트위터에 나타난 미투운동의 키워드 연관성 및 키워드 네트워크 분석

곽수정, 김현희
동덕여자대학교 정보통계학과
e-mail : sujeng21@gmail.com, heekim@dongduk.ac.kr

Analysis of Keyword Association and Keyword Network of #MeToo Movement on Twitter

Soo-Jeong, Kwak*, Hyon Hee Kim*
*Dept. of Statistics and Information Science, Dongduk Womens University

요 약

최근 '미투운동'이 활발히 진행되면서 새로운 페미니즘의 물결을 맞이하였다. 이전의 페미니즘 운동과의 차이점은 SNS 를 통해 익명으로 활동하며 전파속도가 굉장히 빠르다는 것이다. 본 연구는 미투운동의 이러한 특성을 고려하여 실제 트위터 데이터에서 주요 키워드를 파악하고, 해당 키워드의 연관성 및 네트워크 분석으로 사회적 맥락을 알아본다.

1. 서론

'#MeToo'는 미국 할리우드의 유명 영화 제작자 하비 웨인스타인의 성추문 사건 이후 영화배우 알리사 밀라노가 2017 년 10 월 15 일 처음 제안하면서 시작됐다. 해당 운동이 활발해지면서 '미투운동'이란 이름으로 우리나라까지 그 파장이 확산되었고, 권력관계 때문에 불가피하게 성희롱, 성폭행을 당하는 여성들이 SNS 에 '#미투'로 해시태그를 걸고 피해 사실을 고백하여 세상에 알리는 운동으로 자리잡았다.

이전에도 페미니즘에 대한 연구는 계속 되어왔다. 특히 과거부터 끊임없이 이어져 온 페미니즘 운동은 어느 한 나라에 국한되지 않고 특정 사건 또는 계기를 통해 전세계적으로 확산되는 양상을 보인다. 하지만 이전의 확산속도와 달리 현재의 미투운동은 소셜 네트워크의 발달로 눈에 띄게 빠른 전파속도를 보인다.

본 연구에서는 트위터에서 빠르게 확산되는 데이터를 바탕으로 미투운동과 연관된 키워드를 알아보고, 해당 키워드 간의 연관성을 파악하여 현재 미투운동이 가리키는 사회적 맥락을 짚어본다. 이를 위해 연관 규칙을 이용한 키워드간 연관성 분석을 실시하였으며, 키워드 네트워크를 구축하여 영향력이 큰 단어들을 추출하였다.

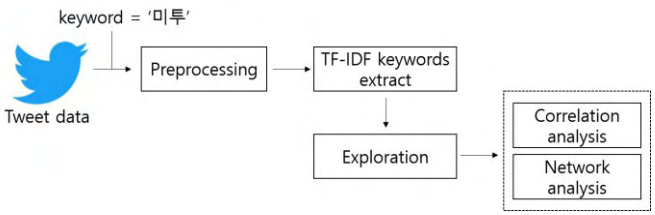
연관 규칙을 이용한 키워드간 분석 결과와 키워드 네트워크 분석 결과는 거의 유사하게 나타났으며, 특히 '미투'는 네트워크 내에서 영향력이 가장 강한 키워드로 다른 키워드들 간의 연관성 또한 높게 나타났다.

지금까지의 페미니즘 운동과 관련한 연구는 이론을 바탕으로 한 사회과학적, 문화적 측면이 대부분이다. 본 연구는 페미니즘 운동을 이러한 국한적인 주제에서 벗어나 소셜 네트워크의 키워드 분석을 통해 접근 방식을 새롭게 넓히고자 한다.

본 논문은 다음과 같이 구성된다. 제 2 장에서는 분석을 위한 실험 설계를 설명하고, 제 3 장에서 키워드 간 연관성 분석 결과를 제시한다. 제 4 장에서 키워드 네트워크 분석 결과를 보여주고, 마지막으로 제 5 장에서 결론을 제시한다.

