반도체디스플레이기술학회지 제23권 제2호(2024년 6월)
Journal of the Semiconductor & Display Technology, Vol. 23, No. 2. June 2024.

방향성을 고려한 밀도 기반 클러스터링 기법에 관한 연구

김진만[*]·국중진[**†]
[*]한국전자정보통신산업진흥회, [**†]상명대학교 정보보안공학과

A Study on Density-Based Clustering Method Considering Directionality

Jinman Kim[*] and Joongjin Kook[†]**
[*]Korea Electronics Association
[**†]Dept. of Information Security Engineering, Sangmyung University

ABSTRACT

This research proposed DBSCAN-D, which is a clustering technique for locating POI based on existing density-based clustering research, such as GPS data, generated by moving objects. This method is designed based on 'staying time' and 'directionality' extracted from the relationship between GPS data. The staying time can be extracted through the difference in the reception time between data using the time at which the GPS data is received. Directionality can be expressed by moving the area of data generated later in the direction of the position of the previously generated data by concentrating on the point where the GPS data is sequentially generated. Through these two properties, it is possible to perform clustering suitable for the data set generated by the moving object.

Key Words : Clustering, DBSCAN, Density, POI, GPS

1. 서 론

최근 스마트폰과 같은 이동 단말기의 확산과 더불어 GPS(Global Positioning System), WSN(Wireless Sensor Network)과 같은 통신 기술의 발달과 모바일 컴퓨팅 기술의 발전으로 인해 자동차나 사람과 같은 이동 객체의 궤적 정보 획득 수단이 다양해지고 있다. 이동 객체란 시간의 변화에 따라 공간적인 위치 및 모양이 연속적으로 변하는 시공간 데이터를 생성하는 주체로 이러한 객체에 의해 생성된 데이터가 시간의 연속성에 의해 객체의 이동 궤적을 나타낸다. 최근 다양한 응용프로그램에서 위치기반 서비스(LBS, Location-Based Services)의 사용이 증가하면서 이러한 데이터에서 의미 있는 정보를 추출하는 것에 대한 관심이 증가하고 있다.

[†]E-mail: kook@smu.ac.kr

위치기반 서비스는 이동통신망이나 위성신호 등을 이용하여 이동 단말기의 위치를 측정하고, 측정한 위치와 관련된 다양한 정보서비스를 제공하기 위한 기술로서 이의 기술체계에는 일반적으로 휴대 단말의 위치를 파악하는 무선측위 기술과 서비스를 위한 핵심 기반기술을 제공하는 LBS 서버기술, 그리고 다양한 LBS 응용기술들을 들 수 있다[1,2].

현재 수많은 스마트폰 어플리케이션에서 사용자의 위치 정보에 따른 서비스를 직간접적으로 제공하고 있다. 이러한 서비스의 대부분은 POI(Point Of Interest) 정보를 사용하고 있다. POI는 (1) 특정인이 관심을 가지는 현실 세계 또는 지도나 도면상의 특정 위치, (2) 차량 운전자가 쉽게 목표 지점을 찾을 수 있도록 제공하는 도로 주변 건물의 위치 정보라는 두 가지 의미를 가지고 있다[3]. POI를 표현하는 방법 중 가장 일반적인 것은 사람들에게 알려지거나 사람들이 많이 모이는 장소를 POI로 선정하는 것이다.

그다음은 GPS와 같은 위치 데이터에서 직접 POI를 추출하는 것이다. 이 두 방법을 혼합하여 POI를 표현하는 방법도 존재한다[4].

이동 객체의 궤적 데이터에서 의미 있는 장소를 찾는 방법은 Mean shift를 이용하여 거리에 기반을 둔 클러스터링[5], K-means을 통한 클러스터링[6]과 같은 것이 존재한다. 가장 일반적으로는 데이터 밀도에 기반하여 클러스터링을 수행하는 방법이 있다. 이 밀도 기반의 기법은 클러스터를 공간상에서 데이터의 밀도 영역에 의해 분리된 고밀도의 영역이라 가정을 한다. 즉 데이터간의 관계로 밀도를 이용해 클러스터와 클러스터에 속하지 않은 데이터를 식별하여 클러스터들을 분리하는 것이다. 이러한 방법을 통해 다양한 모양과 크기의 클러스터를 용이하게 추출할 수 있다. 그러나 밀도 기반 클러스터 방법은 클러스터 결정에 사용되는 파라미터의 선택에 따라 그 모양이 달라지고 밀도차이가 존재하는 그룹이 데이터 집합에 존재할 시에 상대적으로 밀도가 큰 그룹이 클러스터로 인식될 확률이 낮아지는 문제를 가지고 있다[7].

