|  |
| --- |
| COVID-19 Risk Analytics and Safe Activity Assistant (System) with Machine Learning Algorithms  ABSTRACT  (abstract)  Keywords:  머신 러닝 알고리즘을 이용한 COVID-19 Risk 분석 및 Safe Activity 지원 시스템  요 약  요약 내용  키워드: |

# 서 론

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, 사람 클러스터링이란 사람을 이용하여 클러스터링하는 것을 의미한다.

# 관련 연구

관련 연구 내용

# COVID-19 리스크 평가 체계

본 절에서는 COVID-19 리스크 평가에 대한 설명과 필요성 그리고 COVID-19 리스크 평가를 위한 요소들과 메트릭을 소개한다.

## COVID-19 리스크

COVID-19 리스크 평가는 COVID-19 데이터를 이용하여 감염자(COVID-19에 걸린 확진자), 접촉자(감염자와 접촉한 사람) 혹은 장소에 대한 위험성을 계산하고 수치화하여 등급을 산정하는 것을 의미한다. 이 평가의 목적은 COVID-19에 의한 팬데믹(Pandemic)이 시작되고 아직 백신이 개발되지 않은 현 시점에서 지역 사회에서 사회 구성원들이 COVID-19 전염성 및 위험성에 대해 둔감해지는 것을 예방하는 것에 있다. 즉, 전염병의 위험도를 수치화하여 보여줌으로써 개인이 사회의 일원으로서 놓치고 있었던 막연한 경각심을 일깨워 COVID-19에 대해 방어적인 행동을 유발하거나 지속하게 하여 공익 및 지역 사회 안전에 좀 더 기여하도록 돕는데 있다.

## COVID-19 리스크 메트릭

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

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

## COVID-19 개인 리스크 평가 메트릭

### Severity값을 산출하는 메트릭

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.

*sElement*가 될 수 있는 요소로는 감염날짜, 나이, 지병 등이있다. 예를 들어 *IncurredDate*는 사람이 감염된 날짜를 의미한다.

즉, *Severity*값을 계산하기 위해서 여러 요소가 포함될 수 있지만 본 논문에서는 감염날짜로부터 경과한 날의 수만을 메트릭에서 이용한다.

*Severity*값을 산출하는 메트릭에는 감염자에 대한 메트릭과 접촉자에 대한 메트릭이 있다. 먼저 감염자에 대한 메트릭은 다음과 같다.

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

그리고 접촉자에 대한 메트릭은 다음과 같다.

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

예를 들어 감염자가 감염된 후 5일이 경과한 후에 *Severity*를 측정한다면 0.75가 나온다.

현재 메트릭은 감염 혹은 접촉 후 경과한 날짜만 고려한 메트릭으로, 3.1.1)에서 언급한 바와 같이 *SeveritypersonID­*에 요소가 추가됨에 따라 메트릭에 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. 예를 들어 나이만 다르고 동일한 조건의 사람 둘이 동시에 Severity값을 측정하게 된다면 40살의 Severityage는 20살의 Severityage값보다 높게 나오므로 40살의 최종 Severityfinal값이 더 크게 나올 것이다.

### ISR값을 산출하는 메트릭

ISR은 개인이 현재 얼마나 위험한 상태인지 보여주는 척도다. 결과값의 범위는 [0, 1]이며 0이면 무결한 상태를 의미하고 1이면 매우 위험한 상태를 의미한다. ISR은 감염된 상태의 위험도를 의미하는 Severity와 달리 주변 환경을 고려하여 현재 개인에게 얼마나 큰 위협이 있는지를 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 사용자(ISR값을 측정하는 사람)의 Severity를 의미한다. *Severityindivisual*의 범위는 [0, 1]이며 0이면 건강한 상태를 의미하며 1에 가까울수록 그 사람이 위험하다는 것을 의미한다. *RelatedPlaceSet*은 개인과 연관성이 깊은 그룹의 집합을 의미하며 원소인 *rPlace*는 구체적인 특정 장소를 의미한다.