2. 실험 설계

2018 년 3 월 20 일을 기준으로 국내 트위터에서 '미투' 단어가 포함된 20,000 개의 트윗 데이터를 수집하였다. 그리고 전처리과정을 통해 이름명사와 필요 없는 단어를 제외한 2,992 개의 데이터를 바탕으로 분석을 실행하였다.



(그림 1) 연구 절차

그림 1 과 같이 트위터에서 '미투' 라는 키워드가 들어간 실제 트윗만 수집한 후 전처리 과정을 거친다.

그리고 TF-IDF(Term Frequency - Inverse Document Frequency) 기반으로 워드 클라우드를 통해 미투운동을 중심으로 어떤 키워드가 관련이 있는지 알아보고, 키워드 간의 연관 관계를 시각화하여 보기 쉽게 파악

한다. 마지막으로 키워드 네트워크 분석을 통해 영향력이 높은 키워드를 중심으로 미투운동이 소셜 네트워크 내에서 갖는 맥락을 파악한다.

먼저 '미투'와 연관된 키워드를 알아보기 위해 워드 클라우드를 시행하였다. 전처리과정을 거친 트윗 데이터에서 명사만 추출하여 각 단어에 TF-IDF 가중치를 부여하였다. TF-IDF 는 TF 에서 나타나는 단순 빈도에서 벗어나 단어에 가중치를 부여하여 문서 내에서 얼마나 많은 비중을 차지하는지 나타내기 때문에 보다 정확한 중요도를 파악할 수 있다.

(그림 2) TF-IDF 워드 클라우드

그림 2 는 TermDocumentMatrix(이하 TDM)을 매트릭스 형식으로 변환 후, 행의 합을 계산하고 내림차순으로 정렬하여 생성한 워드 클라우드이다. TF-IDF 를 적용했기 때문에 가중치가 높은 순으로 키워드 빈도수가 정렬되어 주요 단어들을 한 눈에 알 수 있다. 중심 키워드로 가중치가 높은 단어는 '미투'이고 그 뒤로 '미투운동', '공무원', '이유', '피해', '여자', '법정', '피해자' 등이 따른다.

3. 키워드 간 연관성 분석

위에서 알아본 '미투'와 연관된 키워드를 바탕으로 해당 연관성을 파악하기 위해 연관 규칙 알고리즘을 TDM 에 적용하였다. 연관 규칙의 평가 척도로는 지지도(Support), 신뢰도(Confidence), 향상도(Lift)를 사용하였다.

지지도란 전체 트랜잭션 중 그 규칙을 따르고 있는 트랜잭션의 비율을 의미한다. 즉, A 와 B 의 지지도는 전체 규칙 중에서 A 와 B 를 동시에 포함하는 규칙의 비율을 통해 해당 연관 규칙이 얼마나 의미 있는지 확인하는 척도이다.

