碩士學位論文

혼합형 데이터에서 유사도 측정을 통한 군집화 방법

濟州大學校 大學院

컴 퓨 터 工 學 科

宋 炯 岷

2016 年 2 月

혼합형 데이터에서
유사도 측정을 통한 군집화 방법

指導敎授 李 尙 俊

宋 炯 岷

이 論文을 컴퓨터工學 碩士學位 論文으로 提出함

2015 年 12 月

宋炯岷의 工學 碩士學位 論文을 認准함

審査委員長 _____ ㊞

委 員 _____ ㊞

委 員 _____ ㊞

濟州大學校 大學院

2015 年 12 月

목 차

목 차 ····· i
그림목차 ····· iii
표 목 차 ····· iv
국문초록 ····· v
Abstract ····· vi

Ⅰ. 서 론 ····· 1

Ⅱ. 관 련 연 구 ····· 3
1. 협업적 필터링(Collaborative Filtering)과 유사도 측정 기법 ····· 3
1) 협업적 필터링 ····· 3
2) 피어슨 상관계수(Pearson's Correlation Coefficient) ····· 4
3) 코사인 유사도(Cosine Similarity) ····· 4
4) 유클리드 거리(Euclidean Distance) ····· 5
2. Gower 유사도 계수(Gower's Similarity Coefficient) ····· 6
3. k-means 알고리즘 ····· 8
1) 군집화 알고리즘 ····· 8
2) k-means 알고리즘 ····· 9

Ⅲ. 혼합형 데이터에서 유사 사용자 군집화 방법 ····· 10

1. 유사도 계산 ····· 10
2. 군집화 ····· 13

Ⅳ. 실험결과 ····· 14
1. 데이터 수집 및 전처리 ····· 14
2. 군집결과 ····· 18

Ⅴ. 결 론 ····· 32

참 고 문 헌 ····· 33

그 림 목 차

그림 1. k-means Algorithm ····· 9
그림 2. 거주지에 대한 다단계 범주형 데이터 구조도 ····· 10
그림 3. 1차 군집결과 중 군집별 출신학교 ····· 20
그림 4. 1차 군집결과 중 군집별 거주지 ····· 21
그림 5. 1차 군집결과 중 군집별 고향 ····· 22
그림 6. 군집화 완료 후 군집별 출신학교 비교 ····· 28
그림 7. 군집화 완료 후 군집별 거주지 비교 ····· 29
그림 8. 군집화 완료 후 군집별 고향 비교 ····· 30

표 목 차

표 1. 유사도 계산을 위한 데이터 예시 ····· 6
표 2. 사용자 프로필 항목과 데이터 타입 ····· 14
표 3. 스포츠에 대한 선호도 ····· 15
표 4. 전처리 데이터 ····· 16
표 5. 1차 군집 결과 중 15번 사용자를 중심으로 한 군집1의 비교 ····· 18
표 6. 1차 군집화 결과 비교 ····· 19
표 7. 2차 군집화 결과 비교 ····· 23
표 8. 군집중심과 소속 사용자 간의 평균 유사도 ····· 24
표 9. 3차 군집화 결과 비교 ····· 24
표 10. 4차 군집화 결과 ····· 25
표 11. 5차 군집화 결과 ····· 25
표 12. 군집화 완료 결과 비교 ····· 26

국 문 초 록

혼합형 데이터에서 유사도 측정을 통한 군집화 방법

컴퓨터공학과 송 형 민
지 도 교 수 이 상 준

정보기술의 발전에 따른 데이터양의 엄청난 증가는 인터넷 사용자가 자신에게 적합한 정보를 찾는 것을 어렵게 만들고 있다. 이런 환경의 변화에 따라 사용자에게 필요한 정보를 걸러서 제공하는 정보 필터링기법이 중요해지고 있다. 인터넷에 존재하는 데이터는 다양한 형태로 존재한다. 하지만 기존의 협업필터링 기법에서 자주 사용되어온 유사도 계산 알고리즘들은 수치형데이터에 적합한 경우가 많고, 범주형 데이터의 경우 부울대수 형태의 극단적인 유사도를 보여준다. 본 논문에서는 Gower 유사도 계수를 사용하여 혼합형 데이터로 이루어진 SNS 사용자 정보의 유사도를 구하며 범주형 데이터의 유사도를 0과 1의 극단적 표현이 아니라 좀 더 완화된 형식으로 계산하는 방법을 제안한다. 제안한 방법은 완전 매칭 방법을 사용한 유사도 계산에 비해 세분화된 계산이 가능하다. 이는 범주형 데이터의 초기 데이터량이 희소한 경우 데이터의 활용도를 높여준다. 이를 활용한 군집화 방법은 SNS나 다양한 추천시스템에서 활용될 수 있다.

주제어 : 유사도, 군집화, 혼합형데이터, Gower 유사도 계수

Abstract

Clustering method based on similarity calculation of Mixed Data

Song, Hyoung-Min
Department of Computer Engineering
Graduate School
Jeju National University

Supervised by Professor Lee, Sang-Joon

The enormous increase of data with the development of the information technology makes internet users hard to find suitable information tailored to their needs. In the face of changing environment, the information filtering method, which provide sorted-out information to users, is becoming important. The data on the internet exists as various type. However, similarity calculation algorithm frequently used in existing collaborative filtering method is tend to be suitable to the numeric data. In addition, in the case of the categorical data, it shows the extreme similarity like boolean algebra. In this paper, we get the similarity in SNS user's information which consist of the mixed data using the Gower's similarity coefficient. And we suggest a method that is softer than radical expression such as 0 or 1 in categorical data. The proposed method is more delicate than exact match method. It also make data meaningful in scarce of initial categorical data. The clustering method using this algorithm can be utilized in SNS or various recommendation system.

Keywords : Similarity, Clustering, Mixed data, Gower's similarity coefficient

Ⅰ. 서 론

‘The Digital Universe of Opportunities[1]’에 따르면 전 세계에서 생산된 데이터 총량은 4.4 ZB에 달하고, 2020년에는 10배 증가하여 44 ZB에 이를 것으로 예측되고 있다. 불과 10년 전만 하더라도 인터넷 포털 사이트에서 검색을 하거나 질문을 입력하면 원하는 정보를 쉽게 얻을 수 있었다. 달리 말하면 정보가 부족하여 원하는 정보를 얻지 못하는 경우는 있어도 너무 많은 정보 때문에 올바른 정보가 무엇인지 고민해야 하는 경우는 드물었다. 하지만 요즘에는 넘쳐나는 정보로 인해 옳은 정보가 무엇인지 사용자가 판단해야만 한다.

정보검색을 어렵게 하는 또 다른 요인으로 블로그 마케팅 등 바이럴 마케팅에 의한 정보의 왜곡이 있다. 블로그 마케팅은 광고비용이 거의 들지 않지만 소비자의 거부감이 덜하고 신뢰성이 높아 중소상공인 뿐만 아니라 대기업에서도 활용하는 추세다. 이런 유행의 반작용으로 전문적인 바이럴 마케팅 기업이 돈을 받고 편향된 내용의 블로그를 작성하거나 다수의 아이디를 이용하여 댓글을 작성하고 순위를 조작하는 등의 부작용이 생겨나고 있다.

이런 정보 과잉과 정보 왜곡의 산물로 사용자에게 불필요한 내용을 제거하고 필요한 정보만을 제공해주는 정보 필터링 기법과 여러 곳에 흩어진 콘텐츠를 하나의 콘셉트나 주제로 모아서 보여주는 콘텐츠 큐레이션 기법이 유행하고 있다[2][3].

일반적으로 추천시스템에 사용되는 정보 필터링기법으로는 협업적 필터링과 내용기반 필터링이 있다. 특히 협업적 필터링에서는 사용자와 성향이 유사한 사용자를 찾기 위해 유사도를 계산해야 하는데 기존 시스템에서 많이 사용되어왔던 피어슨 상관계수(Pearson correlation coefficient), 코사인 벡터(cosine vector) 등의 유사도 계산법은 수치형 데이터에만 활용이 가능하다. 그러나 인터넷상에 존재하는 사용자 정보는 수치형 데이터뿐 아니라 다양한 형태의 데이터로 이루어져있기 때문에 혼합형 데이터의 유사도를 구하는 방법이 필요하다. 본 논문에서는 Gower 유사도 계수(Gower's similarity coefficient)를 사용하여 수치형과

범주형으로 이루어진 혼합형 데이터의 유사도를 구한다. 그런데 Gower 유사도 계수에서는 비교하는 두 범주형 변수의 유사도를 구하기 위해 완전 매칭 방법을 사용한다. 즉 범주형 변수를 서로 비교하여 완전히 매칭이 되면 유사도는 1, 조금이라도 다르면 0과 같은 극단적인 표현을 사용한다. 예를 들어 “사람”과 “원숭이”의 유사도와 “사람”과 “오렌지”의 유사도는 모두 0이다. “사람”과 “오렌지”는 유사성이 없다고 볼 수 있지만 “사람”과 “원숭이”는 같은 포유류로서 어느 정도 유사성을 갖는다고 볼 수 있다. 하지만 완전 매칭 방법을 사용하면 “사람”과 “원숭이”의 유사도가 “사람”과 “오렌지”의 유사도 보다 크다는 정보는 손실되기 때문에 비교하는 범주형 변수 간의 미세한 차이를 반영하지 못한다. 본 논문에서는 Gower 유사도 계수에 부분 매칭 방법을 적용한다. 이는 부울대수 형태의 극단적인 표현을 좀 더 완화된 형식으로 바꾸어 유사도 측정의 정확성을 높일 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 필터링 기법과 유사도 측정 알고리즘의 기본 개념과 계산법 그리고 군집화 알고리즘을 소개한다. 3장에서는 Gower 유사도 계수를 변형하여 범주형 변수의 유사도 측정에 부분 매칭 방법을 적용하는 방법을 제안한다. 4장에서는 실제 SNS 사용자 데이터를 수집하여 제안한 방법에 따라 유사도를 측정하고 군집화한 결과를 분석한다. 마지막으로 5장에서는 결론과 향후 연구를 제시한다.

