한국지능정보시스템학회논문지 제8권 제2호 2002년 12월(pp.177~187)

공간 데이터 마이닝에서 가중치를 고려한 클러스터링 알고리즘의 설계와 구현

김호숙
동의 공업대학 컴퓨터 정보계열 전자계산과
(hskim@dit.ac.kr)

임현숙
LG CNS
(hyunslim@lgcns.com)

용환승
이화여자 대학교 컴퓨터학과
(hsyong@ewha.ac.kr)

공간 데이터 마이닝이란 공간 데이터베이스 내에 함축적으로 존재하는 흥미 있는 관계와 특징을 발견하는 과정이다. 많은 공간 클러스터링 알고리즘이 개발 되었으나, 공간 속성을 기준으로 클러스터링을 수행하면서 동시에 오브젝트의 비 공간적 속성에 대하여 가중치를 부여하는 방법에 대한 연구는 부족하였다. 본 논문은 새로운 공간 클러스터링 알고리즘인 DBSCAN-W를 제안하였다. DBSCAN-W는 밀도 기반 클러스터링 알고리즘인 DBSCAN을 확장한 알고리즘이다. 기존의 DBSCAN에서는 클러스터링을 위해 오브젝트의 위치 속성만을 고려한 반면, DBSCAN-W는 오브젝트의 위치 속성 뿐 아니라 주어진 응용과 관련된 오브젝트의 비 공간 속성들을 함께 고려한다. DBSCAN-W에서 각 오브젝트들은 다양한 크기의 원으로 표현되는 영역을 갖는다. 이때 원의 반지름은 해당 응용 시스템에서 오브젝트가 갖는 중요도를 반영한다. 또한 실험을 통하여 DBSCAN-W 알고리즘이 사용자의 의도를 반영한 다양한 클러스터를 효과적으로 생성하는 결과를 보였다.

1. 서론

클러스터링이란 실제적 또는 추상적인 오브젝트의 집합을 클러스터라는 서로 유사성을 갖는 오브젝트의 분류로 그룹화하는 과정이다. 하나의 클러스터에 속하는 데이터 점들 간에는 서로 다른 클러스터 내의 점들과는 구분되는 유사성을 갖게 된다[1]. 클러스터링은 데이터 마이닝 뿐 아니라 통계학, 생물학, 기계 학습 등으로부터 발전해 왔으며, 최근 효과적으로 대량의 데이터를 취급할 수 있는 많은 방법들이 개발되었다. 클러스터링 방법은 크게 분할(partitioning) 방법[2]과 계층적(hierarchical) 방법[3,4], 밀도 기반(density-based) 방법[5,6], 격자기반(grid-based) 방법[7]과 모델 기반(model-based)방법으로 나눌 수 있다[8]. 많은 데이터 마이닝 알고리즘은 관련된 응용 시스템의 목적에 따라 가중치를 고려한다. 예를 들어서, 텍스트 클러스터링은 공통된 단어를 많이 갖는 문서를 찾아 같은 그룹으로 분류하는 작업이다. 이때 문서의 분류 시 중요한 의미를 갖는 중심어는 다른 단어에 비해 높은 가중치를 갖도록 처리된다[8][9].

공간 클러스터링은 공간 데이터베이스 내에서 오브젝트 간의 거리, 연결성, 밀도를 기반으로 유

사한 오브젝트들을 그룹화 하는 것이다. 지금까지 많은 공간 클러스터링 알고리즘이 개발되었다. 그러나 오브젝트의 공간적인 속성을 고려하면서 동시에 응용 시스템에서 중요하게 취급되는 비공간 속성에 가중치를 두는 클러스터링 방법에 대한 연구는 부족하였다. 그러나 많은 실제적인 응용에서는 아래의 예와 같이 대상물의 공간적 위치 속성을 기반으로 클러스터링을 수행하면서 동시에 응용 프로그램에서 중요한 의미를 갖는 비 공간 속성에 가중치를 부여하는 알고리즘을 필요로 한다.