$$\text{Support}(A \Rightarrow B) = P(A \cap B)$$

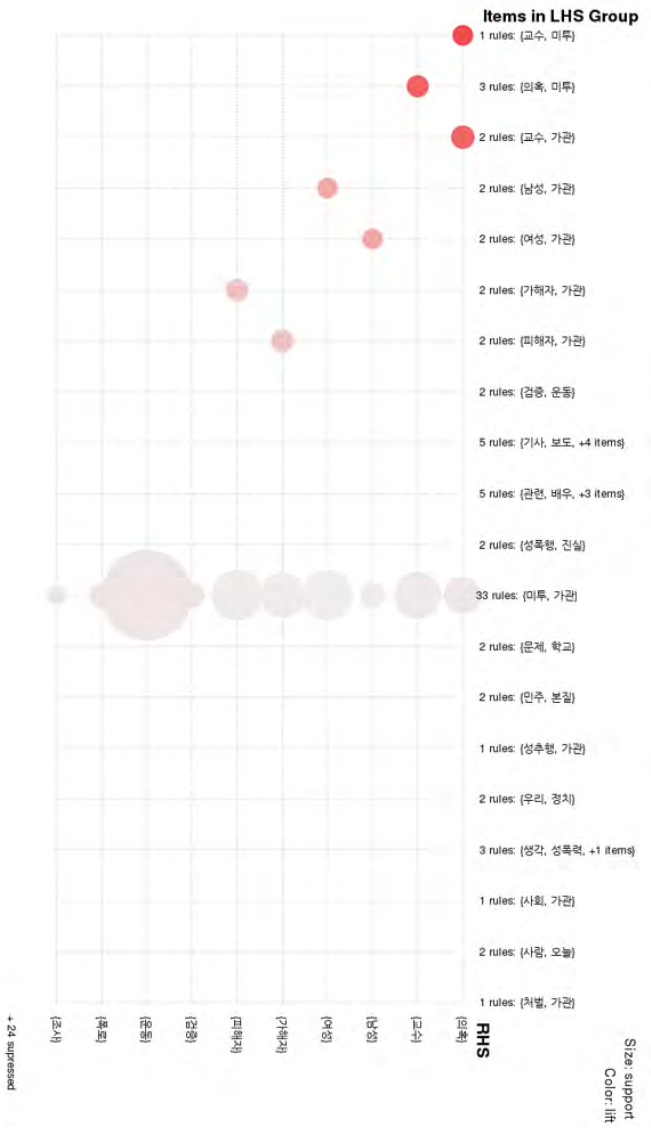
신뢰도는 A 의 트랜잭션 중에서 B 가 포함된 규칙의 비율로 조건부 확률로 정의된다. 즉, 신뢰도는 지지도에 대한 보완적인 연관성 척도이다.

$$\text{Confidence}(A \Rightarrow B) = P(B|A) = P(A \cap B) / P(A)$$

향상도는 A→B 의 연관 규칙에서 임의로(random) B 가 나타나는 경우에 비해 A 와의 관계가 고려되어 나타나는 경우의 비율이다. 즉, 향상도는 연관 규칙이 오른쪽 키워드를 예측하기 위한 능력이 얼마나 향상되었는가를 표현하는 척도이다.

$$\text{Lift}(A \Rightarrow B) = P(B|A) / P(B) = P(A \cap B) / P(A)P(B)$$

이를 바탕으로 빈발항목집합(frequent item sets)*만을 고려하여 연관 규칙을 생성하는 Apriori Algorithm**을 지지도와 신뢰도 모두 0.01 로 하여 TDM 에 적용하였다.



* 최소 지지도 이상을 갖는 항목집합.

** 모든 가능한 항목집합 개수를 줄이는 방식. 모든 항목집합에 대한 지지도를 계산하는 대신에 최소 지지도 이상의 빈발항목집합만을 찾아내서 연관 규칙을 계산한다. 이후 최소 신뢰도를 적용하여 최소 신뢰도에 미달하는 연관 규칙은 제거(pruning)한다.