Ⅱ. 관련연구

1. 협업적 필터링(Collaborative Filtering)과 유사도 측정 기법

1) 협업적 필터링

아마존 닷컴에서 추천시스템을 만들 때 사용한 방법으로 유명한 협업적 필터링 기법은 대부분의 추천시스템에서 가장 많이 사용되는 방법이다. 이 기법은 여러 사용자의 평가 값을 이용하여 특정 사용자의 평가 값을 예측하는 방식으로 크게 메모리기반과 모델기반 방식으로 나뉜다.

메모리기반 방식은 두 사용자 간의 또는 항목 간의 유사도를 측정하여 가장 유사한 사용자들 또는 항목들의 평가치를 반영하여 추천한다[4]. 이런 추천 방법은 실시간으로 쌓이는 데이터를 즉시 선호도 예측에 사용할 수 있는 장점이 있다. 그러나 사용자와 아이템의 수가 많아지면 예측시간이 증가하는 확장성의 문제가 있고 평가치 데이터가 희소할 경우, 유사도 계산이 불가능하거나 낮은 유사도의 이웃을 선택하게 되어 추천 품질의 수준이 나빠진다. 또한 유사도 측정 기법에 따라 추천의 정확성에 많은 영향을 미친다[5]. 사용되는 유사도 측정 방법으로는 피어슨 상관계수, 코사인 유사도 등이 있는데 이러한 방법은 대체로 수치형 데이터에 적합한 계산방법이어서 혼합형 데이터에는 적합하지 않다.

모델기반 방식은 선형대수, 신경망, 군집화 등을 기반으로 사전에 모델을 수립해 두고 추천하는 방식으로 확장성과 희소성에 강하고 신속한 추천이 가능하다. 그러나 모델을 구축하는데 많은 시간이 소요되며 학습시간이 별도로 필요하기 때문에 실시간으로 쌓이는 선호도 데이터를 즉시 반영할 수 없어 사용자의 최신 동향을 반영하지 못하는 단점이 있다[6].

2) 피어슨 상관계수(Pearson's Correlation Coefficient)

협업적 필터링에서 유사도 측정 시 널리 사용되는 피어슨 상관계수는 -1과 1사이의 값을 가지며, 연속적인 숫자열간 일대일 비교를 통해 상관관계를 측정한다. 상관관계가 크면 상관계수는 1에 가까워지며, 관계가 적어지거나 거의 없을 경우에는 0에 가까워진다. 또한 사용자와 비교하는 이웃 간의 경향이 서로 대립하는 상관성을 가질 경우 상관관계는 -1에 가까워진다. 피어슨 상관계수의 정의는 식 1과 같다.

$$r_{ij} = \frac{\sum_{m=1}^{p}(x_{im} - \overline{x_m})(x_{jm} - \overline{x_m})}{\sqrt{\sum_{m=1}^{p}(x_{im} - \overline{x_m})^2 \sum_{m=1}^{p}(x_{jm} - \overline{x_m})^2}} \qquad (1)$$

피어슨 상관계수는 거리척도로 변환하여 $d_{ij} = 1 - r_{ij}$ 로 표현할 수 있고 d_{ij}는 0과 2 사이의 값을 갖는다.

3) 코사인 유사도(Cosine Similarity)

코사인 유사도는 두 벡터 간 각도의 코사인 값을 이용하여 측정된 벡터간의 유사한 정도를 나타낸다. 코사인 계산식은 최저 0에서 최대 1사이의 값을 가지며, 각도가 0°일 때 코사인값은 1로 완전히 유사함을 나타내고 90°의 각을 이룰 경우 코사인값은 0으로 비교하는 데이터는 서로 다른 성향을 갖고 있음을 뜻한다. 만약 각도가 180°를 이룰 경우 코사인 값은 -1로 비교하는 두 데이터가 완전히 반대 성향을 갖고 있음을 나타낸다. 계산 방법은 식 2와 같다.

$$\cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|} \qquad (2)$$
$$= \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

4) 유클리드 거리(Euclidean Distance)

두 점 사이의 거리를 계산하는 방법으로 가장 유명하며 가장 흔히 사용하는 계산법이다. 유클리드 거리 d_{ij}의 정의는 아래 식 3과 같다.

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} \qquad (3)$$

유클리드 거리를 이용한 방법은 세 가지 단점을 가지고 있다. 첫째, 거리측정이 선택변수의 측정단위에 따라 달라진다. 따라서 거리 측정 시 원래 측정 단위(척도)보다 표준편차로 나눈 표준 단위 값을 이용하여 거리를 측정해야 한다. 둘째, 유클리드 거리는 변수들 사이의 상관관계를 무시한다. 변수들 간의 상관관계가 있는 경우 각 집단 사이를 판별하는 데 유용한 변수들이라 할지라도 이들 변수들을 함께 사용할 경우 동일한 변수의 효과가 중복해서 나타나게 된다. 셋째, 극단치에 민감하다. 데이터에서 극단치를 제거하기 어려운 경우 다른 거리 측정방법을 사용하는 것이 좋다.

2. Gower 유사도 계수(Gower's Similarity Coefficient)

일반적으로 유사도를 측정할 때 사용되는 알고리즘은 유클리드 거리, 피어슨 상관계수, 코사인벡터 등이 있다. 이런 종류의 알고리즘은 수치형 데이터를 염두에 두고 개발되었기 때문에 수치형과 범주형 등이 혼합된 데이터에는 적합하지 않다.

Gower 유사도 계수는 혼합형 데이터인 경우에 유사도를 구하기 위해 개발되었다. Gower의 유사도 계수는 식 4과 같이 정의된다[7].

$$S_{ij} = \frac{\sum_{m=1}^{p} w_{ijm} s_{ijm}}{\sum_{m=1}^{p} w_{ijm}} \qquad (4)$$

여기서 $w_{ijm} = 1$ 이고, 제약조건으로 다음의 규칙을 따른다. 첫째, 한 쌍의 레코드들 가운데 하나의 측정치가 알려져 있지 않을 때에는 $w_{ijm} = 0$ 이다. 둘째, 이진이 아닌 범주형 변수에 대하여 만약 레코드들이 동일한 범주에 속해 있지 않을 경우에는 $s_{ijm} = 0$ 이고, 동일한 범주에 속해 있을 경우에는 $s_{ijm} = 1$이다. 셋째, 연속형 변수들에 대해서는 식 5와 같다.

$$S_{ijm} = 1 - \frac{|x_{im} - x_{jm}|}{\max(x_m) - \min(x_m)} \qquad (5)$$

표 1. 유사도 계산을 위한 데이터 예시

사용자	나이	성별	학교	거주지	고향
…	…	…	…	…	…
I	35	1	서울	제주	제주
…	…	…	…	…	…
J	26	1	대구	서귀	경산
…	…	…	…	…	…

<표 1>에서 Gower 유사도 계수를 구하면 식 6과 같이 구하여 진다.

$$S_{ij} = \frac{\sum_{m=1}^{5} w_{ijm} s_{ijm}}{\sum_{m=1}^{5} w_{ijm}} \quad (6)$$
$$= \frac{0.888+1+0+0+0}{1+1+1+1+1}$$
$$= \frac{1.888}{5} = 0.3776$$

계산결과를 보면 수치형 변수인 나이를 제외한 성별, 학교, 거주지, 고향은 비교하는 변수값이 같으면 1 다르면 0을 대입하여 계산한다. 이렇듯 범주형 데이터의 유사도 계산에 있어서 완전 매칭법을 사용하기 때문에 <표1>의 예시데이터에서 사용자 I와 J가 각각 '제주'와 '서귀'라는 유사한 지역에 거주하고 있음에도 이를 계산에 반영하지 못한다. 3장에서는 이런 단점을 보완한 계산방법을 제안한다.

3. k-means 알고리즘

1) 군집화 알고리즘

군집화는 크게 계층형 군집화와 비계층형 군집화 두 가지 방법으로 연구되어 왔다. 일반적으로 계층형 군집화는 수행속도가 느리며 데이터의 극단값이나 오류데이터에 민감한 반응을 보인다. 또한 군집화 이전 단계의 오류를 다시 고칠 수 있는 방법이 없다. 따라서 계층적 군집화로는 처리할 수 있는 데이터에 한계가 있고 시간복잡도 측면에서 비효율적이다. 이런 이유 때문에 비계층적 군집화의 일종인 k-means 기반의 군집화가 주로 사용되어 왔다[8].

계층적 군집화는 하나의 레코드로 구성된 군집으로 시작하여 최종적으로 모든 데이터로 구성된 하나의 군집이 남을 때까지 가장 가까운 2개의 군집들을 단계적으로 병합해 나가는 방법이다. 계층적 군집화는 군집의 수를 명시할 필요가 없고 군집화가 오로지 데이터에 의해 수행된다는 장점이 있다. 그러나 데이터의 집합이 매우 클 경우 n×n 거리행렬을 계산하고 저장하기 때문에 계산횟수가 많아지고 자연히 계산속도가 느려진다. 그리고 단 한 번의 군집화를 진행하기 때문에 초기 단계에서 제대로 분배하지 못한 데이터는 그 이후에도 재분배될 수 없다. 또한 데이터를 재정렬하거나 몇 개의 데이터를 제외시키면 전혀 다른 결과를 보여준다. 계층적 군집화는 거리척도가 바뀐다 하더라도 군집분석의 결과에 큰 변화가 없다. 왜냐하면 군집간의 거리를 선택함에 있어 사용되는 단일 또는 완전연결법이 군집 간 거리의 상대적인 순서가 유지되는 한 거리척도의 변화에 안정적이기 때문이다. 그러나 평균연결법을 사용하면 거리척도에 좀 더 영향을 많이 받게 되어 완전히 다른 군집이 형성될 수도 있다.