DBSCAN[8]은 밀도 기반 클러스터링의 대표적인 기법이다. 이 기법이 소개된 이후로 많은 연구에서 밀도에 기반한 클러스터링의 문제점을 해결하려고 노력하고 있다. 하지만 대부분의 연구에서는 데이터 집합에서 어떻게 클러스터를 생성하는지에 대한, 즉 클러스터 추출에 대한 고찰이 주 대상이다. 본 연구에서는 클러스터 추출 문제를 배제하고 GPS 데이터와 같이 이동 객체로부터 생성되는 위치 데이터 집합에서 POI를 추출하기 위한 클러스터 기법인 DBSCAN-D를 제안했다. 제안된 기법은 GPS 데이터와 같이 방향성 표현이 가능한 데이터 집합에서의 클러스터를 수행하기 위한 것이다.

2. 관련연구

2.1 DBSCAN

DBSCAN은 Ester, Kriegel, Sander, Xu[8]에 의해 최초 제안된 밀도기반 클러스터 기법이다. 이는 클러스터를 데이터 공간상에서 하위 개체(또는 벡터, 객체, 점, 데이터) 밀도 영역에 의해 분리된 고밀도의 영역이라 가정하고, 각 개체들의 위치정보와 주변 데이터들의 밀도를 이용해서 클러스터를 생성해 나간다. 이 방식은 금융 사기와 관련된 특징이나 구매 패턴이 유사한 고객의 특징을 분석하는 것과 같은 금융 및 마케팅 분야뿐만 아니라 교육 및 의료와 같은 다양한 분야에서 사용되고 있다[9].

DBSCAN은 주어진 데이터들을 벡터로 표현하고 벡터들간의 상대적인 관계인 밀도를 이용해 클러스터와 어느 클러스터에도 속하지 않은 데이터 점들인 잡음을 식별하여 클러스터들을 적절히 분리한다. 밀도영역에서 클러스터를 생성하기 위한 6개의 정의와 보조정리 2개를 설명하고 있다[8]. 이를 살펴보면 다음과 같다.

정의 1 (The Eps-neighborhood of a point) : 임의 한 점의 Eps-neighborhood는 그 점으로부터 Eps 반경 내에 있는 이웃의 집합이다. 여기서 EPS는 임의의 점의 밀도를 측정하기 위해 한 점을 중심으로 반경을 지정하는 파라미터이다.

$$N_{Eps}(p) = \{q \in D | dist(p,q) \leq Eps\} \quad (1)$$

식(1)과 같이 한 점 p의 Eps-neighborhood를 $N_{Eps}(p)$로 나타냈을 때 점 q가 데이터 집합 D에 속하고 점 p와 q의 거리 $dist(p,q)$가 Eps 반경보다 작거나 같다면 이때, '점 q는 점 p의 Eps-neighborhood이다'라고 정의할 수 있다.

정의 2 (directly density-reachable)

$$p \in N_{Eps}(q) \quad (2)$$

$$|N_{Eps}(q)| \geq MinPts \quad (3)$$

한 점 p가 점 q의 이웃 객체 집합에 속할 때 $N_{Eps}(q)$(식(2)), 식(3)과 같이 점 q의 Eps-neighborhood가 MinPts 보다 크거나 같다면 점 q를 중심객체라 할 수 있고, 이런 경우 '점 p는 점 q로부터 직접 밀도 도달 가능(directly density-reachable)하다'라고 객체 간의 관계를 정의 내릴 수 있다. (Fig. 1(a) 참조) 여기서 MinPts는 임의의 점의 밀도를 측정하기 위해 한 점을 중심으로 Eps 반경 내에 존재해야 하는 점의 최소 개수를 지정하는 변수이다. 즉, 어느 한 점이 중심객체가 되기 위한 이웃 객체의 최소 개수를 나타낸다.