예를 들어 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. GSR은 3.4에서 후술한다. 가능한 *GroupName*의 범위는 거주지 주변 환경과 현재 위치한 장소의 주변 환경 정보를 포함한다. 예를 들어 *GroupName*은 거주지 근처 상점의 상호나 학교부지 전체를 나타낼 수 있다.

ISR값을 산출하기 위해 *RelatedPlaceSet*과 *SurroundingGSRSet*을 통해 유의미한 GSR을 분류하는 작업이 선행되어야 한다. 개인에게 적합한 그룹을 추출하기 위한 *TargetGSRSet*을 구하기 위한 공식은 다음과 같다.

*TargetGSRSet*은 주변 그룹 중에서도 사용자와 연관성이 있는 그룹을 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 사용자의 *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. *n*은 *tGSR*의 수, 즉 *TargetGSRSet*의 원소의 수를 의미한다. *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.

예를 들어 *Severity*에 더 큰 영향력을 주고자 하면 *ws*에 0.7을, *wg*에 0.3을 부여한다. 사용자가 *SeverityIncurredDate*만 이용할 경우, 감염된 후 6일이 지났다고 가정하면 사용자의 *SeverityPersonID*는 0.7이 된다. 주변에 *tGSR*이 많고 그 거리가 대부분 멀지 않을 경우 시그마항의 값은 1에 가깝게 나올 것이다. 그러면 주어진 식에 따라 *ISRPersonID*는 1에 가까운 값이 도출된다.

## COVID-19 그룹 리스크 평가 메트릭

GSR에서 그룹은 시, 군, 구처럼 행정구역을 의미하는 것뿐만 아니라 건물 혹은 편의시설처럼 국소적인 장소를 의미한다. 즉 GSR은 해당 그룹이 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 그룹, 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. 예를 들어 20명 정도가 적정인 피트니스 센터에 40명이 있고, 그 중 감염자가 2명이라면 *SeveritySet*은 {0.8, 0.7}로 측정될 수 있고 *SafeDensity*는 높게 나올 것이다.

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

그리고 *SafetyDensity*를 구하는 공식은 다음과 같다.

*PeopleNumber*은 해당 그룹에 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 그룹의 총 면적을 의미한다. *min*함수는 입력 값 중에 최소값을 출력하는 함수이다. 즉, 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 *GSRGroupName*의 범위도 [0, 1]이다. 0에 가까울수록 COVID-19의 위험성이 낮고 무결한 장소를 의미하며 1에 가까울수록 고위험군이 있거나 많은 감염자들이 존재했음을 의미한다. 예를 들어 20명 정도가 적정인 피트니스 센터에 40명이 있고, 그 중 감염자가 2명이라면 *SeveritySet*은 {0.8, 0.7}로 측정될 수 있고 *AvgSeverity*는 0.75가 나온다. 그리고 센터의 *TotalArea*가 200이고 *MinimumSafeArea*가 10이라면 *SafeDensity*는 *min*함수에 의해 1을 넘기는 좌항이 아닌 우항 1을 추출할 것이다.

# 사람 및 그룹 클러스터링

본 장에서는 3절에서 언급한 요소들과 메트릭을 통해 나온 결과값으로 클러스터링을 진행한다. 그리고 클러스터링에 사용하는 거리 함수를 일반적으로 사용되는 거리함수와 다르게 가중치를 적용한 거리 함수를 설명한다.

## 가중치를 적용한 거리함수

가중치는 클러스터링에 사용되는 특징(Feature)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 *wx*와 y에 대한 가중치 *wy*를 추가한 변형 공식은 다음과 같다.

## 요소들을 이용한 사람 클러스터링

### 사람 클러스터링의 필요성

지금은 COVID-19에 의한 팬데믹 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, 감염자 및 접촉자를 분류하는 방법도 필연적으로 요구된다. 이때 감염자와 접촉자와 가까운 사람들의 생활 습관 및 행동은 무결한 곳에 사는 사람들에 비해 더 방어적으로 임할 필요가 있다. 사람 클러스터링을 통한 사람 분류는 지역 사회 주민들의 방어적인 행동에 일조하고 지역 전염율을 낮추는데 도움을 줄 수 있다. 예를 들어 이태원 클럽에서 감염자가 존재해서 수많은 접촉자가 생겼을 경우, 이태원 지역 주민 및 접촉자들의 거주지 지역 주민들의 행동은 더욱 조심스러워야 한다. ISR을 통한 사람 클러스터링은 지역 주민들의 방어적인 행동에 일조하여 COVID-19의 지역사회 전파율을 감소시킬 것이다.