2) k-means 알고리즘

군집화 알고리즘 중 분할법에 속하는 k-means 알고리즘은 구현이 쉽고, 패턴의 수가 n일 때 시간 복잡도가 $O(n)$인 장점이 있다. 하지만 초기 군집의 중심에 상당히 종속적이라는 단점이 있다[9]. k-means 알고리즘의 개념은 예를 들어 n개의 데이터가 있다면 이 데이터를 n보다 작거나 같은 k개의 군집으로 분할하고 분할된 군집의 중심과 소속 데이터와의 거리를 최소화하는 것이다. <그림 1>은 k-means 알고리즘을 간략하게 나타낸 것이다.

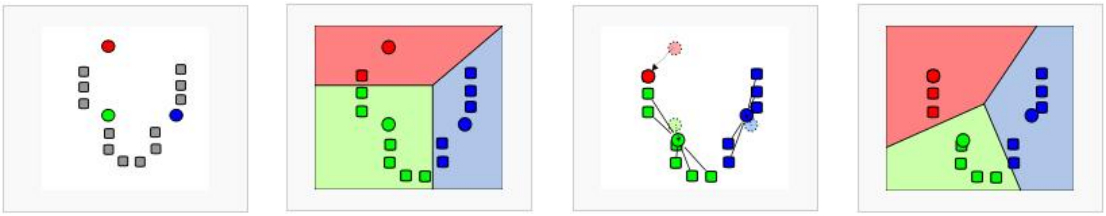
Input

1. k : Number of clusters
2. D : Data set of which the size is n

Output : k Clusters

Algorithm

1. k initial "means" are randomly generated within D.
2. k clusters are created by associating every observation with the nearest mean.
3. The centroid of each of the k clusters becomes the new mean.
4. Steps 2 and 3 are repeated until convergence has been reached.

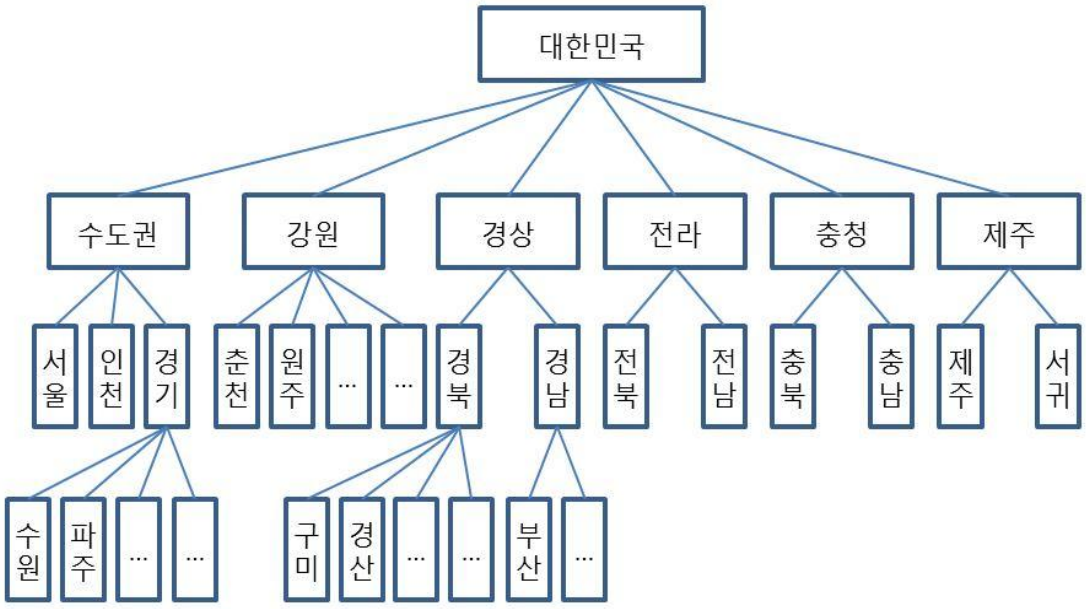


<그림 1. k-means Algorithm[10]>

Ⅲ. 혼합형 데이터에서 유사 사용자 군집화 방법

1. 유사도 계산

본 논문에서는 페이스북 사용자의 프로필 데이터를 가지고 유사도를 구하고 유사 사용자를 군집화 한다. 페이스북 사용자의 프로필 정보는 수치형 또는 범주형 등 다양한 형태의 혼합형 데이터로 이루어져 있다. Gower 유사도 계수는 혼합형 데이터의 유사도 계산에 적합하지만 범주형 데이터의 유사도 계산 방법으로 완전 매칭 방법을 사용하기 때문에 두 개의 데이터가 “같다” 혹은 “아니다”로 극단적으로 표현된다. 이런 부울대수 형태의 표현은 범주형 변수의 값이 완전히 일치 하지 않으면 나머지 유사도에 정보가 손실되는 문제가 있다.



<그림 2. 거주지에 대한 다단계 범주형 데이터 구조도>

본 논문에서는 구조적 관계에 있는 범주형 데이터를 <그림 2>과 같이 다단계 범

주형 데이터 형태로 표현하였다. 이렇게 범주형 데이터를 구조화하고 부분 매칭 방법을 사용하여 유사도를 계산한다면 상위 노드가 같은 노드들은 그렇지 않은 노드보다 유사도가 더 크도록 계산할 수 있다. <그림 2>에서 완전 매칭 방법으로 서울과 인천 사이의 유사도나 서울과 부산 사이의 유사도를 계산한다면 모두 0으로 계산된다. 하지만 부분 매칭 방법을 사용하게 되면 서울과 인천은 수도권이라는 같은 범주의 하위레벨이므로 서울과 부산의 관계 보다 좀 더 유사한 관계가 있다는 정보를 표현할 수 있다.

Gower 유사도 계수의 계산식 중 범주형 변수의 유사도를 부분 매칭 방법으로 구하기 위해 다음과 같은 계산방법을 제안한다. <그림 2>에서 노드와 노드 사이를 잇는 간선의 길이는 1로 정의하고 유사도는 식 7과 같이 정의한다.

$$S_{ijm} = 1 - distance \tag{7}$$

$$distance = \frac{length\ between\ nodes}{\max(length\ between\ nodes)} \tag{8}$$

예를 들어 <그림 2>에서 거주지가 서울인 사람끼리의 거리는 0이고 노드간 거리의 최대값은 6이므로 서울과 인천의 거리는 0.333, 서울과 수원의 거리는 0.5이다. 따라서 서울과 서울의 유사도는 1, 서울과 인천의 유사도는 0.666, 서울과 수원의 유사도는 0.5이다.

그런데 <그림 2>에서 서울과 수원의 거리는 0.5인데 범주가 전혀 다르다고 할 수 있는 수도권과 강원의 거리는 0.333으로 같은 범주내의 항목보다 거리가 더 가깝게 계산된다. 이런 경우가 발생하는 것을 막기 위해 루트 노드를 지나는 거리는 1이라고 정의한다.

이와 같이 범주형 데이터를 완전 매칭 방법이 아닌 부분 매칭 방법으로 변환하여 계산한 후 2장 2절에서 설명한 Gower 유사도 계수 식 2의 s_{ijm}에 대입하면 부울대수 형태의 극단적인 유사도를 좀 더 세분화된 유사도로 표현할 수 있다.

제안한 방식으로 2장 2절의 표 1에서 Gower 유사도 계수를 구하면 식 9과 같이 계산되어 진다.

$$S_{ij} = \frac{\sum_{m=1}^{5} w_{ijm} s_{ijm}}{\sum_{m=1}^{5} w_{ijm}} \quad (9)$$
$$= \frac{0.888+1+0+0.666+0}{1+1+1+1+1}$$
$$= \frac{2.554}{5} = 0.5108$$

만일 학교, 거주지, 고향에 대한 데이터를 다단계 범주형 데이터 구조로 변환하지 않고 완전 매칭 방법으로 Gower 유사도 계수를 구한다면 사용자 I와 사용자 J의 유사도는 0.3776으로 부분 매칭 방법과 비교하여 상대적으로 낮게 구하여진다. 이는 완전 매칭 방법에서는 서울과 대구의 유사도나 제주와 서귀포의 유사도나 모두 0으로 계산되기 때문이다. 즉 완전 매칭 방법과 부분 매칭 방법의 상대적인 유사도 차이만큼 비교하는 변수의 유사도 정보(제주시와 서귀포시가 지리적으로 유사하다는 정보)가 손실된다고 볼 수 있다. 따라서 범주형 데이터를 구조화시켜 부분 매칭 방법을 사용하는 것이 더 정확한 사용자간 유사도를 보여준다고 할 수 있다.

2. 군집화

k-means 알고리즘은 일반적으로 유클리드 거리를 이용해 중심과의 유사도(거리)를 구하지만 본 논문에서는 변형된 Gower 유사도 계수를 사용하여 유사도를 계산한다. 처음 k개의 중심이 선택되면 선택된 k개의 데이터와 모든 데이터와의 유사도를 3장 1절에서 제안한 부분 매칭 방법을 사용하여 계산한다. 새로운 군집이 형성되면 새롭게 형성된 군집의 중심을 다시 계산하는데 수치형 데이터는 산술평균을 사용하여 구하고 범주형 데이터는 최빈값을 사용하여 구한다. 새롭게 형성된 군집의 중심과 모든 데이터와의 유사도를 다시 계산하여 군집을 만드는 과정을 반복한다. 이 반복된 과정은 군집의 중심이 변하지 않을 때까지 계속된다.