정의 3 (density-reachable)

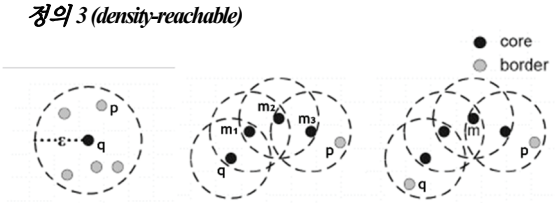

Fig. 1. The concepts (a) directly density reachability, (b) density reachability and (c) density connectedness to determine whether objects are density connected [10].

'한 점 p가 다른 한 점 q로부터 밀도 도달 가능(density-reachable)하다'라는 의미는 두 점 사이에 직접 밀도 도달 가능한 연결이 존재한다는 것이다. 그림1(b)를 예로 들어,

점 m_1이 점 q로부터 직접 밀도 도달 가능하고 m_2가 m_1으로부터, m_3가 m_2로부터, p가 m_3로부터 각각 직접 밀도 도달 가능하다면, 점 p는 점 q로부터 밀도 도달 가능한 것으로 정의할 수 있다. 단 주의할 점은 q로부터 p가 밀도 도달 가능하다고 해도 그 역의 성립은 보장할 수 없다.

정의 4 (density-connected) : '한 점 p와 다른 한 점 q가 밀도 연결(density-connected)하다'라는 의미는 어느 특정한 점을 기준으로 점 p와 q모두 밀도 도달 가능한 점임을 뜻한다. Fig. 2(c)를 예로 들어, 점 p는 m으로부터 밀도 도달 가능하고 점 q도 마찬가지로 점 m으로부터 밀도 도달 가능하다. 따라서 점 p는 점 q로부터 밀도 연결하고 그 반대도 성립한다.

정의 5 (cluster) : 임의의 점 p, q가 있을 때, 점 q가 p로부터 밀도 도달 가능하다면 점 q도 클러스터에 포함된다. 임의의 점 p, q가 클러스터에 속해있다면, 점 p와 q는 밀도 연결하다. 따라서 클러스터는 밀도 연결한 점들의 집합이라고 할 수 있다.

정의 6 (noise) : 앞서 살펴본 용어와 마찬가지로 클러스터에 포함되지 않는 점을 뜻한다. 부연설명을 하자면 다음과 같다. 데이터 집합 D에 한 개 또는 그 이상의 클러스터(C_i)가 존재한다면 noise point는 집합 D에는 속하지만 어떠한 클러스터에도 속하지 않는 점을 나타낸다. 이를 식으로 나타내면 다음과 같다.

$$\text{noise} = \{p \in D | \forall i : p \notin C_i\} \tag{4}$$

보조정리 1: 데이터 집합 D에 속하고 $|N_{Eps}(q)| \geq MinPts$ 을 만족하는 점을 p라 했을 때, $O = \{o | o \in D\}$이고 집합 O의 점 o들이 점 p로부터 밀도 도달 가능하다면 O는 하나의 클러스터가 된다.

보조정리 2 : 클러스터 C에 core point로 p가 있다면, 점 p로부터 밀도 도달 가능한 점들 o로 구성된 집합 O는 클러스터 C와 같다고 할 수 있다.

위에서 살펴본 바와 같이 DBSCAN은 주어진 데이터집합에서 noise point를 배제하고 core point, border point를 추출하는 방식을 통해 클러스터를 생성하여 다양한 모양과 크기의 클러스터를 구분하는데 용이하다.

2.2 DBSCAN 의 확장 연구

앞서 살펴본 DBSCAN 알고리즘은 클러스터 생성 장점에도 불구하고 클러스터 결정에 사용되는 Eps와 MinPts와 같은 파라미터의 선택에 따라 클러스터의 모양이 급격히 변화되고 밀도차가 존재하는 두 그룹에 있어서 밀도가 상대적으로 큰 클러스터 인식률이 떨어지는 문제점이 존재한다[7]. 1996년 DBSCAN이 제안된 후 VDBSCAN(Varied), EDBSCAN(Enhanced), IDBSCAN(Improved), FDBSCAN(Fast), GRIDBSCAN(Grid), KNNDBSCAN, ST-DBSCAN(Spatial-Temporal), GMDBSCAN(Grid and Multi-density)과 같은 다양한 DBSCAN 개선 연구들이 제안되었다. 본 연구에서는 다양한 DBSCAN 연구 중에 가중치를 고려한 연구를 살펴보았다.