### 사람 클러스터링 알고리즘 및 입출력 데이터

클러스터링을 진행하기 위해서는 데이터의 특징이 요구된다. 마찬가지로 사람 클러스터링 알고리즘도 특징이 필요하다는 건 동일하지만 사용자는 원하는 종류를 선택할 수 있다. 예를 들어 필수적으로 포함되는 ISR값을 비롯하여 Age, Gender, Related Disease, Address 등을 선택하여 클러스터링을 진행할 수 있다.

예를 들어 특징으로 ISR값과 Age를 사용할 경우 클러스터링 알고리즘의 입력 데이터는 다음과 같다.

Table 1. People Clustering Input Date Example

|  |  |  |
| --- | --- | --- |
| ID | ISR | Age |
| 1 | 0.01 | 27 |
| 2 | 0.05 | 43 |
| 3 | 0.03 | 71 |
| 4 | 0.65 | 22 |
| 5 | 0.89 | 22 |
| … | … | … |

그리고 4.1의 가중치를 적용한 거리함수를 이용하여 클러스터링을 진행한다. 이때 거리함수 또한 유클리드 거리를 비롯하여 원하는 거리함수를 선택할 수 있으며 가중치 또한 입력 가능하다. 이러한 과정을 통해 나오는 출력데이터 예시는 다음과 같다.

Table 2. People Clustering Output Data Example

|  |  |  |  |
| --- | --- | --- | --- |
| ID | ISR | Age | ClusterID |
| 1 | 0.01 | 27 | 1 |
| 2 | 0.05 | 43 | 1 |
| 3 | 0.03 | 71 | 2 |
| 4 | 0.65 | 22 | 3 |
| 5 | 0.89 | 22 | 3 |
| … | … | … | … |

### 사람 클러스터링의 활용

사람 클러스터링을 진행하면 ISR값 및 선택한 특징 값에 따라 사람들이 클러스터별로 분류된다. 이 클러스터들은 ISR과 선택한 특징에 따라 현재 각 클러스터들에 속한 사람들이 COVID-19에 대해 얼마나 위험한지 가늠할 수 있으며, 미래에 각 클러스터들이 COVID-19에 대해 얼마나 취약해질 수 있는 지를 예측하는데 활용될 수 있다. 예를 들어 서울시 관악구에 감염자 3명이 원인을 알 수 없는 경로로 등장했을 경우, 감염자들과 함께 있던 사람들은 접촉자가 된다. 그리고 감염자들과 접촉자들의 Severity가 높은 값으로 갱신되며 주변 사람들의 ISR과 주변 그룹들의 GSR이 갱신된다. 그러면 클러스터링의 결과가 기존의 값과 달라질 확률이 높아진다. 만약 본인의 클러스터가 변하고 고 위험군이 될 경우 본인의 COVID-19의 감염율을 높이게 되는 활동을 자제할 것이다.

## 요소들을 이용한 그룹 클러스터링

### 그룹 클러스터링의 필요성

현재 CDC(Centers for Disease Control and Prevention, 미국의 질병관리예방본부)에 따르면 COVID-19의 전염 방식은 사람과 사람사이에서 호흡기 비말을 통해 전파되며 공기 중 전파는 밝혀진 감염경로가 아니다. 하지만 WHO(World Health Organization)에 따르면 공기 중 전파의 가능성을 완전히 배제할 수는 없으며 특정 조건 하에서는 공기로 통한 전염이 충분히 가능하다고 경고하며 엘리베이터 같은 좁은 실내에서 짧은 시간내에 감염되는 사례도 보고되고 있다.