Ⅳ. 실험결과

1. 데이터 수집 및 전처리

실험을 위해 페이스북에서 사용자 100명의 프로필 정보를 수집하였다. 수집된 데이터 중 결측치가 많은 32명의 사용자를 제거하고 68명의 데이터를 사용하였으며 데이터 타입에 따라 전처리 과정을 거쳤다. 연속형 데이터는 최소값, 최대값을 정하였고 범주형 데이터는 구조화시킬 수 있는 데이터에 한해 다단계 범주형으로 변환시켜 유사도를 세분화시켰다. 스포츠, 음악, 영화, TV, 책 등에 대한 선호도는 관련페이지를 좋아한 개수를 정규화(Normalization)하여 0~1까지의 연속형 변수로 변환시켜 유사도를 구하였다.

<표 2. 사용자 프로필 항목과 데이터 타입 >

사용자 프로필 항목	데이터 타입
나이	연속형(10~80)
성별	범주형
출신학교	다단계 범주형
거주지	다단계 범주형
고향	다단계 범주형
스포츠	연속형(0~1)
음악	연속형(0~1)
영화	연속형(0~1)
TV	연속형(0~1)
책	연속형(0~1)

출신학교, 거주지, 고향은 구조화된 형태로 표현할 수 있으므로 그림 2에서와 같이 다단계 범주형 데이터로 변환하여 사용하였다. 스포츠, 음악, 영화, TV프로그램, 책에 대한 선호여부는 관련된 페이지를 좋아한 개수를 사용하여 선호도(0~1)로 나타내었다. 선호도는 사용자가 관련 페이지를 좋아한 개수를 평균값의 두 배로 나누어 구하였다. 일반적으로 데이터에서 선호도를 구하는 방법으로는 변수값을 해당 데이터의 최대값으로 나누는 방법이 있다. 이렇게 하면 데이터 내에서 임의의 사용자의 상대적인 선호도를 객관적으로 표현 할 수 있다. 이 경우 최대값은 극단치를 제거하고 사용하는 것이 보통이다. 하지만 극단치의 기준이 애매하기 때문에 본 논문에서는 평균의 2배 이상인 값을 선호도 1로 정하였다. 예를 들어 표 2와 같이 특정사용자 A가 스포츠 관련 페이지를 좋아한 개수가 7개이고 모든 사용자들이 스포츠 관련 페이지를 좋아한 평균 개수가 5라면 평균의 2배인 10으로 나누어 A의 스포즈에 대한 선호도를 7/10로 계산하였다. 그리고 E의 경우와 같이 평균의 2배를 넘는 경우는 선호도를 1로 정의하였다.

<표 3. 스포츠에 대한 선호도 >

사용자	스포츠 관련 페이지를 좋아한 개수	선호도
A	7	$\frac{7}{10}$
B	5	$\frac{5}{10}$
C	8	$\frac{8}{10}$
D	5	$\frac{5}{10}$
E	16	1
F	0	0
G	2	$\frac{2}{10}$
H	4	$\frac{4}{10}$
I	0	0
J	3	$\frac{3}{10}$
합계	50	
평균	5	

<표 4. 전처리 데이터>

사용자	나이	성별	학교	거주지	고향	스포츠	음악	영화	TV	책
1	35	남	경기	제주	제주	0.7	0.5	1	1	1
2	40	남	서울	제주	경기	0.3	0.5	0.3	0.1	0.2
3	46	남	제주	제주	제주	0.3	0.1	1	1	0.4
4	36	남	경기	서울	서울	0.8	0.2	1	0.8	0.6
5	49	남	제주	제주	제주	0.3	0.5	0.5	0.2	0.1
6	16	남	제주	제주	제주	0.9	1	0.2	0.3	0.2
7	46	남	제주	제주	서울	0.1	1	0.4	0.4	0.3
8	35	남	서울	서울	제주	0.3	0.6	0.1	0.8	0.1
9	36	남	경기	서울	울산	0.3	0.3	0.1	0.2	0.5
10	36	남	인천	서울	서울	0.2	0.8	1	0.5	0.5
11	36	남	부산	부산	부산	0.1	1	1	1	1
12	36	남	경기	서울	경기	0.5	0.5	0.4	0.3	0.1
13	32	남	제주	제주	제주	0	0.2	0	0	0.2
14	33	남	서울	서울	서울	0.1	0.5	0.2	0.3	0.2
15	30	여	제주	제주	제주	0.1	0	0.2	0.1	0.4
16	49	남	제주	제주	제주	0.1	0.5	0.2	0.3	0.2
17	33	남	광주	서울	광주	0.2	0.5	0.5	0.5	0.2
18	21	남	제주	제주	제주	0	0	1	1	0.9
19	33	남	제주	제주	제주	0.1	0	0.1	0.1	0.1
20	31	여	제주	제주	제주	0	0.3	0	0	0.5
21	27	여	서울	서울	상파울루	1	0.5	0.3	0.5	0.2
22	36	남	서울	청주	청주	0.1	1	0.8	0.5	0.8
23	33	여	경기	서울	부천	0.2	0.7	0.3	0.2	0.3
24	34	여	제주	제주	제주	0.1	0	0.1	0.5	0.1
25	36	남	서울	부천	토론토	0.5	0.1	0.2	0.7	0.1
26	26	남	대구	서귀포	경산	0.2	1	0.3	1	0.1
27	36	남	경기	인천	인천	0.3	0.4	0.4	0.3	0.2
28	43	남	제주	제주	제주	0.9	1	0.2	0.3	0.2
29	36	남	경기	구미	부천	0.2	0.5	0.5	0.5	0.2
30	20	남	제주	제주	제주	0.5	0.1	0.2	0.7	0.1
31	36	남	경기	광주	경주	0.4	0.6	0.3	0.4	0.4
32	36	남	서울	서울	제주	0.2	0.7	0.4	0.2	0.4
33	36	남	경기	서울	안동	0.2	0.5	0.3	0.2	0.1
34	28	남	서울	서울	서울	0.2	0.5	0.1	0.2	0.5
35	37	여	서울	제주	진주	0	0	0.1	0.1	0.1

사용자	나이	성별	학교	거주지	고향	스포츠	음악	영화	TV	책
36	35	남	서울	서울	제주	0.5	0.3	0.2	0.1	0.1
37	35	여	서울	제주	군산	0.1	0.4	0.2	0.1	0.5
38	38	여	제주	서울	서울	0.2	0.4	0.5	0.5	0.2
39	37	여	경북	제주	제주	0.3	0.5	0.2	0.5	0.7
40	33	여	제주	광주	대구	0.2	0.2	0.5	0.4	0.6
41	39	남	서울	서울	제주	0.7	0.1	1	1	0.7
42	22	남	서울	구리	서울	0.5	0.1	0.2	0.7	0.1
43	33	남	서울	경기	서울	0.2	1	0.3	1	0.1
44	31	남	서울	수원	인천	0.3	0.4	0.4	0.3	0.2
45	25	여	인천	서울	수원	0.4	0.6	0.3	0.4	0.4
46	42	여	경기	서울	서울	0.2	0.7	0.4	0.2	0.4
47	35	남	경기	경기	서울	0.3	0.6	0.1	0.8	0.1
48	40	남	파주	인천	서울	0.3	0.3	0.1	0.2	0.5
49	28	여	경기	경기	서울	0.3	0.1	1	1	0.4
50	36	남	수원	수원	수원	0.8	0.2	1	0.8	0.6
51	24	남	인천	인천	인천	0.1	0	0.1	0.1	0.1
52	23	여	부천	인천	서울	0	0.3	0	0	0.5
53	28	여	서울	서울	서울	0.2	0.5	0.1	0.4	0.4
54	27	남	경주	경주	대구	0.1	0	0	0.1	0
55	27	남	경산	경북	경북	0	0.4	0.1	0.5	0.6
56	26	남	부산	경남	경북	0.5	0.6	0.3	0.2	0.2
57	27	여	경산	경북	경북	0.2	0.1	0.1	0.1	0
58	27	여	경산	경북	경북	0	0.4	0.4	0.1	0.4
59	27	남	경산	경북	대구	0.4	0.8	1	1	0.1
60	27	여	경산	경북	경북	0.1	0.4	0.4	0.4	0.4
61	27	남	경산	대구	경북	0.3	0.2	0.1	0.4	0.1
62	27	남	경산	경북	충남	0.1	0	0	0.1	0
63	27	여	대구	대구	대구	0	0.1	0.1	0.2	0
64	27	여	서울	서울	대구	0.3	0.1	0.1	0.2	0.1
65	27	남	경기	경기	경기	0.2	0.1	0.4	0.4	0.5
66	27	남	경기	울산	경남	0.3	0.4	0.1	0.1	0.1
67	27	남	경기	경기	경기	0.2	0.4	0	0.1	0.4
68	27	남	서울	전남	경북	0.3	0.4	0.4	0.1	0

2. 군집 결과

임의로 초기 k값을 4로 정하고 15, 30, 45, 60번 사용자를 초기 중심으로 지정하여 두 가지 방식의 군집화를 진행하였다. <표 5>는 Gower 유사도 계수와 제안한 방식을 각각 사용하여 1차 군집화한 결과 중 15번 사용자를 중심값으로 설정한 군집1의 비교이다. 내용을 살펴보면 학교, 거주지, 고향 항목의 변수가 Gower 유사도 계수를 사용한 군집에 좀 더 다양하다. Gower 유사도 계수를 사용한 방식에서는 학교, 거주지, 고향 항목을 다단계 범주형으로 변환하지 않고 비교대상이 같으면 1, 다르면 0을 대입하여 유사도를 계산하였기 때문에 제안한 방식에 비해 학교, 거주지, 고향이 유사하지 않다고 판단되는 데이터가 훨씬 많았다. 그래서 나머지 항목이 유사도에 큰 영향을 미치게 되고 범주형 데이터는 군집을 형성하는데 미치는 영향이 작아진다.