2.2.1 DBSCAN-W

논문[11]에서는 클러스터를 생성하기 위해 데이터의 가중치까지 고려한 DBSCAN-W를 제안했다. DBSCAN에서는 공간에 표현되는 점이나 개체들 모두 위치 속성만을 갖는 것으로 표현되어 있어 각 데이터들이 갖는 중요도가 고려되지 않고 있는 반면에 DBSCAN-W는 데이터들의 위치 속성과 더불어 데이터들의 속성값을 클러스터링시에 고려한다. DBSCAN의 정의에 기반한 DBSCAN-W의 정의는 다음과 같다.

정의 1 : 대상 집합의 모든 개체는 해당 응용 시스템에서 그 개체가 갖는 중요도에 따라서 서로 다른 크기의 원으로 표현되는 영역을 갖는다. 즉, 개체를 공간상에 표현할 때 개체의 위치를 중심으로 속성값의 차이를 원의 반지름으로 표현한다. 따라서 개체는 속성값에 따라 서로 다른 크기의 원으로 표현된다.

정의 2 : 한 개체 p가 있을 때 Eps-neighborhood는 p의 중심점으로부터 Eps 반경 안에 각 개체를 표현하는 영역이 겹쳐지는 이웃들의 집합이다.

정의 3 : 클러스터는 밀도 연결된 영역의 최대 집합으로 표현된다.

DBSCAN-W는 개체가 갖는 속성을 표현하기 위해 다음과 같이 총 3단계의 전처리 과정을 수행한다. 첫째 개체의 속성, 즉 가중치를 부여할 비 공간 속성 A를 결정한다. 둘째, 적당한 변형 함수 $F(A_i) = r_i$를 통해 각 개체의 속성 크기로 사용될 반지름 값(r_i)을 결정한다. 셋째, 개체의 위치를 공간 상에 나타내고 2단계에서 결정된 반지름 r_i를 통해 각 개체를 원으로 표현한다.

DBSCAN-W의 클러스터 생성과정은 DBSCAN과 유사하다. 단 DBSCAN은 core point의 조건을 만족하는 점 p를

선택되면 이를 기준으로 Eps와 MinPts를 만족하는 모든 density-reachable한 점들을 찾아 하나의 클러스터로 결정하지만 DBSCAN-W은 점 p로부터 Eps와 MinPts를 만족하는 모든 density-reachable한 점들을 찾는 와중에 각 개체의 Eps-neighbor를 위에서 설명한 2번 정의와 같은 방법으로 결정한다. Fig. 2는 (a)와 같은 분포를 가진 개체들이 공간상에 위치했을 때 DBSCAN과 DBSCAN-W 각각이 Eps-neighborhood를 결정하는 과정을 비교한 것이다. Fig. 2(b)는 DBSCAN에서 점 p의 Eps-neighborhood를 구하는 과정을 나타낸 것으로 점 p의 Eps-neighborhood는 점 p를 중심으로 Eps 반경 내에 포함된 5개의 점들이다. Fig. 2(c)는 DBSCAN-W에서 Eps-neighborhood를 결정하는 과정으로 앞선 전처리 과정을 거쳐 각 점들이 서로 다른 크기의 원으로 표현된다. 이 그림에서 점 q의 경우 점 p의 중심 점과의 거리는 Eps 이상이지만, 점 p를 중심으로 한 Eps 반경 내에 q의 영역이 겹쳐지므로 점 q의 Eps-neighborhood에 속하게 된다. 이처럼 데이터가 가진 속성값의 차이에 따라 해당 데이터의 영역을 달리 표현함으로 주변 개체의 이웃으로 포함될 가능성을 높이고 그 중요한 데이터가 잡음으로 처리되는 확률을 줄어들게 만든다.

