키워드 집합 기반 데이터를 위한 효과적인 군집화 알고리즘

이세희 ㅇ, 정연준, 정성원

서강대학교 컴퓨터공학과

jaydenlee@sogang.ac.kr, fuzzy0427@gmail.com, jungsung@sogang.ac.kr

An Effective Clustering Algorithm for Keyword Set-based Data

Sehee Lee ㅇ, Younjoon Chung, Sungwon Jung

Department of Computer Science and Engineering, Sogang University

요 약

최근 이슈가 되고 있는 소셜미디어 데이터는 이미지, 동영상, 가상현실 등 종류가 매우 다양하기 때문에 일차적으로 키워드(태그)를 통해서 군집화하는 것이 매우 효율적이다. 그러나 지금까지의 군집화 알고리즘은 고정된 차원에 기반하기 때문에 차원이 고정되어 있지 않은 소셜미디어 데이터를 효과적으로 군집화 하지 못한다. 따라서 본 논문에서는 키워드의 종류가 무한하고, 차원이 고정 되어있지 않은 키워드 집합 데이터를 처음으로 정의하고 이를 위한 유사도 측정법과 군집화 알고리즘을 제안한다.

1. 서 론

최근 트위터, 인스타그램과 같은 소셜미디어 서비스를 통해서 전 세계적으로 수많은 사용자생산컨텐츠 (UGC)가 생성되고 있고 그 양이 증가하여 소셜미디어 데이터를 분석하는 기술에 대한 연구와 수요가 증가하고 있다. 특히 군집화 기술은 머신러닝, 인공지능 등의 전처리 과정으로써 사건들의 양상, 혹은 인물에 대한 관점, 생각을 분석하는데 매우 중요하다.

소셜미디어 데이터는 이미지, 동영상 등 종류가 매우 다양하기 때문에 일차적으로 키워드(태그)를 기반으로 군집화를 진행하는 것이 효율적이다. 그러나 기존의 범주속성 데이터 군집화 알고리즘[1]은 차원이 고정 되어있고, 각 차원에 올 수 있는 값이 한정되어 있다는 점에서 차원이 고정되어 있지 않은 소셜미디어 데이터를 정확하게 군집화 할 수 없다. 이러한 문제를 해결하고자 여러 연구가 진행되었다[2][3][4]. 그러나 이러한 알고리즘 역시 단 하나의 대표 키워드와 위치정보를 바탕으로 군집화 하거나 백오브워드(bag-of-word) 모델을 사용한다. 하나의 대표 키워드를 이용하면 대표 키워드를 찾기 위해 상당한 시간이 걸리고 대표 키워드를 제외한 다른 키워드는 고려되지 않는다는 단점을 가지고 있다. 또한 백오브워드를 사용하면 키워드가 많아질 수록 차원이 커지면서 정확한 군집화가 불가능해진다는 단점을 가지고 있다. TBKmedoid[4]는 원이 고정되지 않은 데이터를 군집화 할 수 있으나 키워드가 사용된 위치에 따라 가중치를 부여하고 유사도를 계산하기 때문에 키워드 집합 기반의 데이터를 효율적으로 군집화 하지 못한다.

본 논문에서는 이러한 문제를 해결하기 위하여 키워드를 원소로 하는 집합 기반의 데이터를 처음으로 제안하고, 이 집합 데이터간의 유사도 분석기법과 군집화 알고리즘을 제안한다.

2. 키워드 집합 데이터와 데이터 간의 유사도 측정법

키워드 집합 데이터는 여러 개의 키워드를 가질 수 있으며, 이때 키워드의 개수는 고정되어 있지 않다. 또한, 각 차원이 속성을 지니는 기존의 범주 속성 데이터와는 다르게 이 키워드 집합 데이터는 차원이라는 개념이 적용되지 않는다. 예를 들어 서울 여행 중 찍은 사진 하나의 이미지에 대한 기존의 범주 속성 데이터는 (위도, 경도, 시간, 날짜)와 같은 4차원 데이터로 이루어져 있다면, 키워드 집합 데이터는 “비, 경복궁, 서울, 여행, 한국, 한복”과 같이 여러 개의 키워드가 모여 집합을 이루고 있다.