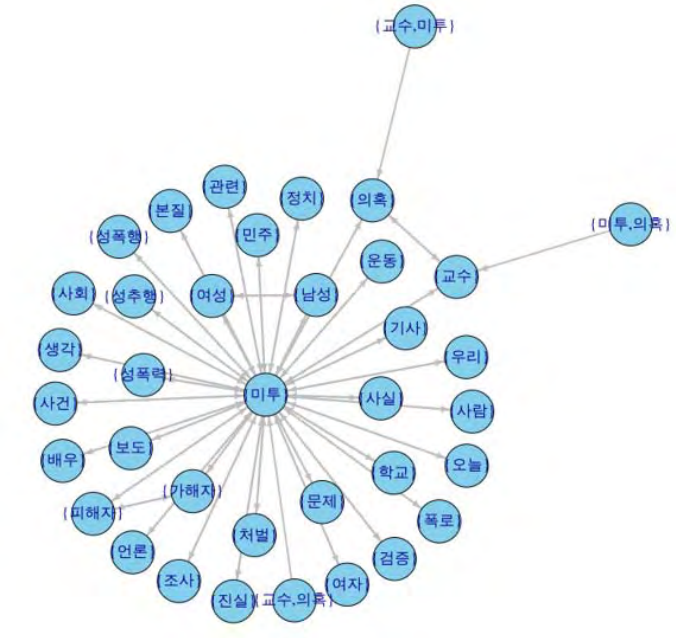
(그림 3) 키워드 연관성 그래프

그림 3 은 TDM 에서 중복 단어를 제거하여 생성한 트랜잭션을 바탕으로 교차표를 작성한 후, 발견된 규칙들을 그래프로 나타낸 것이다. 총 172 개의 규칙들이 발견되었고, 단어가 홀로 등장하는 규칙을 제외한 75 개의 규칙들 중에서 상위 51 개의 규칙들만 시각화 하였다. 그래프에서 원의 크기는 지지도를 의미하고, 색의 진하기는 향상도를 의미한다. 세로축은 규칙의 왼쪽에 위치하는 단어들이고, 가로축은 규칙의 오른쪽에 위치하는 단어들이다.

향상도가 가장 높은 규칙은 {교수, 미투}와 {의혹}이고 이는 '교수'와 '미투'가 언급되면, '의혹'이 등장한다고 해석 할 수 있다. 지지도가 가장 높은 규칙은 {미투, 가관}과 {운동}으로 '미투'와 '가관'이 언급되면,'운동'이 등장한다고 해석된다.

4. 키워드 네트워크 분석

영향력이 높은 단어들을 파악하기 위해서 먼저 그림 4 와 같이 키워드 네트워크를 구축하였다. 여기서 각 노드(node)는 키워드가 되며 노드와 노드는 트랜잭션의 연관 규칙으로 연결되었다.



(그림 4) 키워드 네트워크

키워드 네트워크를 바탕으로 네트워크 내에서 키워드가 가지는 중심성에 대해 연결 정도 중심성(Degree Centrality), 근접 중심성(Closeness Centrality), 위세 중심성(Eigenvector Centrality)으로 분석하였고, 곽기영(2014)에 의해 아래와 같이 정리되었다.

연결 정도 중심성은 네트워크 내에서 각 노드가 다른 노드와 얼마나 많이 인접하고 있는지를 측정하여 중심성이 높은 노드를 파악하는 척도이다. 연결 정도 중심성의 계산 방법은 다음과 같다.

$$C_D(N_i) = \sum_{i=1}^{g} x_{ij}, i \neq j$$

D는 Degree의 첫 알파벳이고 g는 노드의 개수이다. $\sum_{i=1}^{g} x_{ij}$는 노드 i가 (g-1)개의 다른 노드와 갖는 연결관계의 개수이다. 즉, 연결 정도 중심성은 노드의 연결 정도를 자신을 제외한 모든 노드로 나눈 것이다.

근접 중심성은 네트워크 내에서 가장 빠른 경로로 전체 네트워크 내에 다른 노드들까지 접근할 수 있음을 계산한다.

$$C_C(N_i) = \frac{1}{[\sum_{i=1}^{g} d(N_i, N_j)]}, i \neq j$$

노드 i 와 노드 j 간의 최단 경로 거리의 합에 역수를 취해 가장 큰 값을 갖는 노드가 근접 중심성이 높은 노드이다.

위세 중심성은 고유벡터 중심성이라고도 하며, 연결된 노드의 중요성에 가중치를 두고 노드의 중심성을 측정하는 척도이다. 자신의 연결 정도 중심성으로부터 발생하는 영향력과 자신과 연결된 다른 노드의 영향력을 합해 위세 중심성을 결정한다.

