基于 K-Means 聚类的 R-树空间索引方法研究与分析

余冬梅

陕西理工学院数学与计算机科学学院,陕西汉中 723000

摘要　空间聚类和空间索引的结合是当前空间数据库中提高数据检索效率的技术之一。本文从空间聚类和空间索引的存储原理入手,阐述了 K-Means 聚类算法及其改进算法的技术思路,研究了 K-Means 算法在空间数据库中与空间索引方法结合的技术问题;分析了当前基于 K-Means 算法的 R-树系列空间索引技术的研究成果,阐述了它们提高空间检索效率的技术路线及实验结果,研究显示这些技术都能在一定程度上提高数据检索的效率。最后给出了聚类与空间索引结合技术未来的研究方向。

关键词　空间聚类;空间索引;K-Means 算法;R-树

中图分类号　TP301.6　　**文献标识码**　A　　**doi**　10.3981/j.issn.1000-7857.2012.11.011

R-tree Spatial Index Based on K-Means Clustering

YU Dongmei

School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong 723000, Shaanxi Province, China

Abstract　At present, combining spatial clustering with spatial index is one of the techniques that could enhance data retrieval efficiency in the spatial database research. Based on the storage principles of spatial clustering and spatial index, K-Means clustering algorithm and the technical ideas behind improved algorithm are elaborated; the techniques of combining K-Means algorithm with spatial index method in spatial database are studied. Current research results of R-tree series spatial index technology based on K-Means algorithm air analyzed and their technical ideas for improving spatial retrieval efficiency and experiment results are described. The research shows that these techniques could enhance data retrieval efficiency to a certain degree. In the end, the future research trend of technique about combining clustering with spatial index is proposed.

Keywords　spatial clustering; spatial index; K-Means algorithm; R-tree

0　引言

随着地理信息系统、计算机辅助设计与制造、遥感技术以及定位服务系统的应用,空间数据库技术得到了长足的发展,也使空间数据的海量性、对象复杂性等显得更加突出,所以有效解决空间数据库中空间数据的存储、查询等基础问题显得尤为重要。研究发现,空间聚类和空间索引技术可以提高空间数据的处理效率[1],但只用单一的一种技术还不能高效解决以上问题。因此,将空间聚类算法和空间索引结构相结合的技术成了新的研究热点,尤其在 K-Means 聚类算法基础上建立空间索引,可以改善单一技术的缺点,进一步提高空间数据的存储与运算效率。

1　空间聚类与空间索引

空间聚类是将具有相似特性(如空间位置相邻)的数据归为一类(簇),使同一类的数据相似性最大,不同类之间的数据相异性最大,并将同一类的数据尽量存储在存储器相近或同一页面区域,其目的是降低数据访问的 I/O 次数,提高效率。空间聚类的存储结构见图 1。空间索引是在存储空间数据时依据空间对象的位置和形状或空间对象间的某种空间关系等形成概要信息,并将其按照一定顺序排列的一种数据结

收稿日期:2011-11-30;修回日期:2012-04-05

基金项目:陕西理工学院科研项目(SLG0819)

作者简介:余冬梅,讲师,研究方向为空间数据库与地理信息系统,电子信箱:yudongmei609@yahoo.com.cn

构,其目的是在空间数据库中不需逐个查找就能快速定位所需空间对象要素,提高空间操作速度。空间索引的存储结构见图 2。空间索引也可放在空间聚类算法和存储器之间,尤其是当存取海量的区域高密度空间数据时, 先将数据进行聚类,再对每簇数据进行索引存取,能有效提高 I/O 效率。

