空間データ発掘によるクラスタ発見手法の精度評価

3R－1

川原 稔　　*河野 浩之　　*長谷川 利治

京都大学大型計算機センター　　*京都大学大学院工学研究科

1 はじめに

空間データベース (spatial database) の利用において，空間データの検索を効率良く行うことは最も重要となる．しかし，データ相互の関連を求めるデータベースにおける知識発見 (KDD) の視点から，多次元属性をもつ多量のデータの集合操作の必要な空間データマイニング (spatial data mining) が重要となっている．[1]．そこで，文献 [2] における点データのクラスタリング操作を必要とする問合せ処理の高速な実行を実現するための研究を行っている．

我々は，空間インデックスの構成手法によっては，二次元平面上のデータ分布を効率良く保持する性質があることに着目し，空間インデックスを利用したクラスタリング処理を前処理として用いることを提案している [1]．以下，提案した空間インデックスを用いたクラスタリングアルゴリズムに対するクラスタリングの精度や，大量データを扱うこの種の問合せ処理におけるノイズ処理の問題を述べる．

2 空間データマイニング

空間データマイニング手法は，空間属性と非空間属性から構成される空間データベースにおけるデータに対して適用される．例えば，属性指向アルゴリズムを空間データベースに拡張した空間データマイニングの研究として文献 [2] がある．多量の多次元データを扱う際には，二次元平面上に重なりを抑えながら高速に射影する手法を適用することも必要となる．さらに，空間データマイニングにおいて文字属性を扱う手法がルール導出において重要と考えられる．

以下，空間データマイニングの典型的な問合わせ例を示す．さらに，通信ネットワークから収集されるデータのアドレス空間などにおいても，{Src, Dest} などに空間的特性が含まれることを表1に示す．

（例）

```
discover spatial association rules
inside TARGET_REGION
from Table_1 X, Table_2 Y
in relevance to Table_3 Z
where X.attrib1=''Value1'' AND ...
```

2.1 R-tree と PR-Quadtree

空間データ構造 (spatial data structure) を考慮した空間データ特有の操作を効率的に行う空間インデックス法 (spatial indexing method) が数多く提案されており，point Quadtree, k-d-tree, k-d-B-tree などの空間インデックスが代表的な技法である [4]．

特に，バケットに格納可能なデータ数の上限 M，下限 $m(\leq M/2)$ を設定し，以下の条件に合致するまで分割を繰り返す R-tree と，データ空間を規則的に分割する方法である point quadtree インデックス法の変形である PR-Quadtree[4] が多く研究されている．後者の PR-Quadtree インデックスは，挿入されるデータの順序によって木の形状が変化せず，データ分布に従って一意のインデックスが形成される．特に，検索条件を満すクラスターの性質を求める空間データマイニングの特徴となる問合せ処理では，空間的な配置が保存される構造が望ましい．

2.2 インデックスを用いたクラスタリングアルゴリズム

我々は，PR-Quadtree の性質に着目し，データに対して PR-Quadtree によるインデックスが張られている場合には，各バケットに対するバケット密度に比例した重さの点をクラスタリングの対象として扱い，アルゴリズム PAM(Partitioning Around Medoids)[3] を適用することを提案した [1]．PAM において代表点となる k 個の medoid は，ある medoid と，その medoid が属するクラスタ内の他のオブジェクトとの，ユークリッド距離 (Euclidean distance)，あるいはマンハッタン距離 (Manhattan distance) の平均が最小となるように選ぶことにした．

さらに，実際のデータベースに格納されるデータを操作する場合には，どのクラスタにも属さないノイズデータを効率良く除去する操作が必要不可欠となる．そこで，バケット密度の閾値を設定することによってクラスタに属さないノイズデータを含むバケットを除去する方法も提案した．

Evaluation of Index Based Clustering Method in Spatial Database
Minoru Kawahara, Hiroyuki Kawano and Toshiharu Hasegawa
Kyoto University
Kyoto 606-01, Japan

