# 영화 등장인물의 사회관계망에서 시나리오를 기반으로 하는 영화 추천 기법

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

# Recommand Movie Based on Scenario in Movie Characters' Social Networks

Joo-Seong Heo, Tae-Hyeong Kim, Jang-Won Seo , Ye-Young Lee,
Youn-Hee Han*
School of Computer Science and Engineering
Korea University of Technology and Education

요 약

'영화 시나리오를 기반으로 영화를 어떻게 추천할 수 있는가'에서 본 논문에서는 전통적인 사회관계망 분석 지표 중 그래프의 평균 길이와 평균 군집도 그리고 밀도를 이용하여 3차원의 데이터 집합을 산출했고, 산출한 데이터 집합을 기반으로 k-means 군집화 알고리즘을 활용하여 각 k 값에 따른 영화를 추천해보았다. 그 결과 기타 여느 추천들과 다른 추천결과를 도출해냈다.

## 1. 서론

전통적인 사회관계망 분석 지표 중에서는 각 노드들과 엣지들 등의 관계에서 어떤 값을 나타내는 지를 나타내는 여러 지표들이 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 연구들을 바탕으로 지난 춘계학술대회에 영화 등장인물의 사회관계망에서 주연 등장인물을 검출하는 주연검출 기법에 대한 논문을 썼다 [3].

기존 연구 [3]에서는 사회관계망 지표 중 중심도라는 것을 기반으로 연구하였다. 본 논문에서는 이러한 기존 연구를 확장하여 평균 길이(Average Path Length), 평균 군집도(Average Clustering), 밀도(Density)에 대한 사회관계망 분석 기법을 소개한다. 또한, 국내에 상영된 약 130여 편의 국내 영화들에 대하여 제안하는 분석 기법과 함께 k-means 군집화 알고리즘을 적용해 시나리오 구조를 기반으로 영화 추천을 한다. 기존 장르 및 평점에 의한 추천이 아닌 시나리오만으로 추천을 하는데 있어 활용 가치가 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 $v_i$와 $v_j$ 사이에 관계가 있을 때 그 두 개의 등장인물 사이에는 간선 $(e_{ij})$이 존재하며 그러한 간선들의 집합이 $E$이다. 마지막으로 $W$는 그러한 간선에 할당된 가중치$(w_{ij})$들의 집합이다. ■

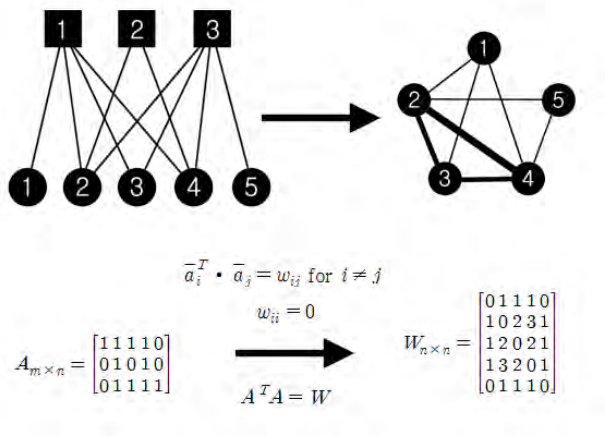
위 정의에서 등장인물 간의 관계는 동일한 장면에 등장하는 것으로 정의되며, 관계가 존재하는 두 등장인물 간의 가중치는 동일한 장면에 등장하는 빈도를 기반으로 0과 1 사이 값으로 정규화하여 계산된다.

임의의 영화 시나리오에 총 $m$개의 장면 ($s_1$, $s_2$, ..., $s_m$)과 $n$명의 등장인물이 존재한다고 가정할 때, 해당 영화에 대하여 $a_{ij}$는 등장인물 $v_j$가 장면 $s_i$에 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.