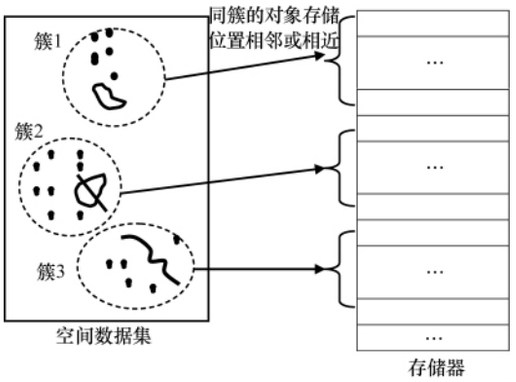

图 1　空间聚类示意图

Fig. 1　Diagram of spatial clustering

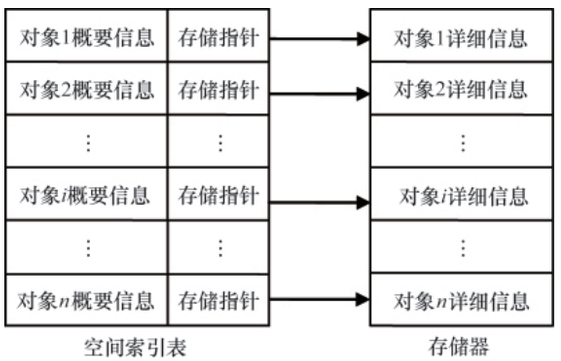

图 2　空间索引示意图

Fig. 2　Diagram of spatial index

目前关于空间索引方法的研究已经有了很多成果,如基于二叉树、基于四叉树、基于 B 树、基于动态哈希的格网和基于空间目标排序等的空间索引技术[1]。关于空间聚类分析的算法也有很多, 如划分法 (Partitioning Methods)、层次法 (Hierarchical Methods)、基于密度的方法 (Density-based Methods)、基于网格的方法(Grid-based Methods)和基于模型的方法(Model-based Methods)。K-Means(K-平均)算法属于基于划分的聚类方法,其简单、快速、易于理解与实现、能够有效处理大规模数据等诸多优点,使之成为最经典、应用最广泛的聚类方法之一。

虽然 K-Means 算法应用非常广泛,但若将其作为空间聚类算法,其自身缺陷(如初始聚类中心的随机选取会导致聚类效果出现偏差;聚类数 K 的经验性确定和数据输入顺序的不同会带来不良的聚类效果)使得实际应用效果受到很大影响。随着空间数据库应用的普及和深入,当数据库中数据量急剧增大时,该算法的缺陷更加突出。为此,很多研究都是在该算法的基础上进行改进或者与该算法一起构成混合聚类算法。本文根据近几年的研究成果,对基于 K-Means 的空间聚类算法及其改进思路, 以及基于 K-Means 算法的改进的 R-树空间索引方法进行梳理和比较分析,以期能对以后的研究有所帮助。

2　经典 K-Means 聚类算法及其改进

2.1　经典 K-Means 聚类算法

设将 d 维数据集 $X=\{x_i|x_i\in \mathrm{R}^d, 1\leqslant i\leqslant N\}$聚集成簇集合 $W=\{w_1,w_2,\cdots,w_k\}$,各簇的质心依次为 $c_1,c_2,\cdots,c_k$。聚类效果的好坏用目标函数 J 衡量,J 是每个数据对象与所在簇的质心的欧氏距离之和,因此 J 值越小,簇越紧凑,相对越独立。算法通过不断优化 J 的值来寻求更优的聚类方案,当 J 取极小值时,对应的聚类方案即为最优方案。

K-Means 算法[2]步骤如下:

第 1 步　从数据集 X 中随机选择 k 个初始参照对象 c_1, $c_2,\cdots,c_k$;

第 2 步　以 $c_1,c_2,\cdots,c_k$ 为参照点(质心),依据如下原则划分集合 X,令

$$d(x_j,c_i)=\sqrt{\sum_{t=1}^{d}(x_{jt}-c_{it})^2}\qquad 1\leqslant i\leqslant k,1\leqslant j\leqslant N$$

对每个参照点 c_i,求使 $d(x_j,c_i)$最小的 x_j,即

$$d_{\text{最小}}=\min d(x_j,c_i)\qquad 1\leqslant j\leqslant N, x_j\neq c_i$$

并将 x_j 划归到以 c_i 为质心的簇中

第 3 步　根据公式

$$c_i=\frac{1}{n_i}\sum_{x_i\in w_i}x_i\qquad 1\leqslant i\leqslant k$$

重新计算簇的质心 $c_1',c_2',\cdots,c_k'$,其中 n_i 是簇 w_i 中数据对象的个数;