예 1 : 한 은행의 관리자가 새로운 automatic teller machine (ATM)을 설치하기 위해서 대상 지역에 속해있는 고객을 클러스터링 하기 원한다. 이때 기계의 위치는 해당 은행을 자주 이용하는 고객들의 위치와 가까운 곳이 바람직하며 이를 위해 클러스터링 시 각 고객의 평균 이용 횟수에 가중치를 두기 원한다.

위와 같은 실제 응용에서의 클러스터링 결과는 공간상에 분포한 데이터에 대하여 다양한 모양과 크기의 클러스터를 구분해 내어야 하고, 클러스터의 구성 조건을 만족하지 못하는 데이터는 잡음(noise)으로 처리되어야 하며, 주어진 가중치를 반영할 수 있어야 한다. 즉 예1에서 고객들의 주소는 공간 속성으로서 공간 클러스터링 알고리즘에서 거리 측정의 기준으로 사용되는 속성이 되고, 평균 이용 횟수는 비공간 속성으로 클러스터링의 가중치로 적용된다.

본 논문에서는 새로운 공간 클러스터링 알고리즘 DBSCAN-W를 제안한다. DBSCAN-W는 밀도기반의 클러스터링 알고리즘인 DBSCAN을 확장한 것이다. DBSCAN에서는 모든 오브젝트의 위치 속성만이 고려된 반면, DBSCAN-W는 각 대상물의 위치 뿐 아니라 응용 분야와 관련된 비 공간 속성을 고려한다. 즉 DBSCAN에서는 모든 오브젝트가 동일한 중요도를 갖는 점으로 표시되는 반면, DBSCAN-W에서의 각각의 오브젝트는 응용 프로그램에서 가중치의 기준으로 선택된 속성 값에 따라서 반지름 길이가 결정되는 서로 크기가 다른 원으로 표현된다. 또한 실험을 통해 DBSCAN-W가 다른 밀도 기반 알고리즘에서 발견하지 못했던 새로운 형태의 클러스터를 생성하는 결과를 보인다.

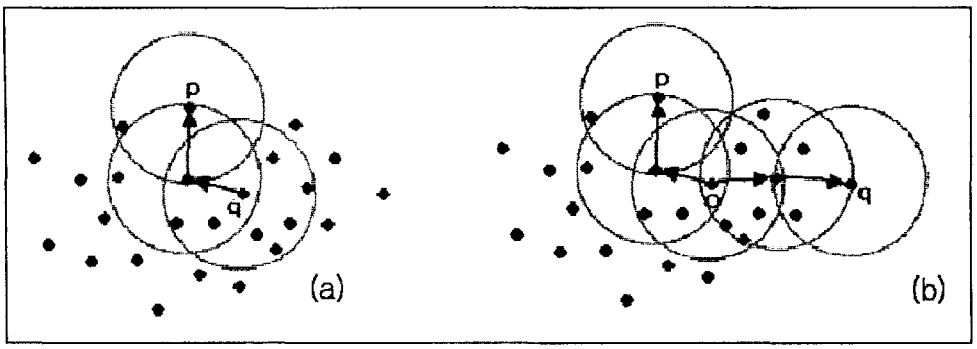
논문의 구성은 다음과 같다. 2장에서는 밀도를 기반으로 하는 공간 클러스터링 알고리즘인 DBSCAN을 소개하고, 3장에서는 DBSCAN-W에서 가중치를 고려하는 방법을 제안하며, 4장에서는 제안된 방법에 대한 구현 결과를 보여준다. 마지막으로 5장에서는 결론 및 향후 연구 과제를 제시한다.

2. DBSCAN : 밀도 기반 공간 클러스터링 알고리즘[5]

DBSCAN(Density Based Spatial Clustering of Applications with Noise)[5]은 밀도 기반 클러스터링 알고리즘으로 잡음을 포함한 공간 데이터베이스 내에서 다양한 모양의 클러스터를 발견해 낼 수 있다. DBSCAN에서는 다음의 몇 가지를 정의한다.

한 점 p의 Eps-neighborhood는 p로부터 반경 Eps내에 있는 이웃(neighborhood)의 집합이다.

