**논문 요약**

**날짜: 2020.12.02**

**이름: 전우진**

|  |  |
| --- | --- |
| 논문제목 | Applications of link prediction in social networks: A review |
| 논문 요약 | 소셜 네트워크는 현대 사회가 인터넷을 통해 사회적 상호작용을 수행하고 정보를 교환하는데 가장 선호하는 수단 중 하나이다.  그것은 사람들이 경계 없이 자신의 내용을 공유할 수 있게 해준다.  사람들은 소셜 네트워크를 통해 빠르고 쉬운 공유 방법을 제공하므로 자신의 의견, 좋아요, 관심사를 소통하고 표현한다.  그 결과, 소셜 네트워크 내에서 복잡한 연결을 생성하며, 이는 큰 그래프로 즉시 시각화 할 수 있다.  지난 10년 동안 링크 예측의 인기는 다른 도메인에서 링크 예측의 추가 개선과 적용을 위해 훨씬 더 많은 연구를 끌어 모았다.  본 논문은 기존 검토의 격차를 줄이기 위해 최신 접근법과 응용프로그램을 포함한 새로운 연계 예측 조사를 제시하고, 대규모 네트워크, 다차원 네트워크, 확장성, 네트워크 동적성 등 현재의 문제를 해결하는 것을 목표로 한다.  결론적으로 소셜 네트워크는 2018년 10월(Kemp, 2018년) 2억 2,3400만 명의 액티브 사용자를 보유한 페이스북이 입증한 것처럼 인적 소통에 중요하고 편리하다.  활동 중인 사용자들은 활동에서 방대한 데이터를 생성하여 연구자들이 다양한 소셜 네트워크 분석을 수행할 수 있는 기회를 제공하였다.  링크 예측은 기존 연결을 기반으로 연결되지 않은 링크 쌍 간의 향후 연결을 예측하는 과정이다. 링크 예측은 미래 연결을 분석 및 예측하는 것 외에도, 주어진 미래 시간인 t'에 시간 간격 동안 네트워크에 나타나거나 사라질 수 있는 누락된 링크를 주입한다(Liben-Nowell and Kleinberg, 2003). 링크 예측은 두 노드 사이의 미래 연결 가능성을 결정하기 위해 기존 방법을 사용하며, 이러한 방법은 그래프 기반 측정 또는 컨텐츠 기반 측정에 따라 결정된다.  그래프 기반 조치는 네트워크의 위상적 구조를 사용하는 가장 기본적이고 사소한 접근법이다. 이 접근법에서는 각 에지의 점수(x,y)를 x와 y 사이의 최단 경로를 기준으로 할당하고 순위를 정한다.  콘텐츠 기반 측정은 정점과 가장자리의 속성을 채택하는 경우가 많다.  예를 들어, 공동저작권 네트워크에서, 노드는 저자인 반면, 가장자리는 저자의 공동저자(저자의 공동저자)이다.  연구자들은 다양한 영역에서 더 높은 예측 성능을 달성하기 위해 그래프 기반과 콘텐츠 기반 수단을 모두 활용하는 몇 가지 뚜렷한 접근법을 적극적으로 개발하고 있다.  서지학, 생물학, 단백질-단백질 상호작용 및 소셜 네트워크와 같은 영역은 미래 연계를 예측하는 데 있어 높은 효율성을 받는다.  소셜 네트워크에는 대규모 소셜 네트워크 데이터를 생성하는 수백만 명의 활성 사용자가 포함된다(Kemp, 2018; Razak et al., 2020).  그러나 소셜 네트워크에서 데이터의 진실성은 데이터 준비 단계에서 예를 들어 불완전하거나 누락된 데이터로 이어진다(Aminzadeh et al., 2015; Liaqat et al., 2017).  이러한 현상으로 인해 소셜 네트워크의 예측 링크 성능에 영향을 미치는 노이즈와 신뢰할 수 없는 데이터 집합이 발생하였다(Liu 등, 2013).  이 문제를 해결하려면 링크를 예측할 수 있는 고품질의 소셜 네트워크 데이터를 얻기 위한 적절한 필터링 및 사전 처리 단계를 포함하는 것이 좋다. 다음 이슈는 머신러닝 분류기를 사용하는 링크 예측에서 자주 발생하는 데이터셋의 불균형이다.  머신러닝 모델의 분류에는 데이터 집합을 교육 세트와 테스트 세트로 나누는 절차가 포함된다.  그러나 소셜 네트워크의 높은 첨단은 분류를 위해 제공되는 훈련 데이터세트의 불균형을 초래하고 결국 예측 과제를 잘못 이끌 수 있다(Moradabadi와 Meybodi, 2018).  따라서, 우리는 미래의 연구가 표준 기계 학습 분류기에 맞게 훈련과 시험 세트를 모두 수정하는 데 집중할 것을 권고한다.  링크 예측을 위한 데이터 집합에 관한 최종 이슈는 제안된 링크 예측 애플리케이션의 평가를 위해 제공되는 데이터 집합이 불충분하다.  분명히 여러 작품이 데이터 집합의 크기를 간과했다(Zhao et al., 2016).  제안된 애플리케이션이 실제 소셜 네트워크 애플리케이션이 아닌 소규모 데이터 집합에서 테스트될 경우, 잘못된 평가 문제로 이어진다.  따라서 이 문제를 극복하기 위해 링크 예측 연구자들에게 데이터 수집 과정을 좀 더 철저하게 문서화하여 다른 연구자들에 의해 쉽게 복제될 수 있도록 할 것을 요청한다.  또한 본 논문에서 검토한 연구의 대부분은 실제로 소셜 네트워크에 이미지, 비디오, 소리 등의 다른 유형의 데이터가 있을 때 텍스트 데이터 유형의 링크만 사용했다.  따라서, 향후 링크 예측 애플리케이션에서 이러한 다양한 유형의 데이터를 고려할 것을 제안한다. |
| 논문 의의 | 최신 접근법과 응용프로그램을 포함한 새로운 연계 예측 조사를 제시한다. 이 논문은 소셜 네트워크의 링크 예측 분석에 대한 추가 개선을 위한 권고사항과 새로운 통찰력을 제공한다.  또한, 이 조사는 소셜 네트워크에 특정한 초점을 두고 서로 다른 맥락에서 링크 예측 애플리케이션을 설명하려는 첫 번째 시도를 나타낸다.  연구결과는 소셜 네트워크에서의 예측 접근과 응용을 연계하는 것이 상당히 어려운 것임을 밝혀냈다. |
| 논문의 한계점 및 기타 | 소셜 네트워크의 기존 링크 예측 분석에 대해서는, 대규모 소셜 네트워크 데이터를 효율적으로 처리할 수 있는 보다 확장 가능한 접근법을 도입하는 등, 개선의 여지가 남아 있다.  링크 예측 분석 접근법도 끊임없이 진화하고 있는 소셜 네트워크의 토폴로지를 고려해야 한다.  소셜 네트워크 데이터에서 시간변동 정보를 유추하기 위해서는 매우 효율적인 방법이 필요하다.  마지막으로, 여러 소셜 네트워크에 걸쳐 링크 예측을 수행하기 위한 더 나은 접근방식이 필요하다. |