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

* 교신 저자 : 한연희

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




(그림 1) 영화 등장인물에 대한 사회관계망 구성

## 3. 시나리오 기반 데이터집합 구성

시나리오를 추천하기 위해서 먼저 각 영화의 사회관계망에서의 여러 지표들을 토대로 데이터집합을 구성할 필요가 있다. 전통적으로 사회관계망 분석기법에서 지표는 여러 가지가 있으나 본 논문에서는 사회관계망의 평균 경로 길이(path length), 군집도(clustering), 밀도(density)를 이용한다.

임의의 영화에서 등장인물 n에 대한 평균 경로길이 $D_i$는 다음과 같다.

$$P=\sum_{s,t\in V}\frac{d(s,t)}{n(n-1)} \tag{2}$$

위 식에서 알 수 있듯이 그래프의 등장인물들 간의 관계에서 평균 경로 길이만큼 넘어가면 서로 알 수 있다는 것을 의미한다.

또 한편, 임의의 등장인물에 대한 군집도는 다음과 같다.

$$C_u=\frac{2T(u)}{\deg(u)(\deg(u)-1)} \tag{3}$$

위 식에서 u는 주변 등장인물들과의 삼각관계가 성립하는 부분을 말한다. T(u)는 삼각관계의 개수, deg(u)는 u의 연결정도(degree)값을 의미한다. 따라서 그래프 전체의 군집도는 각 등장인물들의 값의 평균과 같다.

$$C=\frac{1}{n}\sum_{v\in G}c_v \tag{4}$$

군집도가 높을수록 등장인물들의 관계가 더욱 뭉쳐 있다는 것을 알 수 있다.

마지막으로 임의의 등장인물 n, 등장인물들 사이의 간선 m에 대한 그래프 밀도는 다음과 같다.

$$D=\frac{2m}{n(n-1)} \tag{5}$$

밀도가 높을수록 등장인물들이 서로 관계가 있을 확률이 높음을 알 수 있다.

## 4. 시나리오 기반 추천 기법

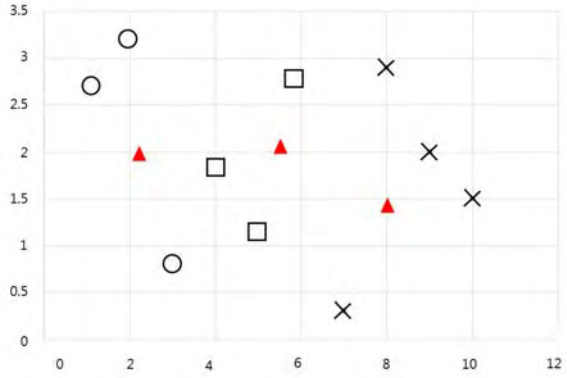
앞서 시나리오에서 구한 세 가지의 값들을 통해 우린 3차원의 데이터 오브젝트 집합 S를 산출했다. 이제 k-means 군집화 알고리즘을 이용해 영화를 그룹화 하는 것이 주어진 문제의 최종 목표이다.

데이터 집합 S에 대하여 k-means 알고리즘이 목적으로 하는 수식은 다음과 같다.

$$\underset{s_1\cdots s_k}{\arg min}=\sum_{i=1}^{k}\sum_{x\in S_i}\|x-u_i\|^2 \tag{6}$$

위 식에서 k는 군집의 개수이며 미리 주어진다. $u_i$는 데이터 집합의 중심점이다. 그러면 군집정도는 주어진 k의 개수에 각 데이터 집합들 간의 중심점의 평균이 가장 최소값인 것이다.

하지만 k-means 군집화 알고리즘은 감독 학습기법으로 중심점과 k를 직접 정해줘야만 한다. 따라서 중심점은 전체 약 100여개의 영화들을 관객 수를 기반으로 정렬한 뒤 k로 나눈 수치를 이용하였다.