MinPts(minimum number of points)는 최소 이웃수로 한 점의 Eps-neighborhood의 수가 최소 이웃수 이상인 경우 이 점을 core object라 한다.



<그림 1> Density-reachability와 density connectivity
p 는 q로부터 density-reachable 하다, o, p, q는 모두 density-connected 하다

주어진 점들의 집합 D에서 한 점 p가 q로부터 directly density-reachable하다는 의미는 p가 점 q의 Eps-neighborhood 내에 있고 q가 core object인 경우이다.

한 점 p가 점 q로부터 density-reachable 하다는 의미는 p에서 시작하고 q에서 끝나는 directly density-reachable 인 연결(chain)이 존재한다는 의미이다.

한 점 p가 점 q로부터 density-connected 하다는 의미는 p와 q로부터 density-reachable한 점 o가 존재한다는 의미이다.

클러스터는 밀집-연결된 (density-connected) 점들의 최대 집합이다.

DBSCAN은 대상 오브젝트들을 1번씩 읽으면서 오브젝트로부터 반경 Eps내의 오브젝트들을 찾아서 하나의 클러스터로 구분하는 과정을 수행한다. 이때 클러스터에 속하지 못하는 오브젝트들은 잡음으로 취급된다. 그러므로 한 오브젝트의 Eps 반경 내에 속하는 이웃을 찾아내는 겹침(overlap) 연산을 log n 시간에 수행할 수 있도록 지원하는 R*-tree 인덱스를 이용하는 경우 DBSCAN 알고리즘에서 대상 데이터의 수가 n 일때 전체 클러스터링을 수행하는 시간 복잡도는 $O(n * \log n)$이다.

3. 가중치를 고려한 밀도 기반의 공간 클러스터링 알고리즘 설계

3장에서는 밀도 기반의 공간 클러스터링 알고리즘에 가중치를 고려하기 위해서 기존의 DBSCAN을 확장한 DBSCAN-W (a DBSCAN algorithm using region expressed as Weight)를 제안한다. DBSCAN에서 모든 오브젝트들은 위치 속성만을 갖는 점으로 표현되고 각 점들이 갖는 중요도는 고려되지 못했다. 반면, DBSCAN-W는 각 대상물의 위치 속성 뿐 아니라 판매 시스템에서의 고객의 구매 총액과 같이 응용 시스템의 분석 시 필요한 비 공간 속성을 클러스터링 시 고려한다. 이를 위해 DBSCAN-W에서는 다음의 몇 가지 개념을 재정의 한다.

정의 1. 모든 오브젝트는 해당 응용 시스템에서 그 오브젝트가 갖는 중요도에 따라서 서로 다른

크기의 원으로 표현되는 영역을 갖는다. 이때 원의 중심 좌표값은 공간상에 오브젝트의 위치 좌표이고, 반지름은 가중치로 선택된 비공간 속성에 의해 결정된다.

정의 2. 한 오브젝트 p의 Eps-neighborhood는 p의 중심점으로부터 반경 Eps 안에 각 오브젝트를 표현하는 영역이 겹쳐지는 이웃들의 집합이다.

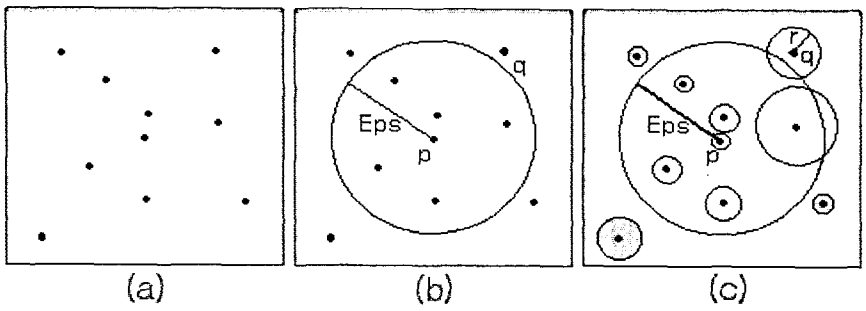
정의 3. 클러스터는 밀집-연결된 영역의 최대 집합(maximal set of density-connected regions)이다.