<표 5. 1차 군집 결과 중 15번 사용자를 중심으로 한 군집1의 비교>

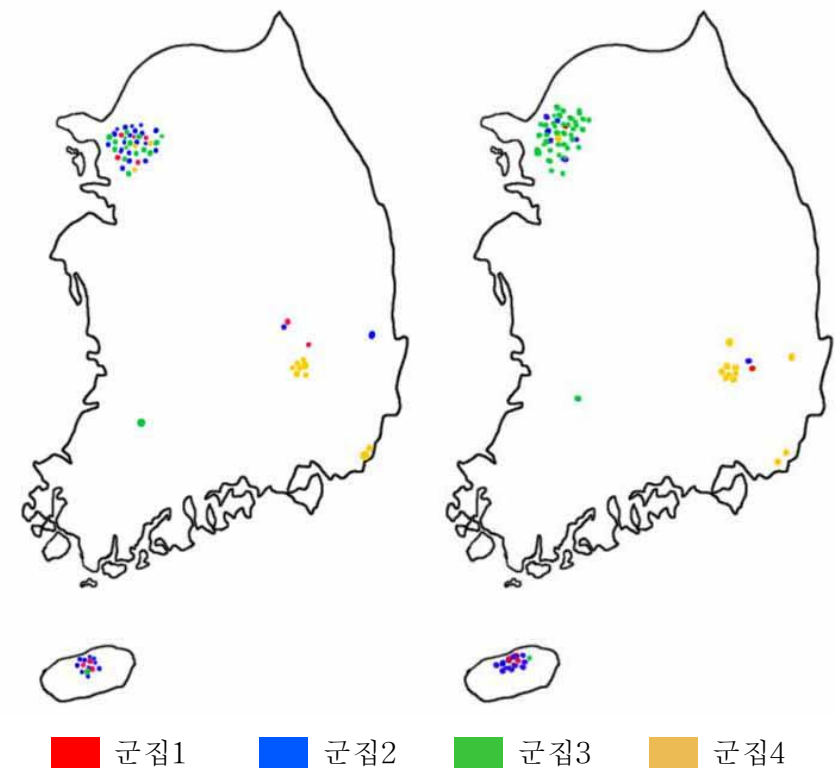
	사용자	나이	성별	학교	거주지	고향	스포츠	음악	영화	TV	책
Gower 유사도 계수를 사용한 군집1	15	30	여	제주	제주	제주	0.1	0	0.2	0.1	0.4
	20	31	여	제주	제주	제주	0	0.3	0	0	0.5
	23	33	여	경기	서울	부천	0.2	0.7	0.3	0.2	0.3
	35	37	여	서울	제주	진주	0	0	0.1	0.1	0.1
	37	35	여	서울	제주	군산	0.1	0.4	0.2	0.1	0.5
	39	37	여	경북	제주	제주	0.3	0.5	0.2	0.5	0.7
	40	33	여	제주	광주	대구	0.2	0.2	0.5	0.4	0.6
	52	23	여	부천	인천	서울	0	0.3	0	0	0.5
	63	27	여	대구	대구	대구	0	0.1	0.1	0.2	0
	67	27	남	경기	경기	경기	0.2	0.4	0	0.1	0.4
	중심 계산	31	여	제주	제주	제주	0.1	0.3	0.2	0.2	0.4
제안한 방식을 사용한 군집1	15	30	여	제주	제주	제주	0.1	0	0.2	0.1	0.4
	20	31	여	제주	제주	제주	0	0.3	0	0	0.5
	24	34	여	제주	제주	제주	0.1	0	0.1	0.5	0.1
	35	37	여	서울	제주	진주	0	0	0.1	0.1	0.1
	37	35	여	서울	제주	군산	0.1	0.4	0.2	0.1	0.5
	39	37	여	경북	제주	제주	0.3	0.5	0.2	0.5	0.7
	중심 계산	34	여	제주	제주	제주	0.1	0.2	0.1	0.2	0.4

<표 6. 1차 군집화 결과 비교>

	Gower 유사도 계수를 이용한 군집 결과	제안한 방식의 군집결과
군집 1	15, 20, 23, 35, 37, 39, 40, 52, 63, 67	15, 20, 24, 35, 37, 39
군집 2	1. 2, 3, 4, 5, 6, 7, 8, 13, 16, 18, 19, 25, 26, 27, 28, 29, 30, 36, 41, 42, 43, 44, 47, 48, 50, 51, 54, 65, 66	1, 3, 5, 6, 7, 8, 13, 16, 18, 19, 25, 26, 28, 30, 36, 41, 42, 66
군집 3	9, 10, 12, 14, 17, 21, 23, 31, 32, 33, 34, 38, 45, 46, 53, 64	2, 4, 9, 10, 12, 14, 17, 21, 22, 23, 27, 29, 31, 32, 33, 34, 38, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 64, 65, 67
군집 4	11, 22, 49, 55, 56, 57, 58, 59, 61, 62, 68	11, 40, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 68

<표 6>은 1차 군집 결과를 보여준다. 1차 군집에서는 두 가지 방식의 군집화가 상대적으로(2차, 3차 군집등과 비교했을 때) 많이 다른 결과를 보인다. 왜냐하면 변수가 완전히 같으면 0, 조금이라도 다르면 1을 대입하는 방법을 사용하면 데이터가 희소하여 같은 범주 값을 갖는 데이터가 별로 없는 경우에 범주형 변수 보다 다른 변수가 유사도에 미치는 영향이 크다. 그래서 제안한 방식에서는 유사하다고 판단된 사용자들이 Gower 유사도 계수를 이용한 군집에서는 각 군집별로 고루 소속된 것을 알 수 있다.

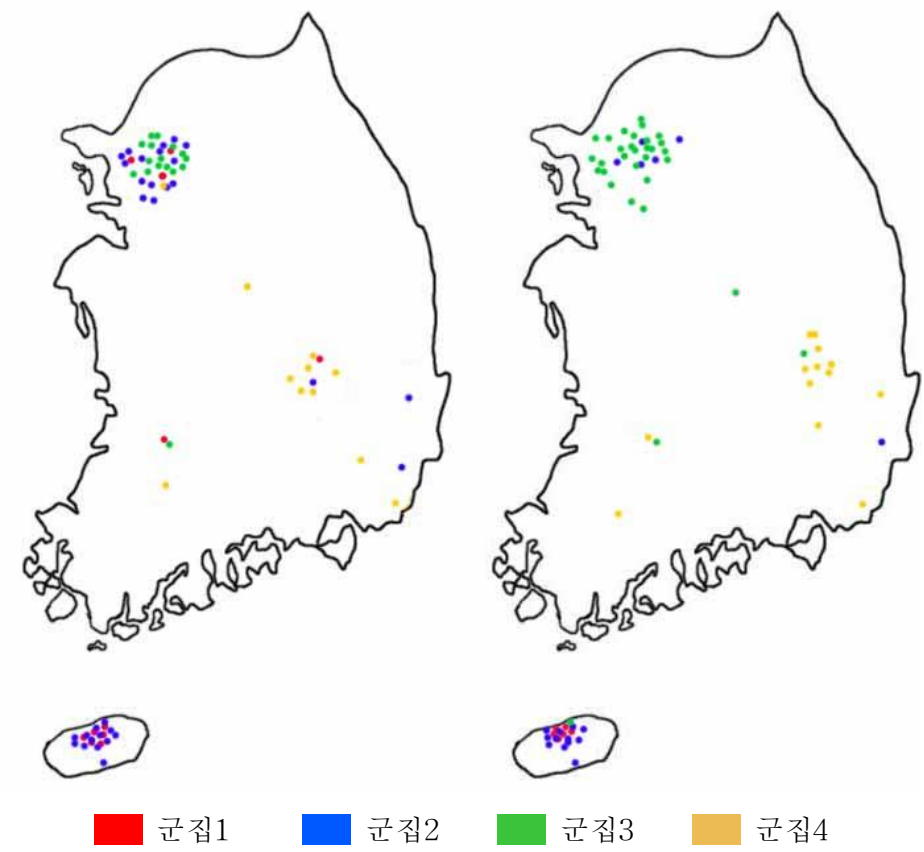
군집 2의 최초 중심값은 30번 사용자로 <표 4>에서 값을 확인하면 (나이 : 20), (성별 : 남), (출신학교 : 제주), (거주지 : 제주), (고향 : 제주), (스포츠 : 0.5), (음악 : 0.1), (영화 : 0.2), (TV : 0.7), (책 : 0.1) 이다. 그런데 Gower 유사도 계수를 이용한 군집에서 군집 2에 소속된 사용자들의 데이터를 <표 4>에서 살펴보면 다양한 출신학교와 다양한 거주지 그리고 다양한 고향을 가진 사용자로 이루어져 있음을 확인할 수 있다. 따라서 군집 2의 최초 중심데이터 중에서 범주형 변수는 군집을 형성하는데 큰 영향을 미치지 않았음을 알 수 있다.



