|  |
| --- |
| COVID-19 Risk Analytics and Safe Activity Assistant with Clustering and Classification Algorithms  ABSTRACT  (abstract)  Keywords:  클러스터링 및 분류 알고리즘을 이용한 COVID-19 위험 분석 및 안전 활동 보조  요 약  요약 내용  키워드: |

# 서 론

# 관련 연구

관련 연구 내용

# COVID-19 안전 위험 평가의 설계

COVID-19 안전 위험 평가는 COVID-19 데이터를 이용하여 감염자, 접촉자(감염자와 접촉한 사람) 혹은 장소에 대한 위험성을 계산하고 안전성에 대해 평가하는 것을 의미한다. 이 평가의 목적은 사회에서 완화되어가는 COVID-19 전염병의 위험도를 수치화하여 보여줌으로써 개인이 사회의 일원으로서 놓치고 있었던 막연한 경각심을 일깨워 방어적인 행동을 더 유발하게 해 공익에 좀 더 기여할 수 있는데 있다.

COVID-19 안전 위험 평가는 두 가지 방식으로 산출할 수 있다. GSR(Group Safety Risk, 집단 안전 위험)과 ISR(Individual Safety Risk, 개인 안전 위험)이다. 그리고 이 두 방식에 공통적으로 들어가는 요소인 Severity(위험도)가 필요하다. 이번 장의 하위 항목에서 GSR과 ISR을 산출하기 위한 요소 및 수식을 제안한다.

## 안전 위험 평가 요소

COVID-19 안전 위험 평가를 계산하기 위해 필요한 요소는 크게 감염자나 접촉자의 Severity 계산을 위한 요소와 GSR 값 계산을 위한 요소 그리고 ISR값 계산을 위한 요소로 나뉜다.

### Severity의 요소

Severity는 감염자 혹은 접촉자가 얼마나 위험한지 보여주는 척도이다. 이러한 Severity값을 산출하기 위한 요소는 아래와 같다. COVID-19의 위험성과 감염성에 대한 연구는 현재 진행형이므로 Severity값 산출을 위한 요소는 추가 및 제거될 수 있다.

sElement가 될 수 있는 요소로는 감염날짜, 나이, 성별, 지병 등이있다. 예를들어 *IncurredDate*는 사람이 감염된 날짜를 의미한다.

### GSR의 요소

GSR에서 Group은 시, 군, 구처럼 행정구역을 의미하는 것뿐만 아니라 건물 혹은 편의시설처럼 국소적인 장소를 의미한다. 즉 GSR은 해당 Group이 얼마나 위험한 상태인지 보여주는 척도다. GSR값의 범위는 [0, 1]이며 0이면 무결한 장소를 의미하고 1이면 감염성이 매우 높은 위험한 상태를 의미한다.

이러한 GSR값을 산출하기 위한 요소는 아래와 같다. COVID-19의 위험성과 감염성에 대한 연구는 현재 진행형이므로 GSR값 산출을 위한 요소는 추가 및 제거될 수 있다.

*SeveritySet*은 Group, 즉 지역, 건물 혹은 편의시설에 있었던 사람들의 Severity값의 집합을 나타낸다. *SeveritySet*의 원소인 *severityi*의 범위는 [0, 1]이며 0이면 건강한 상태를 의미하며 1에 가까울수록 그 사람이 위험하다는 것을 의미한다. *SafetyDensity*는 해당 지역, 건물 혹은 편의시설의 내부 사람들의 밀집도를 나타낸다. 범위는 [0, 1]이며 0에 가까울수록 공간이 넓거나 사람이 적음을 의미하고 1에 가까울수록 비좁은 공간에 사람이 많다는 것을 의미한다.

### ISR의 요소

ISR은 개인이 현재 얼마나 위험한 상태인지 보여주는 척도다. 결과값의 범위는 [0, 1]이며 0이면 무결한 상태를 의미하고 1이면 매우 위험한 상태를 의미한다. ISR을 산출하는 데 필요한 요소들은 아래와 같다. COVID-19의 위험성과 감염성에 대한 연구는 현재 진행형이므로 ISR값 산출을 위한 요소는 추가 및 제거될 수 있다.