DBSCAN-W를 수행하기 위해서 다음과 같은 3단계의 전 처리 과정이 필요하다. 첫째 응용 프로그램의 의도를 반영하기 위하여 가중치를 부여할 비 공간 속성 A를 결정한다. 둘째, 적당한 변형 함수 $F(A_i) = r_i$을 이용하여 각 오브젝트의 반지름 값(r_i)을 결정한다. 셋째 각 오브젝트를 원 타입으로 표현한다. 이때 원의 중심은 공간 상의 위치를 나타내고, 반지름은 2단계에서 결정된 반지름 값(r_i)을 갖는다.

DBSCAN-W가 응용 시스템에서 요구하는 사용자의 의도를 올바르게 클러스터링 결과에 반영시키기 위해서는 각 오브젝트에 적절한 가중치 값을 부여하는 전처리 2단계가 중요한 의미를 갖는다. 이를 위해서 변형 함수 F는 응용 시스템의 특성을 잘 이해하고 있는 전문가에 의해서 결정되어야 하는데 이를 위해서 시스템에서 제공하는 방법은 다음과 같다. 첫째 방법은 사용자로부터 정규화의 범위 N을 입력받아 수치 데이터를 [0, N] 범위로 정규화 하여 이를 가중치 값으로 이용하는 방법이다. 예를 들어서 고객의 평균 거래 금액을 [0,60]의 범위로 정규화하고 그 결과 값을 각 고객의 가중치로 이용한다. 두번째 방법은 수치 데이터에 대해 적용하지만 해당 가중치를 적용하는 범위를 사용자로부터 직접 받아서 결정하는 경우이다. 예를 들어서 고객 평균 거래 금액이 10만원 미만이면 가중치를 1로, 10만원에서 100만원까지는 가중치를 2로 결정하는 등과 같이 가중치의 대상 범위를 사용자로부터 입력받아서 결정할 수 있다. 이 경우는 해당 시스템의 특성에 대해 잘 아는 전문가에 의한 입력 파라메터 설정이 반드시 요구된다. 세번째 방법은 수치 데이터 외의 문자 데이터나 분류 데이터에 대해서 가중치를 적용하는 방법이다. 예를 들어 고객 클래스가 우수 고객이면 가중치를 10으로, 문제 고객이면 가중치를 1로 하는 등의 적용이 그 예이다.

DBSCAN 알고리즘에서 클러스터를 생성하는 과정은 다음과 같다. 먼저 Core-point의 조건을 만족하는 점 p를 선택하여 seed로 잡은 후, 선택한 seed로부터 밀도를 결정하는 범용 파라메터인 Eps와 MinPts를 만족하는 모든 density-reachable한 점들을 찾아 하나의 클러스터로 결정한다. 동일한 방법으로 DBSCAN-W에서도 클러스터를 생성하기 위하여 core-point의 조건을 만족하는 점 p로부터 Eps와 MinPts를 만족하는 모든 density-reachable한 점들을 찾는다. 이를 위해서 각 오브젝트의 Eps-neighbor를 정의2와 같은 방법으로 결정한다.

<그림 2>는 기존의 DBSCAN 알고리즘과 DBSCAN-W에서 오브젝트의 Eps-neighborhood를 결정하는 과정을 비교한 것이다. <그림 2> (a)는 점으로 표시된 오브젝트의 분포를 나타낸 것이다. <그림 2> (b)는 DBSCAN에서 오브젝트 p의 Eps-neighborhood를 구하는 과정이다. 오브젝트 p의 Eps-neighborhood는 오브젝트 p를 중심으로 반경 Eps 내에 포함된 5개의 오브젝트들이다. <그림 2> (c)는 DBSCAN-W에서 Eps-



