영화 등장인물의 사회관계망에서 중요도를 기반으로 하는 주연 등장인물 검출 기법

허주성, 서장원, 김태형, 이예영, 한연희*
한국기술교육대학교 컴퓨터공학부
e-mail:{chil1207, mklmkl2001, matthew409, tripley94, yhhan}@koreatech.ac.kr

Leading Characters Determination Based on Centrality in Movie Characters' Social Networks

Jooseong Heo, Jangwon Seo ,Taehyeong Kim , Yeyoung Lee, Youn-Hee Han*
School of Computer Science and Engineering
Korea University of Technology and Education

요 약

'영화 속에 등장하는 주연들은 어떤 기준으로 선정되는가'에서 본 논문에서는 두 가지 방법을 활용하여 주연들을 추출해보았다. 그 결과 가중치 연결 중심도를 이용한 검출 방법이 공식적인 주연급 등장인물과 일치한다는 것을 도출해냄.

1. 서론

전통적인 사회관계망 분석에서는 각 노드들이 사회관계망에서 어느 정도의 중요성을 나타내는 지를 나타내는 중심도 (Centrality) 지표가 제시되어 왔다 [1][2]. 한편, 최근 영화 시나리오를 기반으로 영화에 등장하는 등장인물 간의 상호관계를 활용한 사회관계망을 구성하고, 전통적인 사회관계망 분석 기법을 활용하여 영화의 여러 가지 요소를 분석하는 연구가 진행 중이다 [3][4]. 이러한 연구들에서는 영화 내 주요 등장인물 (Leading Characters) 및 등장인물 커뮤니티(Community)의 추출과 영화 내용 전개상 구분 (Story Segmentation) 도출 등의 작업을 해왔지만, 영화의 사회관계망 지표 중 중심도의 역할에 대한 비교평가에 대한 연구는 미흡한 편이다.

기존 연구 [4]에서는 가중치 연결 중심도(Weighted Degree Centrality)를 기반으로 영화 내의 주연 등장인물(Leading Characters 또는 Leading Roles)과 조연 등장인물을 구별하는 방법을 연구하였다. 본 논문에서는 이러한 기존 연구를 확장하여 가중치 연결 중심도 뿐만 아니라 가중치 매개 중심도(Weighted Betweenness Centrality)까지 활용하여 주연 및 조연 등장인물을 구분하는 사회관계망 분석 기법을 소개한다. 또한, 국내에 상영된 XX편의 국내 영화들에 대하여 제안하는 분석 기법을 적용해 가중치 연결 중심도가 가중치 매개 중심도 보다 주연 등장인물을 구분해 내는 데 활용가치가 높음을 보인다.

2. 영화 시나리오 기반 사회관계망 RoleNet 구성

본 논문에서는 영화 시나리오의 각 장면(Scene)별 등장인물을 도출하여 동일한 장면에 함께 등장한 등장인물 간에 관계성이 있다고 판단하고 사회관계망을 구성하였다. 본 논문에서 구성하는 사회관계망 구성법은 [4]에서 제시된 바와 동일하며 RoleNet이라는 용어로 아래와 같이 정의한다.

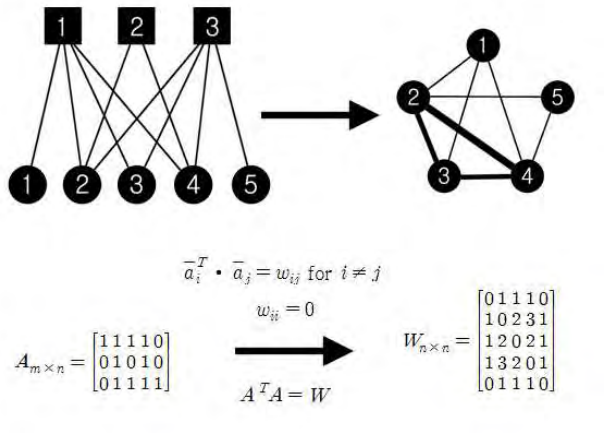
정의: RoleNet은 $G=<V,E,W>$로 표현되는 무방향 가중치 그래프이다. $V=\{v_1,v_2,...,v_3\}$는 등장인물들의 집합이며, 임의의 등장인물 v_i와 v_j 사이에 관계가 있을 때 그 두 개의 등장인물 사이에는 간선 (e_{ij})이 존재하며 그러한 간선들의 집합이 E이다. 마지막으로 W는 그러한 간선에 할당된 가중치(w_{ij})들의 집합이다. ■