$$C_E(i) = \lambda \sum_{i=1}^{N} x_{ij} C_E(j), i \neq j$$

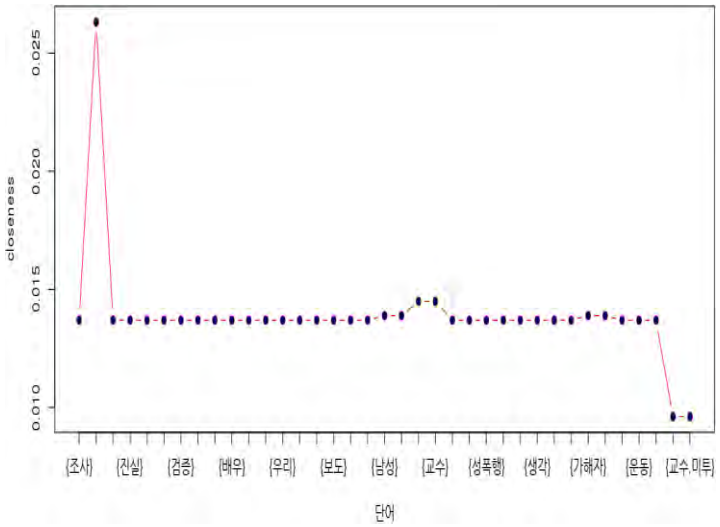
λ는 고유값(eigenvector)이고, x_{ij}는 노드 i와 j간 연결관계의 이진값 또는 계량값이다.

rank	keyword	degree	keyword	closeness	keyword	eigenvector
1	미투	67	미투	0.026315789	미투	1.00000000
2	교수	5	교수	0.014492754	교수	0.20698347
3	의혹	5	의혹	0.014492754	의혹	0.20698347
4	가해자	4	가해자	0.013888889	가해자	0.20517597
5	남성	4	남성	0.013888889	남성	0.20517597
6	여성	4	여성	0.013888889	여성	0.20517597
7	피해자	4	피해자	0.013888889	피해자	0.20517597
8	검증	2	검증	0.013698630	검증	0.17024566
9	관련	2	관련	0.013698630	관련	0.17024566
10	기사	2	기사	0.013698630	기사	0.17024566
11	문제	2	문제	0.013698630	문제	0.17024566
12	민주	2	민주	0.013698630	민주	0.17024566
13	배우	2	배우	0.013698630	배우	0.17024566
14	보도	2	보도	0.013698630	보도	0.17024566
15	본질	2	본질	0.013698630	본질	0.17024566
16	사건	2	사건	0.013698630	사건	0.17024566
17	사람	2	사람	0.013698630	사람	0.17024566
18	사실	2	사실	0.013698630	사실	0.17024566
19	사회	2	사회	0.013698630	사회	0.17024566
20	생각	2	생각	0.013698630	생각	0.17024566

(표 1) 각 중심성 지표의 상위 20 개 키워드

표 1 은 연결 정도 중심성, 근접 중심성, 위세 중심성 분석을 바탕으로 얻은 지표에서 중심성이 높은 상위 20 개의 키워드를 표로 나타낸 것이다. 각 중심성 분석에서 키워드 순서는 동일하게 나타났다. 모두 '미투'가 가장 중심성이 높게 나타났고,'교수'와 '의혹'이 그 뒤를 이었다. 그리고 '가해자', '남성', '여성', '피해자'가 동일한 중심성을 보였다.

아래 그림 5 는 근접 중심성의 지표를 시각화 한 그



래프이다.

(그림 5) 근접 중심성 그래프

그래프 크기 상 단어가 모두 표현되지 못하여 원래 나타나야 하는 단어를 순서대로 정리하자면, {조사}, {미투}, {학교}, {진실}, {본질}, {기사}, {검증}, {폭로}, {오늘}, {배우}, {정치}, {사실}, {우리}, {문제}, {처벌}, {보도}, {여자}, {언론}, {남성}, {여성}, {의혹}, {교수}, {민주}, {관련}, {성폭행}, {사회}, {성추행}, {생각}, {성폭력}, {사건}, {가해자}, {피해자}, {사람}, {운동}, {교수, 의혹}, {미투. 의혹}, {교수, 미투} 이다. 몇 개의 노드들을 제외하고는 거의 비슷한 중심성을 보인다. 그중 근접 중심성이 눈에 띄게 높게 나타나는 단어는 '미투'이다. 키워드 연관 분석 결과와 같이 근접 중심성 결과 또한 '미투'가 다른 단어에 비해 네트워크 내에서 상대적 중요성이 높음을 의미한다. 다음으로 중심성이 높은 단어는 '의혹'과 '교수'이다.