第 4 步　若对于任意 $i\in\{1,2,\cdots,k\}$,都有 $c_i'=c_i$ 成立或者 $J=\sum_{j=1}^{k}\sum_{i=1}^{n_i}d(x_j,c_i)$收敛,则将当前的 $c_1',c_2',\cdots,c_k'$ 作为最终形成的簇质心,算法结束,否则,令 $c_i=c_i'(1\leqslant i\leqslant k)$,返回第 2 步继续执行。

2.2　K-Means 聚类算法的改进思路

对 K-Means 聚类算法的改进思路主要集中在类簇数 k 的选择、初始 k 个聚类中心点的选择、相异程度计算和聚类中心值迭代计算方法等方面,列举如下。

(1) 根据数据自身分布特点(如不同区域的密度)确定 k 值。杨悦[3]在每一维度上沿坐标轴方向建立样本空间的数据直方图, 对整个样本空间在每一维度递归实现最优划分,最终由划分区域数确定 k 值,并非随机确定 k 值。胡伟[4]在初始设定 k 值并 K-Means 聚类的结果中选择簇半径最大的簇及其质心,并用该簇中距离质心最远的样本点 x_1 以及距离 x_1 最远的样本点 x_2 和其他质心迭代进行更细层次聚类,调整 k 值并使其达到合理的聚类。

(2) 根据数据自身分布特点(如不同区域的密度)产生 k 个初始聚类中心。王会青等[5]先确定 k 值,然后利用著名的 Prim 算法选取样本空间中密度最大的 k 个点并将它们作为初始聚类中心。

(3) 改进前 K-Means 算法只能发现球形簇的缺陷,改进后可以发现任意形状的簇。对数据样本空间进行多次采样和 K-Means 预处理产生多组聚类结果,再对各组的交集构造出子簇的加权连通图,合并连通的子簇并最终确定聚类结果[6]。

(4) 对数据样本进行权重修正以调整最终聚类效果。先用构造的权重函数对整个数据变量加权后聚类,然后对不同子簇内的变量进行不同权重的加权后,再利用 K-Means 算法迭代聚类,这样改变样本对簇质心的影响,从而改变对象间的相异(或相似)性[7],达到调整最终聚类的效果。

(5) 迭代过程中优化聚类中心选取方法。利用簇的均值点与聚类种子相分离的思想,在进行第 k 轮迭代的时侯,选择第 $k-1$ 轮的簇 $w_{i(k-1)}$ 中与该轮聚类种子 S 相似度大于 $1-\beta(1-S\_mini_{(k-1)})$ 的数据的均值作为第 k 轮聚类种子,其中 $0<\beta<1$,$S\_mini_{(k-1)}$ 是簇 $w_{i(k-1)}$ 中的数据与该簇的聚类种子 S 之间的最小相似度[8]。该方法消除了孤立点的负面影响。

3 基于 K-Means 及其改进算法的 R-树空间索引方法

经过几十年的研究探索,空间索引方法的研究成果已非常丰硕,其中以 B 树为研究基础的 R-树[9]自 1984 年诞生以来已有很多空间数据库系统将其作为主要的空间索引算法之一。以下主要基于 R-树索引算法,分析空间数据在经 K-Means 或其改进算法聚类后,再进行 R-树空间索引的技术。