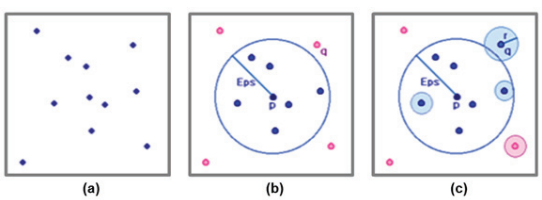

Fig. 2. Eps-neighborhood of object p in DBSCAN and DBSCAN-W[12]. (a) Distribution of objects, (b) Eps-neighborhood of object p in DBSCAN, and (c) Eps-neighborhood of object p in DBSCAN-W.

2.2.2 DBSCAN-SI

DBSCAN-SI는 이웃의 개수를 정하는데 기준이 되는 거리인 Eps 반경을 확장하여 영향력이 큰 값들은 주변 객체들의 이웃이 될 확률을 높여주는 방법(DBSCAN-SI(1))과 이웃들의 영향력 합으로 클러스터의 중심객체를 결정하는 방법(DBSCAN-SI(2))을 병행하는 알고리즘이다[13]. 이는 DBSCAN-W와 유사하게 객체의 속성값을 사용한다. 차이점은 DBSCAN-W는 하나의 대표 속성값을 통해 객체가 가진 영향력을 표현하지만 DBSCAN-SI는 하나 이상의 속성 값과 가중치를 고려한다. DBSCAN-SI(1)에서는 이웃 객체를 결정하기 위해 아래와 같이 2 개의 개념을 재정의 하였다.

정의 1 : 한 점 p의 이웃을 정하는 Eps'의 길이는 DBSCAN과 DBSCAN-W 알고리즘에서의 Eps에 중심점의 반지름(r_p)(객체 p가 가진 영향력)을 더한 값으로 한다.

$$Eps' = Eps + r_p \tag{5}$$

정의 2 : 한 점 p의 Eps'-neighborhood는 p로부터 반경 Eps' 내의 영역과 각 점들의 영향력으로 표현된 원들이 겹쳐지는 점들의 집합이다. 여기에서 $dist()$함수는 공간상의 유클리드 거리를 구하는 함수이다.

$$\begin{aligned} N_{Eps}(p) = \{q \in D | dist(p,q) \le (Eps + r_p + r_q)\} \\ - Eps: dist(p,q) \le Eps + r_p + r_q \\ - Eps': dist(p,q) \le Eps' + r_q \end{aligned} \tag{6}$$

Fig. 3(a)는 DBSCAN-W에서의 Eps를 보여주는 것으로 객체 p로 부터의 거리를 나타내고 있다. 이 그림에서는 중심객체 p를 기준으로 Eps 반경 내에 T, U, Y, V가 이웃 객체임을 알 수 있다. Fig. 3(b)는 Eps'로 객체 p가 갖는 영향력에 비례하여 Eps를 확장한 모습이다. 그 결과 Fig. 3(a)와 달리 T, U, Y, V외에 W, Q, S도 이웃 객체로 포함하고 있다. 즉 객체 W, Q, S는 Eps 반경 밖에 있지만, 중심 객체인 p가 가진 반지름 길이(영향력)에 비례하여 Eps를 확장(Eps')하였기 때문에 이웃 객체에 포함될 수 있다.

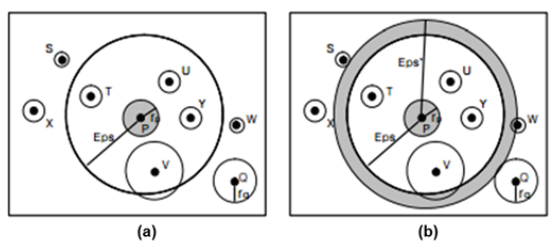

Fig. 3. Eps extension[12]. (a) Eps in DBSCAN-W, (b) Eps in DBSCAN-SI(1).

이웃 객체의 수에 의해 중심 객체가 결정되는 DBSCAN과 달리 DBSCAN-SI(2)는 객체가 갖는 많은 속성들의 영향력을 값으로 나타내고, 이의 합을 중심 객체 결정에 사용한다. 즉, 임의의 한 점이 있을 때 그 이웃들의 영향력 합이 설정된 기준치 이상이라면 그 점을 중심 객체로 결정하는 것이다. 따라서 만일 한 점을 기준으로 이웃 객체의 수가 적더라도 그 이웃들의 영향력 합이 설정된 기준치 이상이라면 그 점이 중심 객체가 된다. 이를 위해 DBSCAN 정의에 더하여 다음과 같은 정의를 나타냈다.