<그림 3. 1차 군집결과 중 군집별 출신학교>

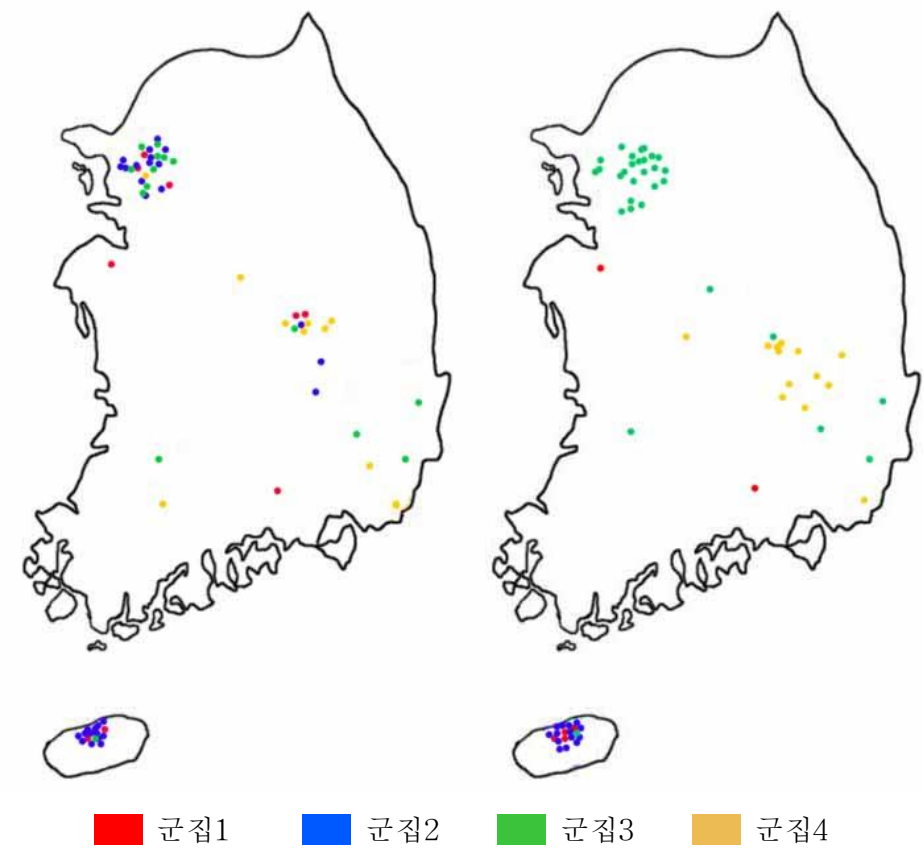
<그림 3>은 출신학교를 지도에 표시 하여 나타낸 것이다. 왼쪽 지도는 Gower유사도 계수를 사용한 결과이고, 오른쪽 지도는 제안한 방식으로 군집을 만든 결과이다. 빨간색 점은 군집1, 파란색 점은 군집2, 녹색 점은 군집3, 황색 점은 군집4를 나타낸다.

<그림 3>을 보면 앞에서 설명한 내용을 가시적으로 확인할 수 있다. Gower 유사도 계수를 이용한 군집화에서는 출신학교라는 범주형 변수가 큰 영향을 못 미쳐서 경상도 지역을 제외하면 수도권과 제주 지역에 각 군집이 고루 분포되었으나, 제안한 방식으로 군집화를 하였을 경우에는 군집 1과 2는 제주 지역에 군집3은 수도권 지역에 군집 4는 경상도 지역에 분포되어 있음이 확연히 드러난다. 즉, 제안한 방식에서는 출신학교라는 범주형 변수가 군집을 형성하는데 영향을 주고 있다.



<그림 4. 1차 군집결과 중 군집별 거주지>

<그림 4>는 1차 군집 결과 중 군집별 거주지를 표시한 것이다. <그림 3>과 마찬가지로 왼쪽 지도는 Gower 유사도 계수를 사용한 군집 결과이고 오른쪽 지도는 제안한 방식을 사용한 군집 결과이다. 군집별 거주지 역시 군집별 출신학교처럼 제안한 방식이 군집의 지역적 특색을 잘 보여주는 것을 확인할 수 있다. Gower 유사도 계수를 이용한 왼쪽 지도에서는 수도권 지역에 각 군집이 고루 섞여 있지만, 제안한 방식을 이용한 오른쪽 지도에서는 수도권 지역에 군집 3이 대부분 차지하고 있다.



<그림 5. 1차 군집결과 중 군집별 고향>

<그림 5>는 1차 군집 결과 중 군집별 고향을 표시한 것이다. 앞서 설명한 <그림 3>, <그림 4>와 마찬가지로 왼쪽 지도는 Gower 유사도 계수를 사용한 군집 결과이고 오른쪽 지도는 제안한 방식을 사용한 군집 결과이다. 군집별 고향 역시 제안한 방식이 군집의 지역적 특색을 잘 보여주는 것을 확인할 수 있다. Gower 유사도 계수를 이용한 왼쪽 지도에서는 수도권에 군집2와 군집3이 많이 섞여 있지만, 제안한 방식을 이용한 오른쪽 지도에서는 수도권이 고향인 사용자가 모두 군집 3임을 알 수 있다.

<표 7. 2차 군집화 결과 비교>

	Gower 유사도 계수를 이용한 군집 결과	제안한 방식의 군집결과
군집 1	15, 20, 24, 35, 37, 39, 40, 52, 63	13, 15, 19, 20, 24, 35, 37, 39, 40
군집 2	1, 3, 5, 6, 7, 13, 16, 18, 19, 28, 30, 50	1, 3, 5, 6, 7, 16, 18, 28, 30
군집 3	2, 4, 8, 9, 10, 12, 14, 17, 21, 22, 23, 25, 27, 31, 32, 33, 34, 36, 38, 41, 42, 43, 44, 45, 46, 47, 48, 49, 51, 53, 54, 64, 66, 67, 68	2, 4, 8, 9, 10, 12, 14, 17, 21, 22, 23, 25, 27, 29, 31, 32, 33, 34, 36, 38, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 64, 65, 66, 67
군집 4	11, 26, 29, 55, 56, 57, 58, 59, 60, 61, 62, 65	11, 26, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 68

<표 7>은 2차 군집 결과를 보여준다. 2차 군집 결과 중 Gower 유사도 계수를 이용한 군집결과를 보면 1차 군집화에서 군집 2에 소속되었던 사용자들의 상당수가 중심이동에 따라 다른 군집으로 이동되었음을 알 수 있다. Gower 유사도 계수를 이용한 방식이 특히 군집 변화가 심한 이유는 유사도를 구하는 데 사용된 변수는 10개 인데 그 중에서 3개의 범주형 변수가 극단적인 결과를 나타내기 때문이다. <표 8>에서 1차 군집 후 최초 k개의 중심과 각 군집에 소속된 사용자 간의 평균 유사도가 제안한 방식 보다 낮다는 것을 확인할 수 있다.

k-means 알고리즘은 군집의 중심과 소속된 개체와의 유사도(거리)를 높여가는 과정이라고 해석할 수 있는데 <표 8>을 보면 제안한 방법은 1차 군집화 단계에서부터 Gower 유사도 계수를 사용한 군집보다 중심과의 평균 유사도가 높게 계산된다. 그 이유는 부분 매칭 방법을 사용하면 s_{ijm}에 대입되는 값이 0이 되는 경우가 상대적으로 적어지기 때문이다. 결과적으로 제안한 방법을 사용하면 군집화가 빨리 이루어질 가능성이 높아진다고도 볼 수 있다. 실제로 <표 8>에서 제안한 방법은 두 번의 중심이동 후 만들어진 3차 군집에서 군집화가 완료된 것을 확인할 수 있다. 반면에 Gower 유사도 계수를 사용한 경우는 네 번의 중심이동 후 5차 군집

에서 군집화가 완료되었다.

<표 8. 군집중심과 소속 사용자 간의 평균 유사도>

구분	군집의 중심과의 평균 유사도							
	Gower 유사도 계수 사용				제안한 방법을 사용			
	군집1	군집2	군집3	군집4	군집1	군집2	군집3	군집4
1차	0.723	0.650	0.650	0.645	0.813	0.722	0.692	0.727
2차	0.778	0.793	0.712	0.712	0.827	0.814	0.774	0.780
3차	0.788	0.753	0.711	0.770	0.826	0.812	0.772	0.782
4차	0.790	0.753	0.728	0.712				
5차	0.790	0.753	0.727	0.713				

<표 9. 3차 군집화 결과 비교>

	Gower 유사도 계수를 이용한 군집 결과	제안한 방식의 군집결과
군집 1	13, 15, 19, 20, 24, 35, 37, 39, 40, 52, 63	13, 15, 19, 20, 24, 35, 37, 39, 40
군집 2	1, 3, 5, 6, 7, 11, 16, 18, 28, 30, 50	1, 3, 5, 6, 7, 16, 18, 28, 30
군집 3	2, 4, 8, 9, 10, 12, 14, 17, 21, 22, 23, 25, 27, 31, 32, 33, 34, 36, 38, 41, 42, 43, 44, 45, 46, 47, 48, 49, 51, 53, 54, 64, 66, 67, 68	2, 4, 8, 9, 10, 12, 14, 17, 21, 22, 23, 25, 27, 29, 31, 32, 33, 34, 36, 38, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 64, 65, 67, 68
군집 4	26, 29, 55, 56, 57, 58, 59, 60, 61, 62, 65	11, 26, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 66

Gower 유사도 계수를 이용한 군집화에서 1차 군집과 2차 군집의 결과가 상대적으로 많은 차이를 보였다면 <표 9>에서 나타난 3차 군집 결과는 Gower 유사도 계수와 제안한 방식 모두 2차 군집 결과에서 크게 다르지 않았다. <표 8>에서 알 수 있듯이 3차 군집부터는 각 군집별 유사도 평균이 이미 0.7을 넘어섰기 때문에 몇몇 개체의 이동만 있을 뿐 이전단계에 비해서 상대적으로 중심의 변화나 군집의 변화가 미미하다.