위 정의에서 등장인물 간의 관계는 동일한 장면에 등장하는 것으로 정의되며, 관계가 존재하는 두 등장인물 간의 가중치는 동일한 장면에 등장하는 빈도를 기반으로 0과 1 사이 값으로 정규화하여 계산된다.

임의의 영화 시나리오에 총 m개의 장면 (s_1, s_2, ..., s_m)과 n명의 등장인물이 존재한다고 가정할 때, 해당 영화에 대하여 a_{ij}는 등장인물 v_j가 장면 s_i에 등장하면 1 값을 지니고 그렇지 않으면 0을 지닌다고 하자. 그러한 a_{ij}값과 함께 행렬 $A=[a_{ij}]_{m\times n}$를 구성할 수 있고, 또 다른 행렬 $W=[w_{ij}]_{n\times n}$은 다음과 같이 구할 수 있다.

$$W=\sum_{k=1}^{m} a_{ki}a_{kj}=\bar{a}_i^T\cdot\bar{a}_j=A^T\cdot A \qquad (1)$$

위 식에서 $\overline{a_i}=(a_{1,}a_{2,}\cdots,a_{m_i})$는 행렬 A의 i번째 열벡터(Column Vector)이다. 그러면, 행렬 W는 등장인물간의 동일 장면 등장 빈도 정보를 지니게 된다. 그림 1은 3명의 장면(사각형)과 5개의 등장인물(원)이 있는 영화에 대한 행렬 A 및 W의 예시를 보여준다. 이러한 행렬 W는 RoleNet에 포함된 각 등장인물 사이의 간선을 형성할 때 할당할 정규 가중치를 구할 때 활용된다.



(그림 1) 영화 등장인물에 대한 사회관계망 구성

3. 주연 검출 기법

주연 등장인물을 검출하기 위해서 먼저 각 등장인물의 사회관계망에서의 중심도(Centrality)를 구할 필요가 있다. 전통적으로 사회관계망 분석기법에서 중심도는 여러 가지가 있으나 본 논문에서는 가중치 연결 중심도(Weighted Degree Centrality)와 가중치 매개 중심도(Weighted Betweenness Centrality)를 이용한다.

임의의 영화에서 등장인물 v_i에 대한 가중치 연결 중심도 D_i는 다음과 같다.

$$D_i=\frac{\sum_{j\in N(i)} w_{ij}}{|V|-1} \tag{2}$$

위 식에서 알 수 있듯이 임의의 등장인물과 관계를 맺고 있는 각 등장인물 사이의 가중치 값이 높을수록 해당 등장인물의 가중치 연결 중심도는 증가한다. 보통 주연 등장인물들이 다른 주연 및 조연 등장인물들과 동일 장면에 함께 등장하는 경우가 많으므로 대체로 주연 등장인물들의 가중치 연결 중심도가 높을 것이다.

한편, n명의 등장인물을 지닌 임의의 영화에서 등장인물 v_i에 대한 가중치 매개 중심도 B_i는 다음과 같다.

$$B_i=\frac{\sum_{v_s\neq v_i\neq v_t\in V, s<t}\frac{\sigma_{st}(v_i)}{\sigma_{st}}}{(n-1)(n-2)/2} \tag{3}$$

위 식에서 σ_{st}는 임의의 등장인물 v_s와 v_t 사이의 가중치 기반 최단 경로의 수이며, $\sigma_{st}(v_i)$는 임의의 등장인물 v_s와 v_t 사이의 가중치 기반 최단 경로 중 v_i를 지나는 경로의 수이다. 임의의 두 등장인물 사이의 가중치가 높다는 것은 두 등장인물이 더욱 밀접한 연관이 있다는 것을 의미하기 때문에, 위 식의 값을 구할 때 각 간선의 가중치는 주어진 간선의 역수를 활용하였다.