정의 3 : MinPts는 최소이웃수를 의미하고, MinInf는 최소 이웃의 영향력의 합을 의미한다. 이때 한 점의 Eps'-neighborhood의 수가 MinPts 이상인 경우이거나, 최소이웃의 영향력의 합이 MinInf 이상인 경우 이점을 중심객체라 한다. 여기에 자신의 영향력 값도 포함한다.

Fig. 4는 중심 객체와 Eps 반경 내 포함된 이웃 객체의 영향력 값을 나타낸 것이다. 이 그림에서 (a)는 점 p를 기준으로 Eps 반경 내의 이웃으로 4개의 객체를 포함하고 있다. 여기서 그 이웃 객체들과 중심점의 영향력 합은 18인 값을 가진다. 그리고 (b)의 경우는 반경 내 이웃 객체의 수가 3이고, 중심 객체를 포함한 이웃 객체들의 영향력 합이 32이다. 여기에서 만일 중심 객체의 조건이 이웃 객체 수가 4이상이거나 이웃 객체의 영향력 값의 합이 30 이상인 경우라면 Fig. 4의 (b)의 점 p도 중심객체가 된다. 이는 이웃 객체의 수가 MinPts 보다 작아도, 객체들이 비중 있는(영향력 값이 큰) 경우라면 중심객체가 되게 하여 클러스터에 포함되도록 하는 효과가 있다. 또한 MinPts를 최대화하는 것처럼 이웃 객체 수의 제한을 걸지 않는 경우에도 이웃 객체의 영향력 합을 통해 중심객체를 선정할 수 있다.

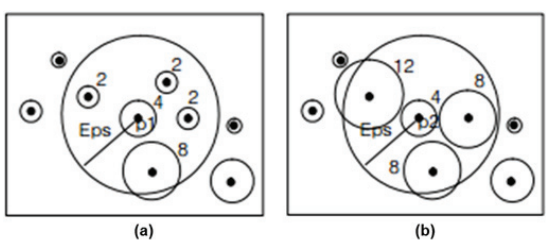

Fig. 4. DBSCAN-SI[12]. (a) core point in DBSCAN-W (MinPts = 4), (b) core point in DBSCAN-SI(MinPts = 3).

3. 방향을 고려한 밀도 기반 클러스터링

3.1 DBSCAN-D

본 연구에서 제안하는 DBSCAN-D는 관련 연구에서 살펴본 DBSCAN-W와 DBSCAN-SI와 같이 DBSCAN에 기반을 두고 있다. 이 세 알고리즘은 데이터 특성이나 도메인에 상관없이 범용적으로 특정 데이터 집합이 주어졌을 때 그 집합에서 각자의 정의에 따라 중심 객체를 기준으로 밀도 연결한 관계를 찾아 클러스터링을 수행한다. 본 연구에서 제안하는 알고리즘은 DBSCAN에 그 기반을 두고 있지만 이들과 다르게 공간 데이터를 다루는 위치기반 서비스 도메인을 적용 대상으로 삼고 있다. 그 이유는 DBSCAN-D라는 이름에서도 알 수 있듯이 방향성을 속성으로 표현할 수 있는 데이터의 집합을 클러스터링의 고려대상으로 삼고 있기 때문이다. 따라서 본 연구에서는 GPS 데이터를 대상으로 클러스터링을 수행하는 알고리즘을 제안했다.

DBSCAN-D는 데이터들이 갖는 중요도를 고려하지 않는 DBSCAN을 확장한 알고리즘들과 유사하게 클러스터링을 수행하기 전 데이터의 크기를 결정하기 위해 데이터가 가진 속성을 분석하는 작업이 선행된다. 이 선행 작업은 다음과 같은 정의를 따른다.

정의 1 : 시간 속성을 가지고 순차적으로 발생하는 데이터의 집합에서 데이터는 데이터 간의 시간차를 이용해 영향력 또는 가중치를 달리 표현할 수 있다.