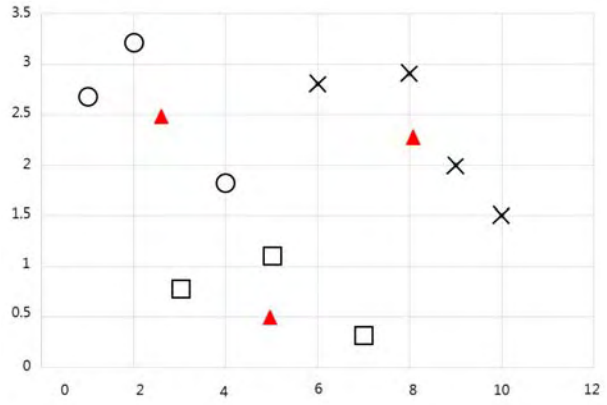



(그림 2) k=3 일 때의 분류 결과

(▲: 중심점, O, □, X : k가 각각 1,2,3인 군집의 데이터)

위 (그림2) 를 보면 k를 3으로 정하고, 10개의 데이터를 3등분 한 뒤 초기 중심점을 위와 같은 빨간 삼각형으로 정한 뒤 세 개의 군집을 분류한 결과이다.

이후 각 집합별 중심점과 집합 내 영화들의 거리의 제곱합이 최소로하는 새로운 집합을 찾는 것이 k-means 군집화 알고리즘이다. 계속된 반복으로 군집 중심을 재조정 한 뒤 더 이상 군집이 변하지 않을 때 추천은 끝이 난다.



(그림 3) 분류가 끝난 뒤의 데이터 집합
(▲: 중심점, O, □, X : k가 각각 1,2,3인 군집의 데이터)

(그림 3)은 분류가 끝난 뒤의 영화들의 군집 결과를 보여준다. (그림 2) 와 비교 했을 때 중심점이 변한 것을 알 수 있다.

| k / 군집 | 5 | 7 | 10 | 12 | 15 |
|---|---|---|---|---|---|
| 1 | 9 | 7 | 10 | 9 | 7 |
| 2 | 9 | 12 | 9 | 9 | 5 |
| 3 | 7 | 15 | 9 | 90 | 9 |
| 4 | 5 | 3 | 20 | 1 | 2 |
| 5 | 99 | 71 | 4 | 1 | 1 |
| 6 | | 12 | 9 | 0 | 2 |
| 7 | | 9 | 8 | 11 | 0 |
| 8 | | | 4 | 1 | 1 |
| 9 | | | 6 | 1 | 3 |
| 10 | | | 51 | 1 | 51 |
| 11 | | | | 4 | 41 |
| 12 | | | | 1 | 5 |
| 13 | | | | | 1 |
| 14 | | | | | 0 |
| 15 | | | | | 1 |

(표 1) 각 k값에 따른 k-means 군집 알고리즘 결과

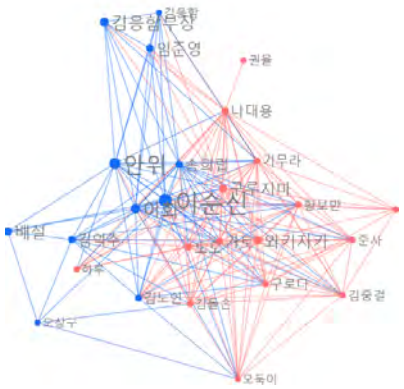
## 5. 실험 결과

본 장에서는 1980년도부터 2015년도까지 국내에 개봉된 약 130여편의 국내 영화에 대하여 이전 장에서 설명한 k-means 군집화 알고리즘을 적용한 결과를 기술한다. k-means 군집화 알고리즘을 적용할 때 초기 중심 값은 영화의 실제 관객 수를 정렬해 k 개수만큼 나눈 데이터 집합에서의 중심으로 정했다. 한편, 실제 관객 수는 한국영화진흥위원회의 정보를 이용했다.