<그림 2> DBSCAN과 DBSCAN-W에서 한 오브젝트p 의 Eps-neighborhood 구하기
(a) 오브젝트들의 분포
(b) DBSCAN에서 오브젝트 p의 Eps-neighborhood 구하기
(c) DBSCAN-W에서 오브젝트 p의 Eps-neighborhood 구하기

neighborhood를 결정하는 과정으로 이때 각 오브젝트들은 서로 다른 크기의 원으로 표현된다. 즉 중심점은 오브젝트의 공간 속성인 위치 값(x,y)이고, 반지름 r은 응용 프로그램의 목적에 따라 사용자가 가중치의 기준으로 선택한 비 공간 속성에 의해서 결정된 값이다. 예를 들어 클러스터링을 수행할 때 가중치의 기준을 구매 총액으로 하는 경우, 구매 총액이 많을수록 높은 중요도를 갖고, 보다 커다란 크기의 원으로 표현된다. <그림 2> (c)에서 오브젝트 q의 경우 오브젝트 p의 중심점과의 거리는 Eps 이상이지만, 점 p를 중심으로 한 반경 Eps 내에 q의 영역이 겹쳐지므로 오브젝트 p의 Eps-neighborhood에 속하게 된다. 이러한 방법은 높은 가중치를 가진 오브젝트를 넓은 영역을 갖는 원으로 표현하여, 주변의 오브젝트의 이웃으로 포함될 가능성을 높이고, 그 결과 잡음으로 처리되는 확률을 줄어들게 만든다.

4. 구현 결과

본 장에서는 3장에서 제안한 DBSCAN-W의 성능을 평가하기 위하여 예1에서 제시한 은행의 ATM 설치를 위한 클러스터링을 수행한 결과를 DBSCAN과 비교한다. 시스템 구현 환경은 다음과 같다. 실험을 위하여 작성된 인위적인(synthetic) 데이터베이스는 Sun사의 Solaris 2.6 기반의 객체-관계 DBMS인 Informix Universal Server를 사용하였고 데이터베이스로부터 공간 데이터를 효율적으로 접근하게 하는 공간 데이터베이스 엔진으로 사용한 Informix Spatial Datablade Module은 2차원 연산을 지원하는 9개의 데이터 타입과 여러 가지 공간 함수를 제공하며 공간 색인을 위해 R-Tree를 지원한다[10,11].

실험에서 사용한 고객 테이블은 <표 1>과 같다.

실험에 사용한 고객 테이블은 7개의 속성을 갖는 총 500개의 데이터로 구성된다. Customer_id 필드는 각 고객 오브젝트를 구분하는 키 값이다. Location 필드는 점 형태로 (x, y) 좌표 값이 각각 1~1000의 범위를 갖는다. Region 필드는 location 필드를 중심점으로 하고, Weight 필드의 값을 반지름으로 하는 원 type 이다.