즉, 근접 중심성이 가장 높게 나온 3 개의 단어가 '교수', '의혹', '미투' 임을 미루어 볼 때, 현재의 미투 운동이 각계 분야 중 학계에서 가장 활발히 움직임을 알 수 있다. 상위 3 개의 단어 다음으로 중심성이 높은 단어는 '남성', '여성'과 '가해자', '피해자'가 나란히 이어진다.

5. 결론

최근 활발히 진행되는 미투운동을 트위터를 활용하여 키워드 연관 분석과 네트워크 분석을 통해 알아보았다. 크롤링 날짜를 가장 최근 날짜로부터 20,000 개의 데이터를 수집했기 때문에 최근 이슈화된 사건 키워드를 중심으로 키워드가 연결되어 있음을 알 수 있다.

미투운동의 장이 SNS 이고 익명성이 보장되어 고백한다는 점에서 사건 키워드가 가해자에 초점이 맞춰져 있고, 미투운동이 권력형 성폭력의 고발이라는 성격을 지닌 점에서 의외로 '여성'과 '남성'의 성별 대립 관계는 또렷하지 않았다. 또한, 키워드 연관성 그래프에서 '가관'이라는 단어가 포함되어 나타나는 규칙이 많았는데 이는 미투운동의 사건을 바라보는 부정적인 시각으로 해석할 수 있다. 그리고 워드 클라우드에서는 '공무원'이 '미투운동' 다음으로 빈도수가 높게 나타났지만 키워드 연관 분석과 네트워크 분석에서 상위 키워드로 나오지 않는 걸 보아 단어 사이의 규칙성은 미미하다는 점을 알 수 있다. 또한, 연결 정도 중심성, 근접 중심성, 위세 중심성에서 모두 동일한 순서로 키워드가 정렬되었다. 이는 모든 키워드가 '미투'를 중심으로 분포되어 있으며, '미투'가 다른 키워드들에게 강력한 영향을 끼친다는 것을 알 수 있다. 마지막으로 '본질', '검증', '의혹', '사실' 이라는 단어의 등장은 이전과는 다른 새로운 형태의 페미니즘 운동인 만큼 미투운동이 피해자가 가해자를 지목하여 고발한다는 점에서 사실 확인이 제대로 이루어지는지 검증이 필요하다는 점을 파악할 수 있었다.

아직 현재 진행형인 미투운동의 앞으로 일어나는 사건에 대해 본 연구를 바탕으로 분석한다면 각 사건에 대한 키워드 분석과 함께 시간의 흐름에 따른 키워드 변화도 알 수 있으리라 생각된다. 또한, 본 연구와 관련하여 소셜 네트워크의 감성분석이 부가된다면 더욱 다양한 접근 방식으로 페미니즘 운동을 연구할 수 있을 것으로 기대된다.

참고문헌

[1] **진설아, 허고은, 정유경, 송민 (2013). 트위터 데이터를 이용한 네트워크 기반 토픽 변화 추적 연구. 정보관리학회지, 30(1), 285-302.**

[2] **나영연, 박준건, 문일철 (2017). 다차원 가우시안 프로세스와 시계열 텍스트 데이터 이용한 대통령 후보자 지지율 분석.한국경영과학회 학술대회논문집, 1151-1156.**

[3] **차준범, 성정웅, 김정근, 박상현 (2016). 트위터 데이터에 기반한 헬-조선 키워드 분석. 한국멀티미디어학회 추계학술발표대회논문집**

[4] **곽기영 (2014). 소셜 네트워크분석. 청람.**