<표 10. 4차 군집화 결과>

	Gower 유사도 계수를 이용한 군집 결과	제안한 방식의 군집결과
군집 1	13, 15, 19, 20, 24, 35, 37, 39, 40, 52, 63	3차에서 종료
군집 2	1, 3, 5, 6, 7, 11, 16, 18, 28, 30, 50	
군집 3	2, 4, 8, 9, 10, 12, 14, 17, 21, 22, 23, 25, 27, 31, 32, 33, 34, 36, 38, 41, 42, 43, 44, 45, 46, 47, 48, 49, 53, 64	
군집 4	26, 29, 51, 54, 55, 56, 57, 58, 59, 60, 61, 62, 65, 66, 67, 68	

<표 11. 5차 군집화 결과>

	Gower 유사도 계수를 이용한 군집 결과	제안한 방식의 군집결과
군집 1	13, 15, 19, 20, 24, 35, 37, 39, 40, 52, 63	3차에서 종료
군집 2	1, 3, 5, 6, 7, 11, 16, 18, 28, 30, 50	
군집 3	2, 4, 8, 9, 10, 12, 14, 17, 21, 22, 23, 25, 29, 31, 32, 33, 34, 36, 38, 41, 42, 43, 44, 45, 46, 47, 48, 49, 53, 64	
군집 4	26, 27, 51, 54, 55, 56, 57, 58, 59, 60, 61, 62, 65, 66, 67, 68	

3차 군집화 결과인 <표 9>와 4차 군집화 결과인 <표 10> 그리고 5차 군집화 결과인 <표 11>을 보면 Gower 유사도 계수를 사용한 경우에 3차 군집화 이후에는 군집 1과 군집 2는 더 이상 변화가 없고 군집 3과 군집 4 사이의 개체 이동만 있음을 확인 할 수 있다.

데이터의 내용을 살펴보면 Gower 유사도 계수를 사용한 방식에서 군집 1과 군집 2가 먼저 군집화가 완료되는 이유를 알 수 있다. 군집 1과 군집 2에 속한 데이터들은 범주형 변수가 일치하는 데이터가 상대적으로 많아서 s_{ijm}에 1이 대입 되는 경우가 많았기 때문이다. 따라서 군집 1과 군집 2는 군집의 중심과 상대적으로 높은 평균 유사도를 갖게 된다. 반대로 군집 3과 군집 4는 범주형 데이터가 상대적으로 다양하고 비록 수도권이나 경상도라는 지역적 유사성을 갖더라도 완전 매칭 방법을 사용하기 때문에 s_{ijm} 값은 0이 된다. 따라서 군집 3과 군집 4는 군집의 중심과 상대적으로 낮은 평균 유사도를 갖는다.

<표 12. 군집화 완료 결과 비교>

	Gower 유사도 계수를 이용한 군집 결과	제안한 방식의 군집결과
군집 1	13, 15, 19, 20, 24, 35, 37, 39, 40, 52, 63	13, 15, 19, 20, 24, 35, 37, 39, 40
군집 2	1, 3, 5, 6, 7, 11, 16, 18, 28, 30, 50	1, 3, 5, 6, 7, 16, 18, 28, 30
군집 3	2, 4, 8, 9, 10, 12, 14, 17, 21, 22, 23, 25, 29, 31, 32, 33, 34, 36, 38, 41, 42, 43, 44, 45, 46, 47, 48, 49, 53, 64	2, 4, 8, 9, 10, 12, 14, 17, 21, 22, 23, 25, 27, 29, 31, 32, 33, 34, 36, 38, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 64, 65, 67, 68
군집 4	26, 27, 51, 54, 55, 56, 57, 58, 59, 60, 61, 62, 65, 66, 67, 68	11, 26, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 66

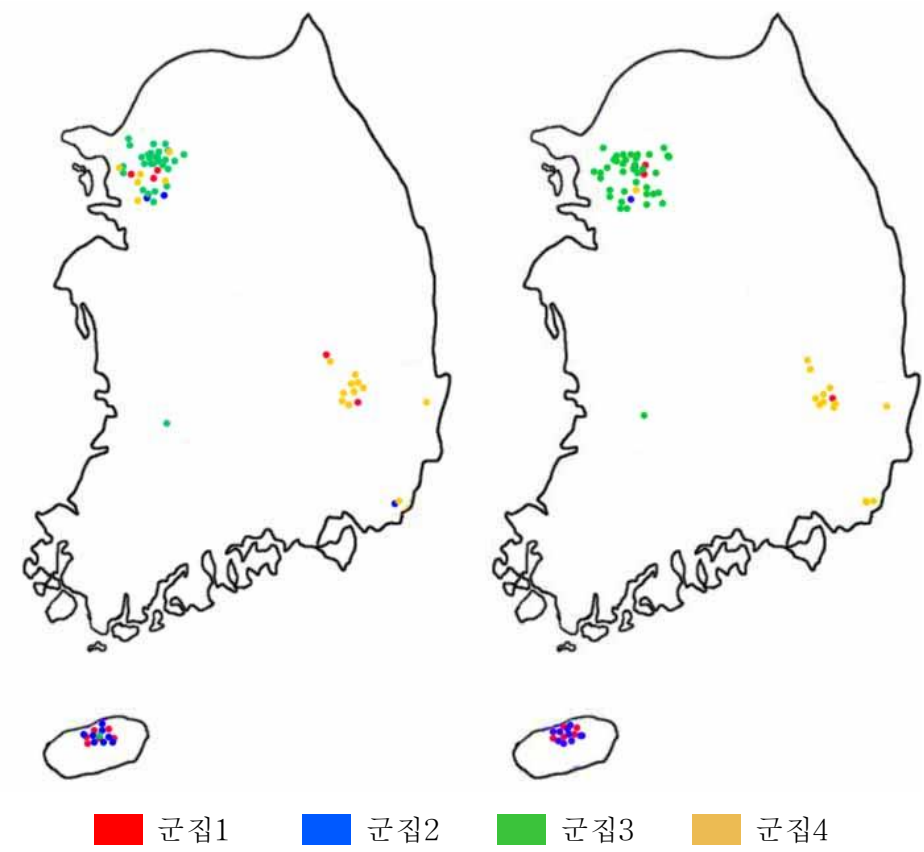
표 12는 군집화 완료 후 제안한 알고리즘을 사용한 군집과 Gower 유사도 계수를 사용한 군집 간의 비교이다. 표 12에서 군집 1의 데이터를 보면 두 가지 방법이 서로 비슷한 군집 결과를 보여준다. 차이점은 Gower 유사도 계수를 사용한 경우에 52번과 63번 사용자가 군집에 추가적으로 포함되었다는 것이다. 52번과 63번 사용자의 데이터를 표 4에서 확인해보면 52번 사용자는 수도권지역 사용자이고 63번 사용자는 대구지역 사용자임을 알 수 있다.

군집 1에서 두 가지 방법에 공통적으로 소속된 나머지 사용자의 특징을 살펴보면 출신학교, 거주지, 고향 항목이 제주와 관련성이 높은 사람들의 집합이다. 즉 Gower 유사도 계수를 사용한 경우에는 걸러내지 못한 52번과 63번 사용자를 제안한 방법에서는 걸러낸 것을 확인할 수 있다. 따라서 제안한 방식으로 군집을 형성하는 것이 더 정확한 군집을 만들어 준다는 것을 알 수 있다.

나머지 군집에서도 다른 결과를 보여주는 사용자를 살펴보면 Gower 유사도 계수를 사용한 방식은 학교, 거주지, 고향 항목에서 전혀 관련성이 없는 사용자를 유사하다고 평가하고 있는 것을 알 수 있다. 예를 들어 군집 4를 살펴보면 제안한 방식의 알고리즘에서는 경상도와 관련된 사용자들 끼리 군집이 형성되었으나 Gower 유사도 계수를 사용한 군집화에서는 수도권과 관련된 사용자 데이터가 섞여서 군집화 되었다.

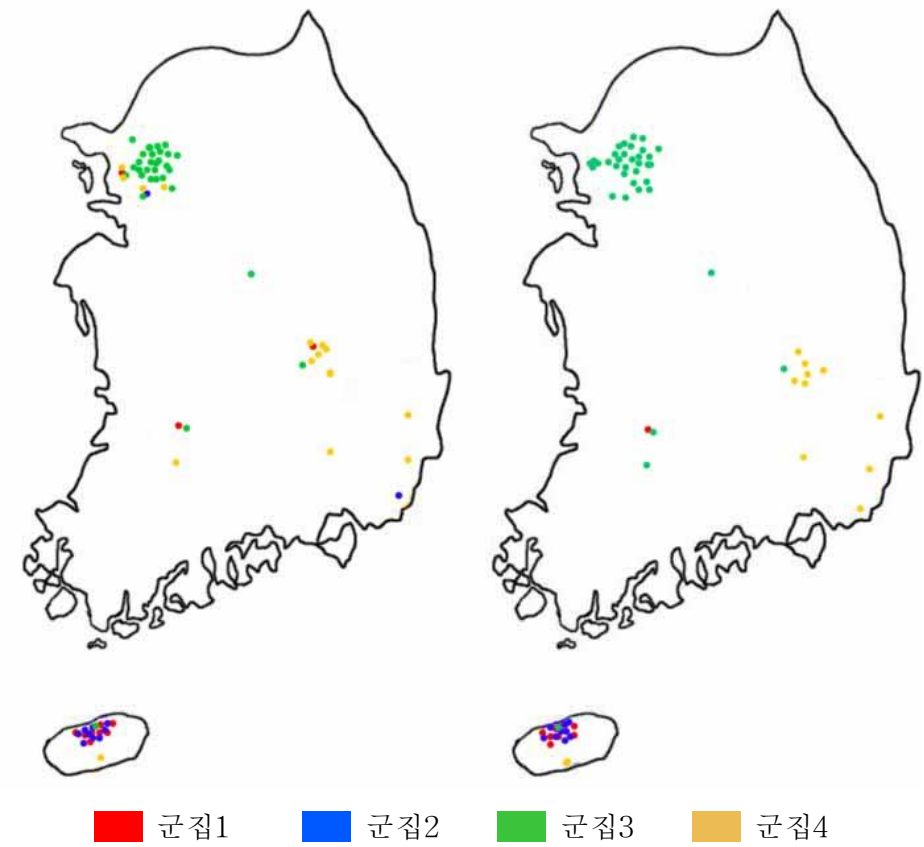
즉 범주형 데이터의 유사도를 완전 매칭 방법으로 구하면 극단적인 결과를 보여주기 때문에 데이터의 양이 적은 경우 범주형 변수들은 유사도 측정에 큰 영향을 주지 못한다. 예를 들어 도서관 대출 데이터를 가지고 협업필터링을 이용한 추천시스템을 만든다면, 어떤 책은 대출정보가 희소하여 선호도 분석에 거의 영향을 주지 않게 될 것이다. 하지만 책의 카테코리를 다단계 범주형으로 구성하면 대출 정보가 희소한 데이터도 의미 있는 정보로 활용이 가능하다.