기존의 유사도 측정법들은 이러한 키워드 집합 데이터를 측정하는데 부족한 부분이 있다. 가장 보편적인 유사도(거리)측정 법인 해밍 거리(Hamming distance)를 포함하는 범주 속성 유사도 측정법은 데이터간의 차원에 기반을 두어 유사도를 측정 하므로 집합 데이터 간의 유사도를 정확하게 분석하지 못한다.

2.1 키워드 집합 데이터

키워드 집합 데이터 총 n개의 속성으로 이루어져 있고 n은 고정되어 있지 않다. 하나의 속성은 2차원 벡터로써 1차원은 키워드이고 2차원은 해당 키워드의 빈도수이다. 키워드 집합 데이터를 정의하면 다음과 같다.

정의 1. 키워드 집합 데이터

어떤 키워드 집합 데이터 sd_i는 총 n개의 속성 $A_1, A_2, \dots, A_n$으로 이루어져있다. 속성 A_l은 2차원 벡터로써 키워드 k_1과 키워드의 빈도수 f_1으로 구성되어있다.

예를 들어 “비, 경복궁, 서울, 여행, 한국, 한복” 이라는 키워드를 가지는 집합 데이터 sd는 $n=6$이고 $A_1=(\text{비},1)$, $A_2=(\text{경복궁},1)$, $A_3=(\text{서울},1)$, $A_4=(\text{여행},1)$, $A_5=(\text{한국},1)$, $A_6=(\text{한복},1)$ 총 6개의 속성을 가진다.

2.2 키워드 집합 데이터간의 유사도 측정법

일반적으로 범주 속성 데이터는 서로 같은 값(키워드)을 많이 포함하면 할수록 유사하다고 볼 수 있다. 그러나 키워드 집합 데이터가 포함하는 키워드의 개수가 고정되어있지 않기 때문에 단순하게 같은 키워드를 많이 공유하고 있다고 유사한 것은 아니다. 예를 들어 sd_1와 sd_2 모두 포함하고 있는 키워드가 10개이고 sd_1 또는 sd_2에 속한 키워드가 100개이면, 전체적으로 10%만 같은 키워드를 가지고 있다고 볼 수 있다. 반면 sd_1 또는 sd_2에 속한 키워드가 10개이면, 두 데이터는 100% 유사하다고 볼 수 있다. 키워드 집합 데이터 간의 정확한 유사도를 측정하기 위해서는 두 키워드 집합 데이터 sd_i와 sd_j에 대해서 데이터 sd_i와 sd_j 모두 속한 키워드뿐만 아니라 sd_i와 sd_j에 대하여 적어도 한 쪽에 속하는 키워드의 개수 역시 고려해야한다. 우리는 두 집합 데이터 sd_i와 sd_j 모두 속한 키워드 데이터 집합을 sd_i와 sd_j의 교집합 $I_{ij}=A_1,A_2,\dots,A_k$라고 하고 다음과 같이 정의한다.

정의 2. 키워드 집합 데이터의 교집합

임의의 두 집합 데이터 sd_i와 sd_j에 대하여 sd_i와 sd_j 모두 속하는 태그를 가진 속성들로 이루어진 집합을 sd_i와 sd_j의 교집합이라 하고 I_{ij}로 표시한다. 즉, 두 속성 $A_k \in sd_i$ 와 $A_l \in sd_j$에 대하여

$$A_r=(t_k,f_k+f_l)\in I_{ij},\qquad where\ t_k=t_l$$

이다.

그리고 두 데이터 sd_i와 sd_j중 적어도 어느 한쪽엔 속한 키워드 데이터 집합을 합집합 $U_{ij}=A_1,A_2,\dots,A_m$ 이라 하고 다음과 같이 정의한다.