가중치 연결 중심도와 가중치 매개 중심도 모두를 대표하는 등장인물 v_i의 중심도를 C_i라고 가정할 때, 집합 Θ_1과 Θ_0을 다음과 같이 정의하자.

$$\Theta_1=\{C_i\,|\,l_i=1\} \tag{4}$$

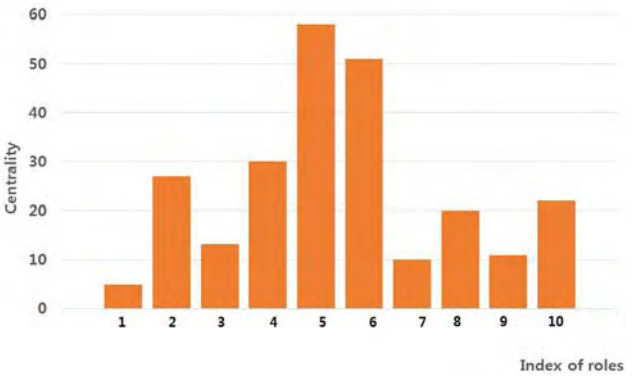
$$\Theta_0=\{C_i\,|\,l_i=0\} \tag{5}$$

위 식에서 l_i는 등장인물 v_i가 주연 등장인물로 판단될 때 1이며, 그렇지 않으면 0이다. 또한, Γ를 $\Gamma=\{l_i\,i=1,2,\ldots,n\}$로 정의할 때 주연 등장인물 검출 문제는 수학적으로 다음과 같이 표현된다 [4].

$$\Gamma^*=\text{argmax}_\Gamma(\min\Theta_1-\max\Theta_0) \tag{4}$$

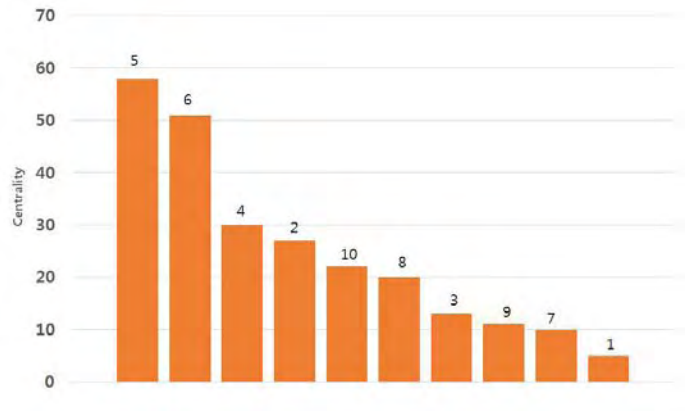
위 식에서, $(\min\Theta_1-\max\Theta_0)$의 의미는 주연 등장인물의 중심도 중 가장 작은 값과 조연 등장인물의 중심도 중 가장 큰 값의 차이다. 그 차이 값이 가장 크도록 만드는 Γ가 이 문제에서 찾고자 하는 Γ^*이다.

Γ^*을 찾기 위해 다음과 같은 기법을 사용한다. 첫 번째로 각 등장인물에 대한 중심도 값을 계산하고 그 값들의 내림차순으로 등장인물 들을 정렬한다. 그 다음 각 인접 등장인물들의 중심도 간의 차이를 계산한 뒤 그러한 차이 값에서 가장 큰 값을 찾고 이것을 기준으로 왼쪽에 위치한 등장인물 들이 주연 등장인물이 된다. 다음은 그에 대한 예시 그래프이다.