*Severityindivisual*은 측정하고 하는 사용자의 Severity를 의미한다. *Severityindivisual*의 범위는 [0, 1]이며 0이면 건강한 상태를 의미하며 1에 가까울수록 그 사람이 위험하다는 것을 의미한다. *RelatedPlaceSet*은 개인과 연관성이 깊은 Group의 집합을 의미하며 원소인 *rPlace*는 구체적인 특정 장소를 의미한다. 일례로 술집의 코로나 확산은 현지 직장인보다 현지 중고등학생에게 더 적게 영향을 끼칠 것이며, 대구의 코로나 확산은 태백의 시민보다 서울 시민에게 더 민감하게 작용할 것이다. *SurroundingGSRSet은* 개인의 주변 환경에 위치한 GSR의 집합을 의미하며 원소인 *GSRgroupName­*은 *groupName*의 *GSR*값을 의미한다. 이때 가능한 *groupName*의 범위는 거주지 주변 환경과 현재 위치한 장소의 주변 환경 정보를 포함한다.

## 안전 위험에 대한 수식

본 절에서는 3.1의 요소들로 만든 수식을 제안한다.

### Severity값을 산출하는 수식

*Severity*값을 계산하기 위해서 여러 요소가 포함될 수 있지만 본 논문에서는 감염날짜로부터 경과한 날의 수를 수식에서 이용한다.

*Severity*값을 산출하는 수식에는 감염자에 대한 수식과 접촉자에 대한 수식이 있다. 먼저 감염자에 대한 수식은 다음과 같다.

이때 *DaysAfterInfection*은 양의 정수이며 범위는 [0, 14]이다. 15째일부터는 *Severity*는 0으로 간주된다.

그리고 접촉자에 대한 수식은 다음과 같다.

이때 *DaysAfterContact*는 양의 정수이며 범위는 [0, 14]이다. 15일째부터는 감염자의 경우와 마찬가지로 0으로 간주된다.

현재 수식은 감염 혹은 접촉 후 경과한 날짜만 고려한 수식으로, 3.1.1)에서 언급한 바와 같이 *SeveritypersonID­*에 요소가 추가됨에 따라 수식에 변동사항이 있거나 추가적인 연산이 있을 수 있다. 예를 들어 Age가 Severity의 결과값에 영향을 줄 경우 최종 Severity를 구하는 연산은 아래 식과 같이 변동될 수 있다.

이때 *Severityage*값과 *Severitydays* 값의 범위는 [0, 1]이므로 *Severityfianl*값의 범위도 [0, 1]이 된다.

### GSR값을 산출하는 수식

GSR값을 산출하기 위해 *SeveritySet*과 *Safety-Density*를 수치화하기 위한 계산이 선행되야 한다. 이를 위해 요구되는 수식은 다음과 같다. 먼저 *SeveritySet*의 평균을 구한다. *SeveritySet* 평균은 다음과 같이 나타낼 수 있다.

그리고 *SafetyDensity*를 구하는 공식은 다음과 같다.

*PeopleNumber*은 해당 Group에 있었던 사람 수를 의미한다. *MinimumSafeArea*는 COVID-19의 전염을 막기 위해 사람과 사람 사이에 유지되어야 하는 거리를 기반으로 계산한 면적을 의미한다. 이 거리는 COVID-19 연구에 따라 유동적으로 변할 수 있다. *GroupTotalArea*는 해당 Group의 총 면적을 의미한다. 정수 1과 *min* 함수로 *SafetyDensity*의 최대값은 1이 된다.

*AvgSeverity*와 *SafetyDensity*를 기반으로 GSR값을 구하는 공식은 다음과 같다.

*AvgSeverity*의 범위는 [0, 1]이고 *SafetyDensity*의 범위 또한 [0, 1]이므로 GSRGroupName의 범위도 [0, 1]이다. 0에 가까울수록 COVID-19의 위험성이 낮고 무결한 장소를 의미하며 1에 가까울수록 고위험군이 있거나 많은 감염자들이 존재했음을 의미한다.

### ISR값을 산출하는 수식

ISR값을 산출하기 위해 RelatedPlaceSet과 SurroundingGSRSet을 통해 유의미한 GSR을 분류하는 작업이 선행되어야 한다. 개인에게 적합한 group을 추출하기 위한 TargetGSRSet을 구하기 위한 공식은 다음과 같다.

TargetGSRSet은 주변 Group중에서도 연관성이 있는 Group을 tGSR로 추출하여 원소로 가지고 있는 집합을 의미한다. tGSR은 GSR과 동일하며 범위와 의미 또한 동일하다.