정의 3. 키워드 집합 데이터의 합집합

임의의 두 집합 데이터 sd_i와 sd_j에 대하여 적어도 sd_i또는 sd_j 한 쪽에 속하는 태그를 가진 속성들로 이루어진 집합을 sd_i와 sd_j의 합집합이라 하고 U_{ij}로 표시한다. 즉, 두 속성 $A_k \in sd_i$와 $A_l \in sd_j$에 대하여

$$A_r=(t_k,f_k+f_l)\in U_{ij},\qquad where\ t_k=t_l$$
$$A_k\in U_{ij}\ and\ A_l\in U_{ij},\qquad where\ t_k\neq t_l$$

이다.

두 키워드 집합 데이터의 교집합과 합집합을 통해서 두개의 집합 데이터 $\mathrm{sd_i}$와 $\mathrm{sd_j}$의 유사도를 측정한다. 두 키워드 집합 데이터 $\mathrm{sd_i}$와 $\mathrm{sd_j}$의 유사도는 다음과 같다.

$$s(sd_i,sd_j)=\frac{\sum_{l=1}^{n(I_{ij})}f_l}{\sum_{l=1}^{n(U_{ij})}f_l}$$

유사도 값의 범위는 $[0,1]$이고, 0이면 교집합이 공집합, 즉 공통된 것이 하나도 없다는 의미이고, 1이면 교집합과 합집합이 같아 두 키워드 집합 데이터는 완벽하게 같다는 의미이다. 예를 들어 $A_1=(\text{비},1)$, $A_2=(\text{경복궁},1)$, $A_3=(\text{서울},1)$, $A_4=(\text{여행},1)$, $A_5=(\text{한국},1)$, $A_6=(\text{한복},1)$ 인 sd_1과 $A_1=(\text{한식},1)$, $A_2=(\text{서울},1)$, $A_3=(\text{여행},1)$, $A_4=(\text{한국},1)$ 인 sd_2의 유사도 $s(sd_1,sd_2)$는 0.6이 된다.

3. 키워드 집합 기반의 데이터 군집화 알고리즘

최근 키워드와 태그를 기반으로 하는 군집화 알고리즘에 대한 연구가 활발하게 진행되고 있다. 하지만 이 알고리즘들은 온전하게 키워드 만으로 군집화하지 않고 있다. 키워드 중 대표 키워드를 추출하여 위치정보를 결합하여 군집화하거나 키워드의 카테고리를 정의한 뒤에 군집화 한다. 이러한 알고리즘 들은 키워드 사용에 한계가 있을 뿐만 아니라 키워드 외의 정보가 없는 경우에는 사용하지 못한다. 따라서 키워드 집합 데이터를 군집화 할 수 있는 알고리즘이 필요하다. 본 섹션에서는 k-mode[1]처럼 반복을 통한 군집화하는 새로운 알고리즘인 RCASK(**R**epresentative **C**lustering **A**lgorithm for **S**et of **K**eyword data)를 제안한다.

3.1 키워드 집합 데이터로 이루어진 군집의 대푯값

키워드 집합 데이터들의 군집을 대표하는 값 역시 집합 기반으로 정의한다. 그래야 앞서 정의한 유사도 기법을 통한 유사도 측정이 가능하다. 따라서 군집의 대푯값을 다음과 같이 정의한다.

정의 4. 키워드 집합 데이터 군집의 대푯값

키워드 집합 데이터들의 군집 $C_i=\{sd_1,sd_2,\dots,sd_n\}$가 있을 때 이 집합 C_i의 대푯값은 $Q_i=\{q_1q_2,\dots,q_m\}$이다. 대푯값의 속성 q_i는 집합 데이터의 속성 A_i와 같이 2차원 벡터로 태그 k_i과 태그의 빈도수 f_i로 구성되어있다. 다만, 대푯값의 속성 q_i는 집합 데이터의 속성과는 다르게 빈도수 f_i가 임계값 t보다 반드시 커야 한다. 즉

$$f_l\geq t,\qquad \forall f_l\ of\ q_l\ in\ Q_i$$

이다.