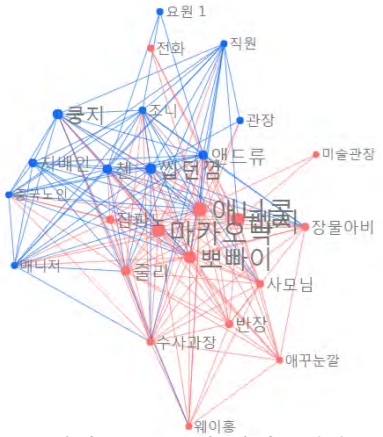
k-means 군집화 알고리즘을 적용하기 위한 도구로는 Python 2.7을 활용하였으며, 데이터 집합을 구성하기 위한 분석 도구는 Python 2.7의 NetworkX 1.9.1 모듈[5]을 활용하였다. NetworkX 모듈은 그래프 객체에 대해 각종 사회관계망 분석 API를 제공하고 있어서 본 논문에서 제시하는 세 가지 척도를 계산하는 데에 매우 적합하다.

실험은 총 k를 5개, 7개, 10개, 12개, 15개로 나누어 총 5번 실험 했으며, 그 결과는 조금씩 모두 차이가 나는 것을 알 수 있었다. 아래 (표 1) 을 보면 개수가 0개 인 것부터 99개인 것까지 차이가 많이 나는 것을 알 수 있다.

주목 할 만한 점은 모든 추천 결과에서 k-means 군집화 알고리즘은 영화 '명량'을 본 관객은 영화 '도둑들'을 함께 보면 좋은 연관된 영화로 추천했다. 영화 '도둑들' 외에도 '명량'과 함께 추천된 영화는 k가 7일 때의 추천 결과를 예로 '베스트셀러', '싸이보그지만 괜찮아', '스승의 은혜', '미술관 옆 동물원', '해안선' 등이 추천 되었다.



(그림 4) 영화 '명량'의 사회관계망 그래프
(pathlength: 1.418, clustering: 0.129, density: 0.591)



(그림 5) 영화 '도둑들'의 사회관계망 그래프
(pathlength: 1.403 clustering: 0.141769, density: 0.603)

(그림 4)와 (그림 5)* 는 2장에서 설명한 'RoleNet'을 구성하고 시각화한 모습으로 영화 '명량'과 '도둑들'의 사회관계망을 구성한 그래프이다. 데이터 집합의 값들이 매우 유사하며 또한, 단순하게 직관적으로만 봤을 때 또한 그래프의 모양이나 색깔 등이 비슷한 것을 알 수 있다. 따라서 두 영화가 함께 추천될 것이라는 것 또한 유추해 볼 수 있다.

## 6. 결론

본 논문에서는 영화 등장인물 사회관계망 그래프의 밀도, 평균 길이, 군집도를 데이터 집합으로 하는 k-means 군집화 알고리즘을 이용하여 시나리오 기반의 영화 추천 기법에 대해 소개하고 국내 영화들에 대해 적용 및 실험하였다 .그 결과 k를 5, 7, 10, 12, 15개로 하였을 때 각각 추천된 값이 다른 것을 알 수 있었다.

결과 값들이 대체로 특정 군집으로 몰려있는 것을 알 수 있었는데 이것은 데이터 집합을 구성하기 위한 값들이 영화별로 크게 차이가 나지 않아 전체적으로 데이터가 몰려 있어서 이런 결과가 나옴을 생각해 볼 수 있다.

실제 영화를 보고 영화의 스토리나 장르 등으로만 보면 전혀 관계가 없어 보이는 두 영화지만 시나리오 구조를 기반으로 분석한 결과는 두 영화가 비슷한 구조의 영화임을 나타내고 있다. 하지만 영화 추천이라는 것이 상당히 주관적이고 명확하지 않은 사실이기 때문에 확실한 정답도 결론도 내릴 수 없는 것도 사실이다.

향후에는 사회관계망 지표 중 좀 더 값들의 편차가 있는, 그렇지만 시나리오 구조를 잘 표현하는 값들을 연구하여 의미 있고 이유 있는 추천 알고리즘을 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] 허주성, 서장원, 김태형, 이예영, 한연희 "영화 등장인물의 사회관계망에서 중요도를 기반으로 하는 주연 등장인물 검출 기법",정보처리학회 춘계학술대회, 2015

[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] 영화진흥위원회, hhttp://www.kofic.or.kr/

* 사회관계망 분석 기법을 활용한 영화 시나리오 분석 웹 서비스 http://movietween.com