이러한 TargetGSRSet을 바탕으로 산출되는 ISR값은 다음과 같다.

*ws*와 *wg*는 각각 Severity에 대한 가중치와 tGSR에대한 가중치를 의미하며 두 가중치의 합은 1이고 *Severity*와 *tGSR*이 ISR값에 미치는 영향을 조절한다. *SeveritypersonID*는 ISR을 사용자(ISR값을 측정하는 사람)의 *Severity*값을 의미한다. Severity와 범위와 의미가 동일하다. *LivingDistance*는 사용자의 생활 반경을 의미한다. 사용자가 입력할 수 있는 값이다. *tGSRDistance*는 사용자와 tGSR과의 거리를 의미한다. *LivingDistance*와 *tGSRDistance*로 사용자로부터 거리에 따른 가중치를 조절할 수 있다. *tGSR*은 *TargetGSRSet*의 원소로 GSR과 범위와 의미가 동일하다. 이 요소들을 계산해서 나온 ISR값의 범위는 [0, 1]이며 0에 가까울수록 사용자가 COVID-19의 감염 위험에서 안전하다는 것을 의미하며 1에 가까울수록 생활 반경 내에 COVID-19 감염 위험이 많다는 것을 의미한다.

# 인적데이터 및 Group 클러스터링

## 요소들을 이용한 인적데이터 클러스터링

클러스터링(Clustering)이란 거리 함수(Distance Function)와 데이터들의 특징(Feature)을 이용하여 데이터들 간의 거리를 계산하고 이를 기반으로 여러 개의 클러스터를 생성하여 데이터를 클러스터별로 분류하는 것을 의미한다. 즉, 인적데이터 클러스터링이란 인적데이터를 이용하여 클러스터링하는 것을 의미한다.

지금은 COVID-19에 의한 팬데믹(pandemic) 시대로 지금껏 인류가 경험하지 못한 새로운 시대로 접어들고 있다. 팬데믹 시대에는 1명의 감염자 혹은 접촉자는 슈퍼 감염자가 될 수 있는 위험한 가능성이 있으며 이에 따른 전염병을 통제하는 방법이 요구되는 만큼, 전염자 및 접촉자를 분류하는 방법도 필연적으로 요구된다. 본 논문에서 제시하는 클러스터링을 통한 인적데이터 분류는 현 시대에 COVID-19를 통제하는데 큰 도움을 줄 수 있다.

## 가중치를 준 클러스터링 거리 함수

본 논문에서는 클러스터링에 사용하는 거리 함수이다. 일반적으로 사용되는 거리함수와 다르게 본 논문에서는 가중치를 적용한 거리 함수를 제안한다. 가중치는 클러스터링에 사용되는 특징(Features)들에 적용되며 각 특징에 영향력을 조절하게 된다. 또한 가중치의 합은 1로 유지되어 가중치 간의 비율을 유지하여 일부 특징이 과도하게 영향력을 가지게 되는 것을 방지한다.

예를 들어 유클리드 거리 함수를 이용하는 경우, 특징 x, y 대한 점 p(px, py)과 q(qx, qy)이 있고 두 점사이의 거리를 *distancepq*라 하면 *distancepq*의 값을 구하는 공식은 다음과 같다.

그리고 가중치 x에 대한 가중치 wx와 y에 대한 가중치 wy를 추가한 변형 공식은 다음과 같다.

## 클러스터링 결과 및 활용

본 논문에서는 COVID-19 안전 위험 평가에 필요한인적데이터로 클러스터링을 진행한다. 이런 특징들로 클러스터링을 진행한 결과로 인적데이터를 클러스터별로 나눌 수 있다. 이 클러스터들은 사용한 인적데이터에 따라 현재 각 클러스터들이 COVID-19에 대해 얼마나 위험한지 가늠할 수 있으며, 미래에 각 클러스터들이 COVID-19에 대해 얼마나 취약해질 수 있는 지를 가늠할 수 있다.

Clustering Results

Clustering 결과의 활용

## Clustering Groups using Factors

Group Clustering 필요성

Clustering Algorithm

Clustering Results

Clustering 결과의 활용

# Design of Safe Activity Assistant

Query processing

….

….

# Experiments and Assessment

# Conclusion

References

1. …
2. …
3. …