예를 들어 $t=0.5$ 인 경우 $A_1=(\text{비},1)$, $A_2=(\text{경복궁},1)$, $A_3=(\text{서울},1)$, $A_4=(\text{여행},1)$, $A_5=(\text{한국},1)$, $A_6=(\text{한복},1)$ 인 sd_1 과 $A_1=(\text{한식},1)$, $A_2=(\text{서울},1)$, $A_3=(\text{여행},1)$, $A_4=(\text{한국},1)$ 인 sd_2로 이루어진 군집 C_1의 대푯값 Q_1은 $\{q_1=(\text{서울},0.5),\ q_2=(\text{여행},0.5),\ q_3=(\text{한국},0.5)\}$이 된다.

3.2 RCASK 설명

기존에 연구된 군집화 알고리즘은 키워드의 종류가 다양하고 데이터 하나가 가지는 키워드의 개수가 고정되어있지 않은 키워드 집합 데이터를 효율적으로 군집화하지 못한다. 이번 장에서는 이러한 문제를 해결하기 위하여 효과적인 키워드 집합 데이터 군집화 알고리즘인 RCASK를 제안한다. 기본적인 RCASK의 동작은 k-mode와 같다. 최대 군집의 개수 k와 임계값 t를 입력받은 뒤에 k개의 대푯값은

전체 데이터 중에서 임의로 선택한다. 그리고 각각의 키워드 기반 집합 데이터에 대해서 각 군집의 대푯값과 유사도를 계산하여 유사도가 가장 높은 군집에 데이터를 포함 시킨다. 모든 데이터를 가장 유사한 군집에 포함 시켰으면, 각 군집의 대푯값을 업데이트한다. 만일 대푯값이 공집합인 군집이 있으면 해당 군집은 제거하고 제거한 군집의 개수를 k에서 뺀다. 새롭게 대푯값이 계산되었으면, 새로운 대푯값을 기준으로 각 데이터를 가장 유사한 군집으로 재배열한다. 이 과정은 군집의 대푯값, 군집의 개수, 각 군집이 포함하는 데이터 모두 변화가 없을 때까지 반복한다.

4. 성능평가

RCASK의 성능을 확인하기 위하여 우리는 인스타그램 데이터[5]를 수집하여 사용했다. 총 500,000개의 인스타그램 데이터를 수집 하였고, 모든 데이터는 이미지와 태그로 이루어져있다. 태그란 "#"과 키워드의 결합으로 각 데이터에서 태그를 추출하고 "#"을 지워서 인스타그램 데이터를 키워드 집합 데이터로 변환하였다. 예를 들어 태그 "#비 #경복궁 #서울 #여행 #한국 #한복"을 가지는 데이터를 $A_1 = (\text{비}, 1)$, $A_2 = (\text{경복궁}, 1)$, $A_3 = (\text{서울}, 1)$, $A_4 = (\text{여행}, 1)$, $A_5 = (\text{한국}, 1)$, $A_6 = (\text{한복}, 1)$ 총 6개의 속성을 가지는 집합 데이터로 변환한 뒤에 실험하였다. 성능 측정은 실제데이터에 대한 군집화가 잘 이루어졌는지를 평가하는 척도인 "*Silhouette coefficient*" 를 각 데이터에 대하여 구하고 이 값의 평균을 비교하여 평가한다. 비교 알고리즘은 TBKmedoid[4]로 한다. 실험을 위해서 RCASK의 k는 100, t 는 0.6으로 실험을 진행하였다. 그리고 TBKmedoid는 RCASK 수행 후 나온 군집의 개수를 k 값으로 설정하여 실험하였다. 인스타그램 데이터 500,000개를 군집화 한 결과 *Silhouette coefficient*값은 그림1과 같다.

