4L-05

配列 DBMS における空間スキャン統計量の計算手法

安田 健人 † 河井 悠佑 †† 趙 セイ ††† 杉浦 健人 ††† 石川 佳治 ††

† 名古屋大学工学部電気電子・情報工学科 †† 名古屋大学大学院情報学研究科

††† 名古屋大学大学院情報科学研究科

1 はじめに

近年膨大な量のデータが利用できるようになり，大規模データの活用が課題となっている．特に，どのようにしてデータの中から必要な情報を取り出すかというデータマイニングは大きな注目を浴びている．データマイニングの 1 つである空間クラスタリングの方法は多く提案されているが，その中でも代表的なものとして空間スキャン統計量が挙げられる．空間スキャン統計量は空間的に過密な領域を検出するために使用され，新たな病気の流行の発見などに役立っており，計算の高速化など多くの研究が行われている [1,2].

一方，多次元配列の大規模データ分析に特化した配列指向 DBMS が開発されており，注目を浴びている [3]. そこで, 本研究では配列指向 DBMS の 1 つである SciDB を活用し，空間スキャン統計量を計算する手法について検討を行う．また，計算した空間スキャン統計量を用いて，空間的に過密な領域を検出する方法についても述べる．

2 配列指向 DBMS

本研究では配列指向 DBMS の 1 つである SciDB [3,4] を用いる．SciDB は多次元配列形式のデータを格納，処理できる大規模並列 DBMS である．配列は次元と属性値をそれぞれ複数持つことができる．図 1 は 2 つの属性値を持つ 2 次元配列の例である。配列 DBSM ではその要素の並び方にも意味があり，例えば，位置座標や時系列を表している．

2.1 SciDB の特徴

SciDB の最大の特徴はチャンクを使った処理である．チャンクは SciDB におけるデータ処理の最小単位であり，配列の各要素は，固定長のチャンクへと分割して配置される．隣接した要素は同じチャンクに配置されるため，意味的に近いデータ同士が物理的にも近い位置に格納され，効率的な処理が可能となる．チャンクの大きさは配列を生成する際に次元ごとに設定することができ，参照されるチャンクの数が少なくなるようにデータの性質に合わせてチャンクを適切な大きさに設定することによって，より効率的に演算処理を行える．チャンクは重ね合わせることもでき，重なるセルの数についても設定が可能である．また，SciDB では並列処理が取り入れられている．チャンクを複数のインスタンスに振り分けることで効率的に処理が行われる．

i \ j	1	2	3	4
1	0,1	0,0	4,6	1,2
2	2,3	2,3	1,2	1,6
3	1,4	5,5	1,2	0,3
4	0,4	3,5	2,2	1,3

図 1 SciDB の 2 次元配列の例

2.2 SciDB の演算

SciDB では，配列に格納されたデータに対して様々な演算処理を実行でき，配列の作成や削除，ファイルからの出入力，抽出，集約などの演算が使える．例えば，抽出演算では各次元について抽出範囲の上限と下限を指定し，範囲内の要素を抽出することができる．apply 演算ではすでに存在している配列の 1 つに新たな属性値を追加できる．集約演算では，指定した範囲の集約関数を実行することで集約値を計算し，結果を新たな配列として出力する．集約値を計算する集約関数としては合計値や最大値などがある．集約演算の 1 つである window 演算では集約関数を実行する領域の大きさを次元ごとに指定することで，指定した大きさのすべての部分領域における集約値を計算できる．例えば、図 1 の属性値のうち左の属性値の最大値を求めたい場合，2 × 4 という大きさを指定することで，2 × 4 の大きさの 3 つの部分領域すべてについて集計演算が実行され，3 つの最大値 (5,5,4) を得ることができる．本研究では，このような SciDB の様々な演算を組み合わせて使うことにより，空間スキャ

Computing Spatial Scan Statistics in an Array DBMS

Kento Yasuda†, Yusuke Kawai††, Jing Zhao†††, Kento Sugiura†††, and Yoshiharu Ishikawa††

†Department of Information Engineering, School of Engineering, Nagoya University

††Graduate School of Informatics, Nagoya University

†††Graduate School of Information Science, Nagoya University

ン統計量の計算を実現する．

3 空間スキャン統計量

空間スキャン統計量 [2,5] はある測定値 (measurement) m と基準値 (baseline measure) b から計算できる非単調密度尺度の 1 つである．空間的に過密な領域を検出するためによく用いられており，疫学や生物学の分野で広く知られている．空間スキャン統計では m は平均 qb のポアソン過程によって生成されていると仮定している．ここで q はある定数である．例えば，m は稀な病気の症例数であり、q は病気になる確率 ($\frac{m}{b}$ の期待値)，b は病気のリスクがある人の数である．このとき，空間スキャン統計量を計算することにより，病気の流行を早期に検出することができる．