데이터 크기는 GPS와 같은 위치 데이터 집합을 대상으로 속성을 추출하고 각각의 데이터가 가진 중요도 또는 가치를 그들의 크기로 표현한다. GPS 데이터는 경/위도(Longitude/Latitude)와 같은 기본적인 위치 정보 외에 시간(UTC Time), 방위(N/S, E/W Indicator), 위성수(Satelites Used), 고도(Altitude)와 같은 부가적인 정보로 표현된다. 제안된 알고리즘에서는 데이터의 크기를 결정하기 위해 GPS 수신 데이터 간의 수신 시간(UTC Time) 차이를 이용한다. 즉 한 공간상의 특정 지점에 머무른 시간을 그 지점의 크기로 사용한다. GPS 데이터는 순차적으로 데이터가 쌓이기 때문에 한 지점과 그 다음 지점이 수신된 시간을 알면 그 차이를 구해 한 지점에서 머무른 시간으로 사용할 수 있다. 따라서 GPS 데이터 집합에서 마지막으로 수신된 위치를 제외하고 모든 위치에서 머무른 시간을 알 수 있다. 이를 그림으로 나타내면 Fig. 3(c)와 같이 다양한 크기를 가진 형태로 표현될 수 있다.

DBSCAN-D에서 클러스터링을 위해 사용되는 데이터 속성 중 또 다른 하나는 바로 방향성이다. 이는 GPS 데이터가 가진 고유 속성인 방위와는 차이가 있다. 이 방향성은 앞서 설명한 크기를 결정하는 것과 마찬가지로 두 지점간의 관계를 통해 찾을 수 있다. 이를 정의로 나타내면 다음과 같다.

정의 2 : 시간 속성을 가지고 순차적으로 발생하는 데이터의 집합에서 점 p가 점 q보다 이전에 발생된 점이라면 점 q의 영향력의 위상은 점 p의 위치로 최대한 가까이 이동한다. 이때 이동의 범위는 q의 영향력 위상이 q의 좌표값을 벗어나지 않는 최대구간까지이다.

Fig. 5(a)는 GPS 데이터 집합으로 화살표를 통해 일부 데

이터가 생성된 순서를 나타내었다. Fig. 5(b)는 DBSCAN-W의 정의에 의해 각 데이터가 가진 가중치, 여기서는 그 지점에 머무른 시간에 비례하여 데이터의 크기를 나타낸 것이다. Fig. 5(c)는 방향성을 적용하여 데이터의 크기로 표현되는 영역을 이동시킨 것이다. 여기서 이동은 정의 2에 따라 한 점을 기준으로 그 점 이후에 발생된 점들의 크기 영역을 기준이 되는 점 방향으로 원의 둘레에 점의 위치가 놓일 때까지 이동을 시킨 모습이다. 따라서 만일 중심객체의 조건이 Eps 반경 내에 MinPts가 5이상이라면 Fig. 5(b)는 반경내 포함된 이웃개체가 4개로 중심객체가 될 수 없다. 이와는 다르게 5(c)는 정의 2에 따라 기준 점 p이후에 생성된 데이터들의 영향력 위상이 p위치 쪽으로 기울어져 있어 (b)와 다르게 q점이 Eps 반경 내에 들어가게 된다. 따라서 이웃객체 수가 5개로 점 p는 중심객체가 된다. 이처럼 데이터가 가진 속성값의 차이에 따라 해당 데이터의 영역을 달리 표현하고 방향성을 통해 주변 개체의 이웃으로 포함될 가능성을 높여 주요 데이터가 잡음으로 처리되는 경우를 줄어들게 만들었다.

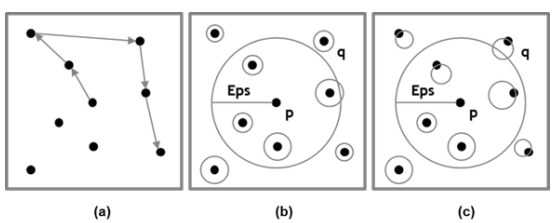

Fig. 5. Eps-neighborhood of object p in DBSCAN-D. (a) Distribution of objects, (b) Weight area according to the staying time of objects in DBSCAN-D, and (c) Directional representation of objects in DBSCAN-D.