<표 1> 고객 테이블 정의

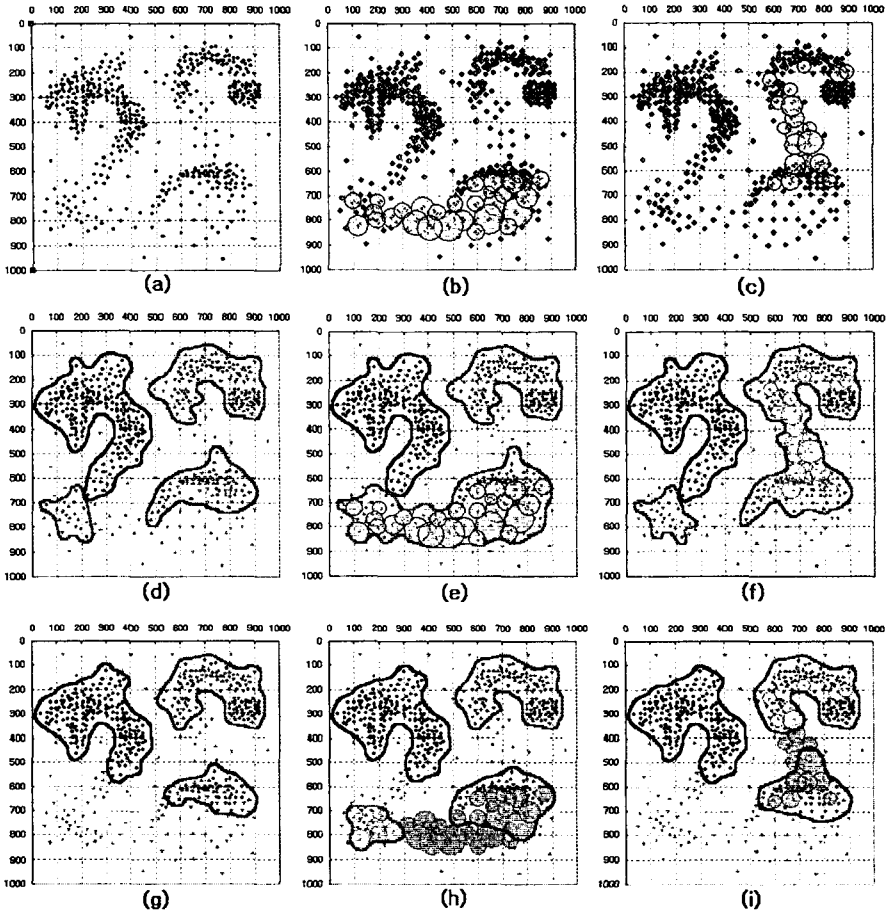
```
Create table Customer
( Customer_id      smallint not null,  /*고객 ID */
ocation        sp2Pnt,    /*고객의 위치 속성을 2차원 좌표값으로 나타낸 값 (x, y) */
Region         sp2Circ,   /* 가중치를 반지름으로 반영한 원 타입의 영역 (x, y, r) */
Avg_trade_frequency smallint, /* 평균 방문 횟수 */
Avg_trade_amount  smallint, /* 평균 거래 금액 */
Weight         smallint,  /* 비공간 속성에 의해 계산된 가중치 값 */
Cluster_id       smallint  /* 클러스터링 결과 */
);
```

Avg_trade_frequency와 Avg_trade_amount 필드는 고객의 평균 방문 횟수와 평균 거래 금액으로 각각 고객 데이터의 비 공간 속성이다. DBSCAN-W는 클러스터링을 수행하는 응용 시스템의 사용 목적에 따라서 오브젝트의 비 공간 속성 중에서 한 속성을 선택하여 클러스터링의 수행 시 가중치의 기준으로 삼는다. Weight 필드는 사용자가 가중치의 기준으로 선택한 비공간 속성을 변형 함수 F를 이용하여 계산한 결과로 Region 필드의 반지름으로 적용되는 값이다. 마지막으로 Cluster_id 필드는 클러스터링 수행 결과 결정된 클러스터 결과를 저장한다.

본 실험에서 사용한 입력 파라메터는 기준 반경 Eps과 최소 이웃 수 MinPts 그리고 가중치의 기준이 되는 속성 필드 3가지이며 전 처리 과정에서 사용한 변형 함수 F는 가중치의 기준으로 선택된 비 공간 속성의 값을 [0,60] 사이의 값으로 정규화하여 그 결과를 이용하였다. 입력 변수를 다양하게 변경하면서 기존의 DBSCAN과 DBSCAN-W를 비교한 결과는 <그림 3>과 같다.

(a) Location으로 나타낸 데이터의 분포
(b) Avg_trade_frequency를 기준으로 가중치를 계산하여 Region을 표시한 데이터의 분포
(c) Avg_trade_amount를 기준으로 가중치를 계산하여 Region을 표시한 데이터의 분포
(d) Eps =50, MinPts =5 적용 시 DBSCAN의 결과
(e) Eps =50, MinPts =5, 가중치 기준= Avg_trade_frequency 적용 시 DBSCAN-W의 결과
(f) Eps =50, MinPts =5, 가중치 기준= Avg_trade_amount 적용 시 DBSCAN-W의 결과
(g) Eps =50, MinPts =10 적용 시 DBSCAN의 결과
(h) Eps =50, MinPts =10, 가중치 기준= Avg_trade_frequency 적용 시 DBSCAN-W의 결과
Eps =50, MinPts =10, 가중치 기준= Avg_trade_amount 적용 시 DBSCAN-W의 결과