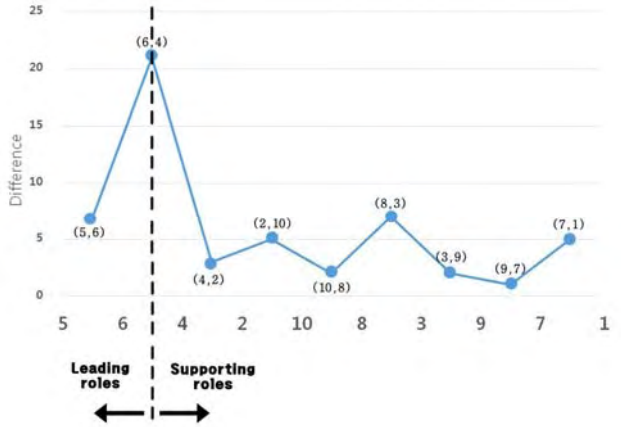


(그림 2) 각각 등장인물들의 중심도 값 분포

(그림 2)는 10명의 등장인물과 그에 대한 중심도 값을 나타낸다. 그림에서 알 수 있듯이 등장인물 5, 6의 중심도가 상대적으로 높은 것을 알 수 있다. 이를 명확하게 하기 위해 (그림 3)처럼 중심도 값을 기준으로 내림차순 정렬하여 배치한다.



(그림 3) 중심도를 내림차순으로 정렬하여 재배치한 그래프



(그림 4) 인접 중심도의 차이로 주연 등장인물과 조연 등장인물의 구분

(그림 3)에서 각 인접 중심도 값들의 차이를 계산하면 (그림 4)와 같이 (5,6)=7, (6,4)=21, (4,2)=3 등이 되며, 등장인물 6과 등장인물 4 사이의 중심도 값 차이가 가장 큰 것을 알 수 있다. 이로부터 등장인물 5와 6이 주연 등장인물이고, 나머지는 조연 등장인물로 분류한다.

4. 실험 평가

본 장에서는 1980년도부터 2015년도까지 국내에 개봉된 약 100여편의 국내 영화에 대하여 이전 장에서 설명한 주연 등장인물 검출 기법을 적용한 결과를 기술한다. 주연 등장인물 검출 기법을 적용할 때에 가중치 연결 중심도 및 가중치 매개 중심도로 나누어 실험하여 두 가지 중심도 중 어떠한 중심도가 주연 등장인물 검출에 더욱 유리한지를 판단한다. 분석 도구는 Python 2.7의 NetworkX 1.9.1 모듈[5]을 활용하였다. NetworkX 모듈은 그래프 객체에 대해 각종 사회관계망 분석 API를 제공하고 있어서 본 논문에서 제시하는 두 가지 척도를 계산하는 데에 매우 적합하다. 한편, 각 영화에 대한 실제 주연 등장인물 데이터는 Naver 영화 정보 사이트[6]에서 획득한 것을 활용하였다.

기법 / 영화제목	가중치 연결 중심도(D)	가중치 매개 중심도(B)	Naver
7번방의 선물	용구, 예승, 방장, 춘호, 만범, 봉식	용구 예승 민환	용구, 예승 방장, 춘호 만범, 봉식 서노인
살인의 추억	두만, 태윤	두만, 태윤	두만, 태윤
반창꼬	미수, 강일 용수, 하윤 반장, 현경	미수, 강일 용수, 간호사	미수, 강일
박쥐	상현, 태주	상현	상현, 태주
국가대표	방코치, 밥 홍철, 수연 재복, 봉구 칠구	밥	방코치, 밥 홍철, 수연 재복, 봉구 칠구

(표 1) 각각의 중심도와 Naver 주연배우 비교 표

임의의 영화 시나리오의 총 등장인물의 수가 n이라고 가정하고, l_i^N는 해당 영화의 임의의 등장인물 v_i에 대해 Naver 영화 정보 사이트에서 주연 등장인물이라고 지칭하고 있으면 1을 지니고, 그렇지 않으면 0을 지니는 파라미터이다. 이 때, 가중치 연결 중심도를 기반으로 하는 주연 등장인물 검출 기법에 대한 성능 R_D는 다음 수식과 같다.

$$R_D = \frac{n - \sum_{i=1}^{n} |l_i^N - l_i^D|}{n} \quad (5)$$

위 식에서 l_i^D는 등장인물 v_i가 가중치 연결 중심도를 기반으로 검출한 주연 등장인물이면 1을 이고, 그렇지 않으면 0이다. 또한, 가중치 매개 중심도를 기반으로 하는 주연 등장인물 검출 기법에 대한 성능 R_B는 다음 수식과 같다.