4. 결 론

본 연구는 기존의 밀도 기반 클러스터링 연구에 기반하여 GPS 데이터와 같이 이동 객체에 의해 생성되는 위치 데이터를 대상으로 이에 적합한 POI를 찾기 위한 클러스터링 기법인 DBSCAN-D를 제안했다. 이 방법은 GPS 데이터간의 관계를 통해 추출된 머무른 시간과 방향성을 기본으로 설계되었다. 머무른 시간은 GPS 데이터가 수신된 시간을 이용하여 데이터간의 수신 시간차이를 통해 추출될 수 있고 방향성은 GPS 데이터가 순차적으로 생성되는 점에 집중하여 이전 생성된 데이터의 위치 방향으로 이후 생성된 데이터의 영역(머무른 시간에 비례)이 이동함으로 표현할 수 있다. 이와 같은 두 속성을 통해 이동 객체에 의해 생성된 데이터 집합에 적합한 클러스터링을 수행할 수 있다. 따라서 이 제안된 기법은 기존의 클러스터링 기법에 비해 범용성은 떨어지지만 반대로 위치기반 서비스 도메인의 특성을 잘 반영할 수 있을 것으로 기대된다. 본 연구는 제안 연구로 실제 위치 데이터에 적용하여 제안된 기법의 유용성을 제시하고 특히 방향성에 대해 좀 더 이론적인 배경을 통해 정의할 의무가 남아 있다.

참고문헌

1. H. O. Choi, “LBS, Location-Based Services,” TTA Journal, vol. 86, pp. 59-69, 2003.
2. Yong-hwan, L,. (2020) Image Clustering using Geo-Location Awareness, Journal of the Semiconductor & Display Technology, 19(4), 135-138.
3. G. W. Lee and H. U. Son, Glossary of Geo Spatial Information System, 1th ed, Goomi Seokuan, Seoul, 2016.
4. Y. K. Heo, J. S. Oh, P. Paudel, P. Thapa, H. J. Jeon, M. A. Jeong, and S. R. Lee, “Density Based system for Recommendation of Hybrid POI,” Proceeding of the Conference of the Institute of Electronics Engineers of Korea, pp, 1318-1322, 2015.
5. S. Khetarpaul, R. Chauhan, S. K. Gupta, L. V. Subramaniam, and U. Nambiar, “Mining GPS data to determine interesting locations,” Proceeding of the 8th International Workshop on Information Integration on the Web: in conjunction with WWW 2011, ACM, p. 8, 2011.
6. A. J. Dou, V. Kalogeraki, D. Gunopulos, T. Mielikinen, V. Tuulos, S. Foley, and C. Yu, “Data clustering on a network of mobile smartphones,” Proceeding of the IEEE/IPSJ 11th International Symposium, IEEE, pp. 118-127, 2011.
7. A. Kirmse, T. Udeshi, P. Bellver, and J. Shuma, “Extracting patterns from location history,” Proceeding of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, pp. 397-400, 2011.
8. M. Ester, H. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large geospatial databases with noise,” Proceeding of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, pp. 226–231, 1996.
9. K. Santhisree, A. Damodaram, S. V. Appaji, and D. NagarjunaDevi, “Web usage data clustering using DBSCAN algorithm and set similarities,” Proceeding of the Data Storage and Data Engineering (DSDE) 2010 International Conference, IEEE, pp. 220-224, 2010.
10. N. Schlitter, T. Falkowski, and J. Lässig, “Dengraph-ho:

Density-based hierarchical community detection for explorative visual network analysis.” Research and Development in Intelligent Systems XXVIII. Springer London, pp. 283-296, 2011.
11. H. S. Kim, H. S. Lim, and H. S. Yong, “Design and development of the clustering algorithm considering weight in spatial data mining,” Journal of Intelligence and Information Systems, Vol. 8, No. 2, pp. 177-187, 2002.
12. H. S. Lim, A Density-based Spatial Clustering Algorithm Considering Weight and Obstructed Distance, Master’s Thesis of Ewha Institute of Science and Technology, 2002.
13. B. C. Kim, “Design and Development of Clustering Algorithm Considering Influences of Spatial Objects,” Journal of The Korea Contents Association, Vol. 6, No. 12, pp. 113-120, 2006.

접수일: 2024년 5월 15일, 심사일: 2024년 6월 17일, 게재확정일: 2024년 6월 21일