<그림 3> (a)는 2차원 좌표상에 500개의 고객 데이터의 위치가 분포한 형태이다. <그림 3> (b)와 (c)는 고객 테이블의 두 가지 비 공간 속성인 Avg_trade_frequency와 Avg_trade_amount를 각각 가중치의 기준으로 설정하고 그 값을 region 필드의 반지름 크기로 적용하여 오브젝트를 원으로 표현한 결과이다. 실험에서 사용한 데이터의 경우 하단에 위치한 고객들은 평균 방문 횟수가 많은 고객들로서 그 결과는 <그림 3> (b)와 같다. 또한 우측에 위치한 고객들은 평균 거래 금액이 많은 고객들로서 그 결과는 <그림 3> (c)와



<그림 3> 실험 결과

같다. <그림 3> (d)는 입력 파라메터로 Eps =50, MinPts =5 을 적용 시 DBSCAN 방법에 의해 클러스터링을 수행한 결과이다. 모두 4개의 클러스터가 구성되고, 나머지는 noise로 남겨진 것을 볼 수 있다. <그림 3> (e)는 Eps =50, MinPts =5 이고 고객의 평균 방문 횟수 속성을 가중치의 기준으로 할 때 DBSCAN-W를 수행한 결과이다. 실험에서 사용한 데이터의 하단에 위치한 고객들은 평균 방문 횟수가 많은 고객들로서 그 결과는 <그림 3> (d)에서와 같이 DBSCAN에서 우측 하단과 좌측 하단에 분리되어 존재하던 두개의 클러스터와 그 주위의 잡음으로 취급되었던 많은 오브젝트들이 하나의 큰 클러스터로 통합된 것을 보여준다. <그림 3> (f)는 Eps =50, MinPts =5 이고 고객의 평균 거래 금액 속성을 가중치의 기준으로 할 때 DBSCAN-W를 수행한 결과이다. 그 결과 우측 상단과 우측 하단의 두개의 클러스터가 주위의 잡음들을 포함하는 하나의 큰 클러스터로 통합된 결과를 보여준다. <그림 3> (g)~(i)는 클러스터의 밀도를 결정하는 입력 파라메

터인 기준 반경과 최소 이웃 수를 변경하여 DBSCAN과 DBSCAN- W에 각기 적용한 결과이다. MinPts 값이 증가함에 따라 클러스터를 구성하는 조건이 강화되어서 <그림 3> (d)~(f)에 비해서 작은 클러스터가 형성된 결과를 볼 수 있다. 또한 기준이 되는 가중치가 변화 함에 따라 클러스터 모양이 변경되는 것을 보여준다.

실험 결과 동일한 위치에 존재하는 오브젝트에 대한 공간 클러스터링도 적용되는 가중치에 따라 서로 다른 모양의 클러스터를 생성해 내는 것을 볼 수 있다. 또한 밀도를 결정하는 입력 파라메터에 따라서도 클러스터의 모양이 달라지는 것을 볼 수 있다. 공간 마이닝의 응용 시스템에 따라서는 오브젝트의 위치 속성을 기반으로 하지만, 클러스터를 결정할 때 비 공간 속성이 갖는 영향을 고려해야 하는 경우가 있다. 이때 본 논문에서 제시한 DBSCAN-W는 기존의 DBSCAN과는 달리 사용자의 의도를 반영한 새로운 클러스터링 결과를 도출할 수 있다.