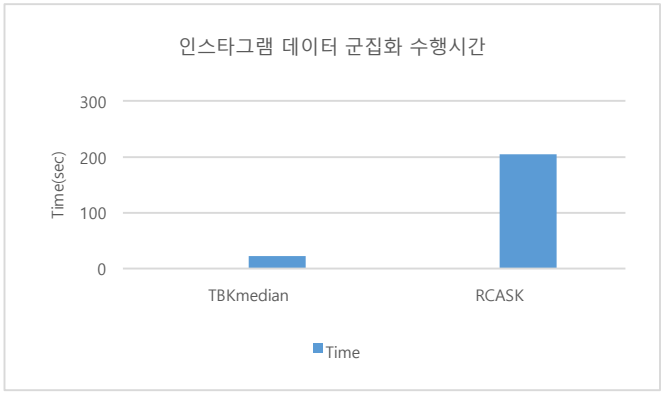

그림 1. 각 알고리즘 별 수행시간

*Silhouette coefficient*는 -1부터 1까지의 값이 나오는데 1에 가까울 수록 데이터가 잘 분리되었다는 의미이다. TBKmedoid는 *Silhouette coefficient*값의 범위가 [0.01, 0.73]이고 평균은 0.31이고 수행 시간은 22.4초였다. RCASK의 실험 결과 범위는 [0.36, 0.74]이고 평균은 0.56이 나왔다. 그러나 수행 시간은 204.2초로 TBKmedoid보단 많은 시간이 소요되었다. 이는 TBKmedoid가 분산 병렬처리를 하기 때문에 차이가 나타났지만 정확도 측면에서는 RCASK가 기존의 군집화 알고리즘 보다 더 정확하게 군집화가 가능하다는 사실을 확인할 수 있었다. RCASK를 분산처리를 통해서 실행하면 TBKmedoid만큼 빠르면서 더 정확한 군집화가 가능할 것으로 보인다.

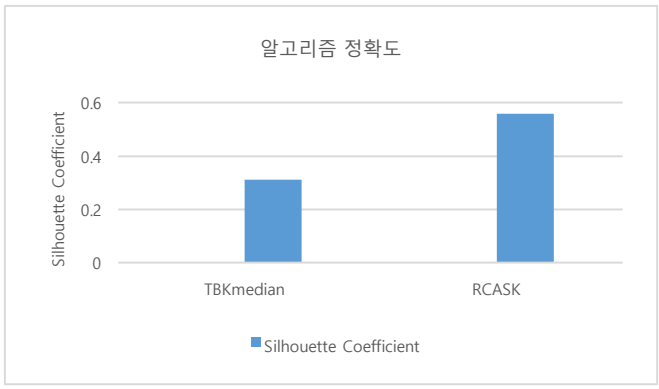

그림 2. 군집화 알고리즘의 정확도

5. 결론

본 논문에서는 기존엔 없었던 소셜미디어를 포함한 키워드 집합 데이터를 군집화하기 위한 유사도 기법과 새로운 알고리즘인 RCASK를 제안하였다. 실험 결과 실제 데이터에서 기존 알고리즘보다 높은 정확도를 보였다. 또한 초기에 정확한 군집의 개수가 필요한 다른 알고리즘과는 다르게 최대값만 알면 된다는 점에서 RCASK는 군집의 개수를 모르는 키워드 집합 데이터를 군집화 하는데 효과적이다. 다만 시간이 많이 걸린다는 문제가 있는데 이는 TBKmedoid와 같이 맵리듀스를 활용하여 시간을 단축하는 후속 연구가 필요하다.

본연구는 2016년도 한국연구재단의 지원을 받아 수행된 이공분야기초연구사업(NRF-2015R1D1A1A01058001)임.

참고문헌

[1] Huang, Z; Ng, MKP, "A fuzzy k-modes algorithm for clustering categorical data", IEEE Transactions on Fuzzy Systems, v. 7 n. 4, p. 446-452, 1999

[2] Takamu Kaneko, Keiji Yanai, "Event photo mining from Twitter using keyword bursts and image clustering", Neurocomputing 172, pp. 143-158, 2016.

[3] Oren Tsur, Adi Littman, Ari Rappoport, "Efficient Clustering of Short Messages into General Domains", ICWSM, 2013

[4] Saeyoung Kim, Jaehwan Lee, Sehee Lee, Sungwon Jung, "A MapReduce-based Clustering Algorithm for Social Media Data with Tags", KIISE, 2015

[5] http://www.instagram.com