<그림 8>, <그림 9>, <그림 10>은 최종 군집화 완료 후 군집별 출신학교, 거주지, 고향을 표시한 것이다. 왼쪽지도는 Gower 유사도 계수를 사용한 결과이고 오른쪽 지도는 제안한 방법을 사용한 결과를 보여준다.



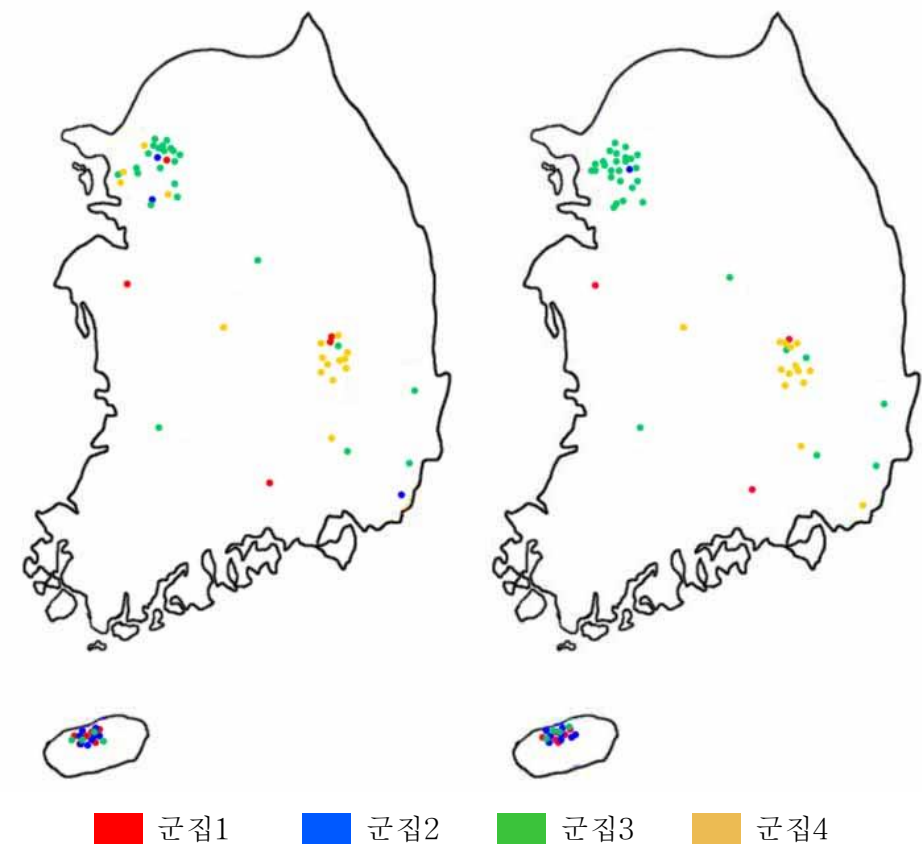
<그림 6. 군집화 완료 후 군집별 출신학교 비교>

군집화 완료 후 군집별 출신학교를 비교하면 Gower 유사도 계수를 사용한 왼쪽 지도는 수도권 지역에 여러 군집이 혼합된 모습을 보여준다. 그러나 제안한 방식을 사용한 오른쪽 지도에서는 상대적으로 수도권 지역에 군집 3에 속한 사용자가 많음을 확인할 수 있다. 달리 설명하면 제안한 방식에서는 범주형 변수의 특색이 군집을 이루는데 상대적으로 더 많이 반영되었다고 할 수 있다.



<그림 7. 군집화 완료 후 군집별 거주지 비교>

군집별 거주지 비교에서는 두 가지 방식의 차이가 더욱 명확해 보인다. <그림 7>의 오른쪽 지도에서 군집 1과 군집 2는 제주지역, 군집3은 수도권 지역 그리고 군집4는 경상도 지역으로 Gower 유사도 계수를 사용한 방식보다 뚜렷한 군집을 보여준다.



<그림 8. 군집화 완료 후 군집별 고향 비교>

<그림 8>의 군집별 고향 역시 제안한 방식이 더 선명한 지역적 특징을 보여 준다.

실험 결과 제안한 방식의 군집화 방법은 다음과 같은 장점이 있음을 알 수 있었다. 첫째, 완전 매칭 방법을 사용한 경우 보다 정확하고 세분화된 유사도를 구할 수 있다. 그 이유는 완전 매칭 방법을 사용했을 경우 손실되는 유사도 정보를 사용할 수 있기 때문이다. 둘째, 데이터가 희소할 경우 데이터의 활용성을 높여준다. 데이터가 희소할 경우 어떤 범주형 변수와 완전히 일치하는 데이터가 존재하지 않을 경우가 있다. 이럴 경우에 완전 매칭 방법에서는 유사도는 항상 0이 된다. 그러나 부분 매칭 방법에서는 상대적으로 유사도가 0 이상이 될 가능성이 높아진다. 즉 희소한 데이터에서도 유사도를 구할 수 있는 가능성이 높아지기 때문에 데이터의 활용도가 높아지게 된다. 셋째, k-means 알고리즘을 수행할 경우 군집 중심과의 평균 유사도가 높아지는 효과가 있어서 알고리즘의 수행이 빨리 완료될 가능성이 높다. 부분 매칭 방법을 사용하면 Gower 유사도 계수에서 s_{ijm}에 0 이상의 값이 대입될 가능성이 상대적으로 높아지기 때문에 완전 매칭 방법에 비해 유사도가 높아지는 효과가 있다. 이는 군집의 중심과 소속 개체와의 평균 유사도를 높여준다. k-means 알고리즘은 군집의 중심과 소속된 개체와의 유사도(거리)를 높여가는 과정이라고 볼 수 있기 때문에 중심과의 평균 유사도가 높아지면 알고리즘이 빨리 완료될 가능성이 높아진다.

V. 결 론

인터넷이나 SNS에 존재하는 데이터들은 여러 형태의 혼합형 데이터로 이루어져 있는 경우가 많다. 본 논문에서는 혼합형 데이터의 유사도를 구하고 사용자를 군집화 하기 위하여 Gower 유사도 계수를 변형하여 사용하였다. 제안한 방식은 구조화가 가능한 범주형 데이터를 트리 형태의 다단계 범주형 데이터로 변환하고 부분 매칭 방법에 의하여 유사도를 구하였다. 그리고 군집화 방법으로 k-means 알고리즘을 이용하였다.

제안한 방식은 크게 세 가지 장점이 있다. 첫째, 극단적인 유사도를 세분화하여 정확성을 높이는 효과가 있다. 둘째, 초기 데이터가 희소할 경우 극단적인 유사도 측정은 범주형 데이터에서 유사한 군집이 거의 발행하지 않지만 부분 매칭 방법을 이용하면 초기 데이터가 희소할 경우에도 범주형 데이터의 활용성을 높여주는 효과가 있다. 이는 초기 데이터가 부족한 추천시스템에서 활용 가능하다. 셋째, k-means 알고리즘 수행 시 군집중심과의 평균 유사도가 높아지는 효과가 있다. 군집 중심과의 평균 유사도가 높아지면 알고리즘의 수행 속도가 빨라질 가능성이 있다.

본 논문에서는 혼합형 데이터에서 군집의 중심값을 구하기 위해 수치형 변수는 산술평균을 중심으로 하였고, 범주형 변수는 최빈값을 중심으로 설정하였다. 그러나 최빈값이 전체 군집에서 차지하는 비중이 낮다면 중심값으로써 의미가 약해진다. 향후에는 범주형 데이터로 이루어진 군집의 중심값을 구하는 방법에 대한 추가적인 연구가 필요하다.

참 고 문 헌

[1] “The Digital Universe of Opportunities”, EMC & IDC, 2014.

[2] 이수진, 전태룡, 백경동, 김성신, “협업적 필터링 및 퍼지시스템 기반 사용자 성향분석에 의한 영화평가 예측 시스템”, 한국지능시스템학회 논문지 2009, Vol. 19, No. 2, pp. 242-247.

[3] 윤여광, “포털 사이트의 콘텐츠 큐레이션에 관한 연구”, 한국엔터네인먼트산업학회논문지, 2014. 12., Vol. 8, No. 4, pp. 31-43.

[4] Ahn Hyung Jun, “A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem”, Information Sciences, Vol. 178, No. 1, pp.37-51, 2007.

[5] 김형도, "잠재 요인 모델의 원리를 이요한 협업 태그 기반 추천 방법", 한국전자거래학회지, 2009. 11., Vol. 14, No. 4, pp.47-57.

[6] 권형준, 홍광석, “협업 필터링 기반 추천 시스템을 이용한 LBS의 개인화”, 한국인터넷정보학회논문지, 2010. 12., Vol. 11 No. 6, pp.1-11.

[7] Galit Shmueli, Nitin R. Patel, Peter C. Bruce, "Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Exel with XLMiner", Wiley, p.299, 2006.

[8] 박동진, 황인극, 안태훈, “범주형 데이터 위주의 데이터베이스를 위한 클러스터링 알고리즘”, 대한산업공학회 ‘98 추계 학술대회 논문집, 1998. 10., pp.355-362.

[9] 이신원, "K-Means 클러스터링에서 초기 중심 선정 방법 비교", 한국인터넷정보학회논문지, 2012. 12., Vol.13, No.6, pp.1-8.

[10] https://en.wikipedia.org/wiki/K-means_clustering