그렇기에 COVID-19의 영향력이 사람으로부터 나오기도 하지만 감염자나 접촉자가 머문 장소로부터 감염될 수 있기 때문에 감염자나 접촉자에 대한 분류만 아니라 이들이 머무르거나 지나간 장소에 대한 분류에 대한 요구도 필연적이라 할 수 있다. 예를 들어 술집에서 감염자와 술을 마신 접촉자가 지하철을 타고 편의점을 방문한 후 집에 도착했다면, 해당 술집과 지하철 그리고 편의점에서 추가적인 감염자가 발생할 수 있다. 그러므로 이런 그룹들은 즉시 방역대상이 포함되며 방역 전까지는 다른 무결한 그룹들과 다르게 관리되어야 할 것이다. 그룹 클러스터링은 이러한 경우에 핵심적인 역할을 할 것이다.

### 그룹 클러스터링 알고리즘 및 입출력 데이터

사람 클러스터링과 마찬가지로 그룹 클러스터링 알고리즘도 특징이 요구되며 사용자는 원하는 종류를 선택할 수 있다. 예를 들어 필수적으로 포함되는 GSR값을 비롯하여 BusinessCategory, DailyDensity, RelatedGroup등을 선택하여 클러스터링을 진행할 수 있다. 예를 들어 특징으로 ISR값과 BusinessCategory를 사용할 경우 클러스터링 알고리즘의 입력 데이터는 다음과 같다.

Table 3. Group Clustering Input Date Example

|  |  |  |
| --- | --- | --- |
| ID | GSR | BusinessCategory |
| 1 | 0.33 | A Bar |
| 2 | 0.89 | B Bar |
| 3 | 0.78 | K Restaurant |
| 4 | 0.11 | J Laundry |
| 5 | 0.01 | T Apartment |
| … | … | … |

그리고 4.1의 가중치를 적용한 거리함수를 이용하여 클러스터링을 진행한다. 이때 거리함수 또한 유클리드 거리를 비롯하여 원하는 거리함수를 선택할 수 있으며 가중치 또한 입력 가능하다. 이러한 과정을 통해 나오는 출력데이터 예시는 다음과 같다.

Table 4. Group Clustering Output Data Example

|  |  |  |  |
| --- | --- | --- | --- |
| ID | GSR | BusinessCategory | ClusterID |
| 1 | 0.33 | A Bar | 1 |
| 2 | 0.89 | B Bar | 2 |
| 3 | 0.09 | K Restaurant | 3 |
| 4 | 0.11 | J Laundry | 3 |
| 5 | 0.01 | T Apartment | 3 |
| … | … | … | … |

### 그룹 클러스터링의 활용

그룹 클러스터링을 진행하면 GSR값 및 선택한 특징 값에 따라 그룹들이 클러스터별로 분류된다. 이 클러스터들은 GSR과 선택한 특징에 따라 현재 각 클러스터들이 COVID-19에 대해 얼마나 위험한지 그룹인지 가늠할 수 있으며, 미래에 각 클러스터에 속한 그룹들이 COVID-19에 대해 얼마나 취약해질 수 있는 지를 예측하는데 활용될 수 있다. 다만 사람 클러스터링과 다르게 방역시스템에 의해 바로 0으로 초기화 될 수 있다. 예를 들어 대구의 한 교회에서 대규모 집단 감염으로 해당 교회 및 주변 지역 일대의 GSR이 매우 높은 값을 가지게 된다면 해당 그룹들의 클러스터군은 고위험군 클러스터로 분류될 것이다. 이 수치들은 지역 사회의 주민들의 ISR값 갱신 및 주민들의 클러스터에 큰 영향을 미치며 주민들의 행동에 큰 영향을 준다. 감염자 및 접촉자들이 분리되고 지역 일대에 방역이 마무리되면 GSR값은 크게 감소되고 지역 주민들의 방어적인 행동은 조금 완화될 수 있을 것이다. 다만 지역 내 자가 격리된 지역 주민들이 여전히 존재하니 ISR은 0이 되지 않으며 지역 주민들에게 COVID-19 확산에 대해 지속적인 주의를 요할 것이다.

# Design of Safe Activity Assistant

Query processing

….

# Experiments and Assessment

실험 관련 내용 및 평가 관련 내용

# Conclusion

결론 내용

References

1. …
2. …
3. …