5. 결론 및 향후 연구과제

지식탐사 프로세스의 핵심적인 역할을 담당하는 데이터 마이닝 단계에서는 여러 가지 목적에 따라 알고리즘을 선택하여 사용한다. 최근 통계, 비즈니스, 전자 상거래, 의학, 생물학 등의 분야에서 데이터 마이닝 기술이 적극적으로 활용되고 있으며 이를 위해 다양한 알고리즘들이 계속해서 연구 개발되고 있다. 본 논문은 대상 오브젝트들이 공간적인 위치 속성을 갖는 공간 데이터 마이닝을 수행할 때 데이터의 밀도를 고려하면서 시스템의 특성에 따라 오브젝트의 비공간 속성에 가중치를 부여하는 DBSCAN-W를 제안하였다. 또한 실험을 통해 기존의 DBSCAN 알고리즘에서 찾아내지 못한 사용자의 의도를 반영한 새로운 형태의 클러스터링 결과를 도출하는 것을 보였다. 향후 각각의 실제 응용에 적합한 가중치를 부여하기 위해서 오브젝트의 영역을 결정하는 방법에 대한 실제적이고 다양한 연구가 필요하다. 또한 우리는 이 알고리즘이 강이나 고속도로와 같은 장애물이 존재하는 경우를 다룰 수 있도록 확장할 것이다.

참고 문헌

[1] Michael J. A Berry, and Gorden Linoff, Data Mining Techniques : For Marketing, Sales, and Customer Support, John Wiley & Sons, Inc., 1997.

[2] Raymond T. Ng, and Jiawei Han, "Efficient and Effective Clustering Method for Spatial Data Mining," In Proc. of the VLDB Conference, Santiago, Chile, pp. 144-155, September 1994.

[3] Tian Zhang, Raghu Ramakrishnan, and Miron Livny, "BIRCH : An Efficient Data Clustering Method for Very Large Databases," In Proc. of the ACM SIGMOD Conference on Management of Data, Montreal, Canada, pp. 103-114, June 1996.

[4] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim, "CURE : An Efficient Clustering Algorithm for Large Databases," In Proc. of the ACM SIGMOD Conference on Management of Data, Seattle, Washinton, USA, pp. 73-84, May 1998.

[5] Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," In Proc. of ACM SIGMOD 3rd International Conference on

Knowledge Discovery and Data Mining, pp. 226-231, AAAI Press, 1996.

[6] Mihael Ankerst, Markus M. Breuning, Hans-Peter Kriegel, and Jorg Sander, "OPTICS: Ordering Points To Identify the Clustering Structure," In proc. of ACM SIGMOD International Conference on Management of Data, Philadephia, Pennsylvania, USA, pp. 49-60, June 1999.

[7] W.Wang, J.Yang, and R.Muntz, "STING :A statistical information grid approach to spatial data mining", In Proc. 1997 Int. conf. Very Large Data Bases(VLDB'97), Athens, Greece, pp.186-195, August 1997.

[8] Jiawei Han, and Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann publishers, 2001.

[9] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey, "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections," In Proc. of the 15th Annual International ACM SIGIR Conference, pp. 318-329, June 1992.

[10] Informix, Informix Universal Server Guide to SQL: Tutorial Version 9.1, Informix Press, 1997.

[11] Informix, Informix Spatial Datablade Module: User's Guide, Informix Press, 1997.

Abstract

Design and development of the clustering algorithm considering weight in spatial data mining

Ho-Sook Kim* · Hyun-Sook Lim** · Hwan-Seung Yong***

Spatial data mining is a process to discover interesting relationships and characteristics those exist implicitly in a spatial database. Many spatial clustering algorithms have been developed. But, there are few approaches that focus simultaneously on clustering spatial data and assigning weight to non-spatial attributes of objects. In this paper, we propose a new spatial clustering algorithm, called DBSCAN-W, which is an extension of the existing density-based clustering algorithm DBSCAN. DBSCAN algorithm considers only the location of objects for clustering objects, whereas DBSCAN-W considers not only the location of each object but also its non-spatial attributes relevant to a given application. In DBSCAN-W, each datum has a region represented as a circle of various radius, where the radius means the degree of the importance of the object in the application. We showed that DBSCAN-W is effective in generating clusters reflecting the user's requirements through experiments.

Key words: 공간 마이닝, 공간 클러스터링, 가중치

* DIT of Computer Information, Dongeui Institute of Technology
** LG CNS
*** Dept. of Computer Science and Engineering, Ewha Womans University