空間スキャン統計量の計算手順について述べる．ここで，G を $N \times N$ の正方形グリッドとし，G に含まれる最小の各グリッド g_{ij} が，測定値 m_{ij} と基準値 b_{ij} を持つとすると，

$$M = \sum_{g_{ij} \in G} m_{ij} \tag{1}$$

$$B = \sum_{g_{ij} \in G} b_{ij} \tag{2}$$

のように m と b それぞれに関してグリッド全体での総和を求めることができる．また，G の部分矩形領域 S について，

$$m_s = \frac{\sum_{g_{ij} \in S} m_{ij}}{M} \tag{3}$$

$$b_s = \frac{\sum_{g_{ij} \in S} b_{ij}}{B} \tag{4}$$

のように m と b それぞれに関して全体の総和のうち部分領域が占める割合を計算すると，空間スキャン統計量は，

$$d(m_s, b_s) = m_s \log\left(\frac{m_s}{b_s}\right) + (1 - m_s) \log\left(\frac{1 - m_s}{1 - b_s}\right) \tag{5}$$

のように計算することができる．ただし，$m_s \leq b_s$ のとき，$d(m_s, b_s) = 0$ とする．この値が大きい部分矩形領域ほど，空間的に過密な領域であると判断できる．

4 実装

本研究では空間スキャン統計量の計算を，SciDB の演算を用いて実装する．ここでは，$N \times N$ の正方形グリッド G に集約されたデータが，SciDB 上の配列に格納されている状況を考える．分析に用いるデータは次元 $\{i, j\}$，属性値 $[m, b]$ の 2 次元配列である．i と j はそれぞれグリッドの座標に，m は測定値，b は基準値に対応している．

分析するデータが事前に格納された 2 次元配列を用いて，部分矩形領域ごとに空間スキャン統計量の計算を行い，その中で空間スキャン統計量が最大である部分矩形領域を求めることにより，G の最大密度領域を見つけ出せる．具体的には, 最初に配列に対し集約演算を実行し，m と b それぞれについて，すべての要素の合計値 M, B を計算する．そして，その値を使って全体の総和のうち各グリッドの要素が占める割合をすべてのグリッドに関して求め，新たな属性値 m_g，b_g として配列へ格納する．次に，window 演算を用いてグリッド上のすべての部分矩形領域に関して順に m_g と b_g の合計値 m_s，b_s を求める．その値を用いて各部分領域の空間スキャン統計量を順に計算し，その結果を 4 次元配列に格納する．この 4 次元配列は，部分矩形領域の左上の座標と大きさを次元として持つ。最後に，4 次元配列から空間スキャン統計量が一番大きな領域を抽出することで，密度が最大の領域を見つけ出す．ここでは，要素を降順に並び替えるためにソート演算と，一番目の要素を抽出するための抽出演算を用いる．

5 まとめと今後の課題

本研究では配列 DBMS を用いた空間スキャン統計量の計算手法について検討を行い，その値が最大となる領域を抽出することで最大密度領域を検出した．今後の課題としては，最大密度領域が統計的に有意であるかを判別できるようにすることや，空間スキャン統計量の計算の高速化が挙げられる．

謝辞

本研究の一部は，科研費（16H01722）および CREST「大規模・高分解能数値シミュレーションの連携とデータ同化による革新的地震・津波減災ビッグデータ解析基盤の創出」による．

参考文献

[1] D. B. Neill and A. W. Moore, “Rapid detection of significant spatial clusters,” *In SIGKDD 2004*, pp. 256–265, 2004.

[2] D. Agarwal, A. McGregor, J. M. Philips, S. Venkatasubramanian, and Z. Zhu, “Spatial scan statistics: Approximations and performance study,” *In SIGKDD 2006*, pp. 1481–1496, 2006.

[3] M. Stonebraker, P. Brown, J. Becla, and D. Zhang, “SciDB: A database management system for applications with complex analytics,” *IEEE Computational Science & Engineering*, vol. 15, no. 3, pp. 54–62, 2013.

[4] “Paradigm4: Creators of SciDB a computational DBMS.” http://www.paradigm4.com/.

[5] M. Kulldorff, “A spatial scan statistics,” *Comm. in Stat.: Th.and Meth.*, vol. 26, pp. 24–33, 1997.