3 クラスタリング性能の評価

本研究では，PR-Quadtreeの木の深さ制限や，データの含むノイズ量を変化させた場合のクラスタリング精度を評価する．

サンプリングを用いたクラスタリング手法との比較

無作為抽出によるサンプリングを行った後に，PAMを適用する方法によって得られたクラスタリングの平均距離を1とし，我々の提案したクラスタリングによる性能と比較を行った．表2における実験例に対する最適な木の深さとして，深さ5を採用したが，領域に対して記述しなければならないクラスターの構造に合わせた動的な木の深さの変更も必要である．

ノイズ除去に対する性能評価

ノイズ除去を含むクラスタリング性能の評価のために，各点の属するクラスタが既知であるデータ集合を生成し，そのデータに対してアルゴリズムを適用した．図1に示したように，各々500個の点データよりなる4個のクラスタによるデータを実験で用いた．また，ノイズデータはランダムに，2.5%, 5%, 10%発生させ，クラスタリングを行った．

PR-Quadtreeによるインデックスを張った図1において，網のかかっていない密度の低いバケットはノイズデータのみを含むものとして除去し，網掛け部であるバケットがクラスタリング対象となる．

上記のデータ集合に対してノイズの除去を含んだクラスタリングを行った結果を表3に示す．誤って除去されるデータ量を3%以内に押さえながら，90%以上のノイズを除去することができたことが分かる．この結果は，本研究で提案したノイズ除去操作が，空間データマイニングに利用するに十分な性能を持つことを示している．

4 結論と今後の課題

以上，主記憶上に格納可能なインデックスによって，空間データマイニングの基礎であるクラスタリングのための二次記憶上のデータ読み出し回数を抑制し，大幅に計算コストの低減を可能とした．また，我々のアルゴリズムは，ノイズの除去操作を含むクラスタリングを精度良く行うことも可能としている．

今後，データベースにおいて用いられる各種インデックス構造のもつ情報量を評価し，効率良い集合操作を行うために必要なインデックス手法を明らかにする必要がある．

謝辞

本稿の一部は文部省科学研究費重点領域「分散発展型データベースシステム技術の研究(08244103)」のもとでの研究成果による．

表 1 空間的特性をもつデータ例

Lnth	Proto	Src	Dest	Src port	Dst port
98	udp	20.65	20.69	983	111
60	tcp	20.192	42.1	4453	119
60	tcp	42.1	20.192	119	4453
88	udp	43.1	43.30	1028	53
60	tcp	28.1	31.80	4009	25
...	...	...	...	...	...

表 2 提案手法の無作為抽出に対する比較

クラスタ数	データ数	平均距離(index)
3	300	0. 9862
5	500	1. 0107
10	1000	0. 9948

表 3 ノイズ除去性能

ノイズ	除去率(%)	不正除去率(%)	クラスタ率(%)
50	94.0	2.70	97.30
100	94.0	1.85	98.15
200	91.0	2.30	97.70

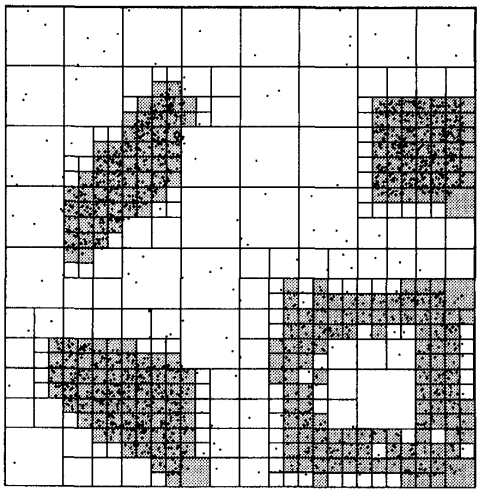

図 1 ノイズを含むデータ例

参考文献

[1] 伊藤, 河野, 長谷川, "空間データマイニングにおけるクラスタ発見とインデックス構造の利用," 人工知能学会全国大会論文集, pp231–234, (平成8年6月).

[2] R.T. Ng, and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," Proc. 20th VLDB, pp.144–155, 1994.

[3] L. Kaufman and P.J. Rousseeuw, "Finding Groups in Data: an Introduction to Cluster Analysis," John Wiley & Sons, 1990.

[4] H. Samet, "Spatial Data structures," *Modern Database Systems,* (W. Kim, ed.), ACM Press, New York, pp.361–385, 1995.