

- 큰 그래프에서 outperform
- 동시에 업데이트하는 병렬화 가능

Multi-label classification task에서 좋은 성능

- 60% 적은데이터로 5~10% 수준의 Micro F1 향상 보임

Questions

1. 기존의 방식과 다르게 라벨과 노드 구조를 분리해서 노드 구조만 사용해서 학습하는데, multi-label classification에서 왜 lable 노드 수의 변화에 따라 성능이 변화하는지 이해가 안됨. 라벨은 학습 시에 사용하지 않는게 아닌가?
    a. classifier를 학습할 때 라벨링 데이터를 사용함. 비교한 다른 논문들도 Unsupervised learning이지만, classifier를 학습함으로써 라벨 데이터의 비율에 따라 성능이 바뀜
    b. 반면 GCN같은 경우 라벨을 사용하므로, 학습 단계에서 반영

Terminology

- latent representation: 복잡한 고차원 데이터를 단순한 저차원 공간으로 변형한 벡터 표현
- Social representation of vertices