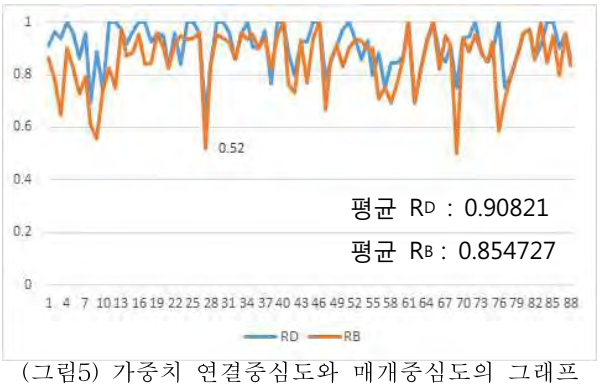
$$R_B = \frac{n - \sum_{i=1}^{n} |l_i^N - l_i^B|}{n} \quad (6)$$

마찬가지로, 위 식에서 l_i^B는 등장인물 v_i가 가중치 매개 중심도를 기반으로 검출한 주연 등장인물이면 1이고, 그렇지 않으면 0이다. R_D와 R_B 모두 0과 1 사이의 값을 지니며, 1에 가까울수록 해당 중심도를 기반으로 하는 주연 등장인물 검출 성능이 좋다고 볼 수 있다.

5. 결론

본 논문에서는 가중치 연결 중심도와 가중치 매개 중심도를 이용하여 영화 내 주연 등장인물 검출 방법에 대해 소개하고 국내 영화들에 대해 적용 및 실험 평가하였다.

그 결과 가중치 연결중심도가 가중치 매개중심도 보다 영화 내 주연 등장인물 검출 신뢰도가 높은 것을 알 수 있었다.



(그림5) 가중치 연결중심도와 매개중심도의 그래프

(그림 5)에서는 Naver 영화 정보 사이트의 주연 배우를 기준으로 한 가중치 연결 중심도와 매개중심도의 값들의 그래프이다. 평균값이 Rd = 0.90821, Rb = 0.854727로 가중치 연결중심도의 검출 성능이 더 좋다는 것을 알 수 있다. 가중치 매개중심도가 가중치 연결중심도 보다 주연 등장인물 검출 확률이 더 낮은 이유는 대게 영화 속에서 주연이 핵심 역할을 맡고 자주 등장하기 때문에 가중치 연결 중심도가 높다. 하지만 조연 또한 경우에 따라 영화 내에서 가중치 매개중심도가 높게 나타날 수 있기 때문에 가중치 연결 중심도보다 조연을 검출할 확률이 높은 것을 생각해 볼 수 있다.

간혹 값들이 크게 차이가 나는 것들이 있는데 이것은 해당 영화의 특수성이 반영된 것으로 주연 배우가 10여명 이상으로 아주 많거나 주연 배우의 역할의 특수성 때문에 오류가 나타남을 생각해 볼 수 있다.

향후에는 주연 배우 검출 성능을 높이는 방법을 연구하면서 본 논문에서 활용한 이론을 바탕으로 영화 등장인물 사회관계망과 영화 흥행도와의 특정 값, 요소 등의 관계 또한 도출해낼 계획이다.

참고문헌

[1] Lei Tang and Huan Liu, "Community Detection and Mining in Social Media," Synthesis Lectures on Data Mining and Knowledge Discovery, Vol.2, No.1, 2010.

[2] 신수진, 김용환, 김찬명, 한연희, "사회관계망에서 매개중심도 추정을 위한 효율적인 알고리즘," 정보처리학회논문지, Vol.4, No.1, pp.37-44, 2015.

[3] S Gil, L Kuenzel, S Caroline, "Extraction and Analysis of Character Interaction Networks from Plays and Movies," Standford, 2011

[4] Chung-Yi Weng, Wei-Ta Chu, and Ja-Ling Wu, "Rolenet: Movie Analysis from The Perspective of Social Networks," IEEE Transactions on Multimedia, vol. 11, No. 2, pp. 256 - 271, 2009.

[5] Python NetworkX, https://networkx.github.io

[6] Naver 영화 정보, http://movie.naver.com