3.1 聚类的 R-树空间索引算法

在 R-树算法中,不同的空间索引目录矩形之间的面积是允许重叠的,但是随着数据量的增大或者数据密度的增大,这种重叠量会迅速上升,进而增大检索路径,检索失败的路径也增多,使检索效率快速下降。鉴于此,王锡钢等[10]改进的索引算法是从缩小 R-树中间结点的最小外接矩形 MBR(也称为目录矩形)的空白面积,从而减小 MBR 重叠面积角度进行的。首先在对 R-树进行构建或插入新结点时,采用 K-Means 聚类算法的"距离最小"准则来代替原 R-树算法的"面积增量最小"准则进行对象聚集。针对点状要素采用点间最小距离,对于面状要素采用两个矩形之间的最近点距离,而非两个矩形的中心点距离。根据该算法,数据聚集后 MBR 中的空白区域比改进之前有明显缩小。这样每个 MBR 在所包含的数据量不减的情况下,面积减小了,也就降低了不同 MBR 之间相互重叠的机会,即使重叠了,重叠面积也会明显缩小,其理论效果优于原始 R-树索引算法。空间查询运算实验测试结果表明,改进后的算法的实际效率提高了 4—5 倍,说明改进方法能有效提高检索 R-树算法的速度。

3.2 聚类的 Hilbert R-树空间索引算法

Hilbert R-树是利用 Hilbert 曲线把空间的大量高维数据对象映射到一维存储器空间的过程,即对空间数据对象的 MBR 的中心点进行 Hilbert 码排序,然后把有序的数据对象按照前面的排序顺序递归生成 R-树。Hilbert R-树的空间利用率较 R-树高一些,但高维数据按照其中某一维进行存储也破坏了空间数据在空间上相对相邻的特点,即在空间某维相邻的对象,其真正物理存储位置可能不再相邻,这容易造成 MBR 的交叠,当数据对象分布不均匀且数量增大时,MBR 的空白区域率会升高,使检索数据的效率降低。

为解决此问题,何小苑[11]、韩秋英[12]、王宝祥[13]等提出在对空间数据对象的 MBR 进行 Hilbert 码排序前,先对其进行聚类预处理,即以 MBR 为中心,利用 K-Means 算法或改进的 K-Means 算法对其进行聚集形成 K 个类,再分别对每个类中对象的 MBR 进行 Hilbert 码排序,并按 R-树对结点中对象个数的要求组成不同的叶结点,之后逐层进行 Hilbert 排序聚集形成中间结点,直至最终形成混合聚类的 Hilbert R-树。根据空间查询运算实验测试结果,与改进前 Hilbert R-树的算法相比,基于 K-Means 的 Hilbert R-树在测试集为均匀状态下效率提高约 30%,为正态分布状态下的效率提高约 70%,为极不均匀状态下的效率提高约 2 倍。实验总体说明,该方法能够有效提高空间数据的检索效率,尤其对不均匀分布的空间数据集,通过聚类使相邻数据提前聚集并以邻居的形式存储在一起,减少了 I/O 时间,因此查询性能的提高更加有效。

3.3 改进聚类的 R-树空间索引算法

对 M 阶的 R-树进行构建和插入新结点时,会因为结点内索引项个数 M 的限定,使得结点数目达到 M 时要进行分裂,分裂的基本原则是分裂后的两个目录矩形面积最小。刘彦宾[14]认为,按照面积最小原则分裂后的两个结点,不能很好地体现出聚类效果,对于给定对象数目的数据空间,会因为 MBR 划分的个数不同而影响整个 R-树的空间检索效率。刘彦宾进行的算法改进主要是针对 R-树结点分裂个数而言的,即通过 K-Means 算法确定出 R-树中间结点要分裂成的结点个数 K,然后根据此 K 值来分裂该结点。在确定 K 值初值时,$N/M \leqslant K \leqslant N/m$,这里的 N 为当前待分裂结点的子结点个数,m 为 R-树结点中拥有的索引项数的最小值。在调用 K-Means 算法迭代时,附加了评价函数 $P(K)$,最终取能使 $P(K)$ 函数值最小的 K 值作为最优划分的簇数,并依此聚类数对 R-树进行分裂。这样不是按照原始 R-树的分裂方法中一分为二的思想,而是根据子结点中索引项的聚类的簇数来确定分裂数目,使得分裂后的结点内部更紧凑,某种程度上可以改善检索效率。以空间查询进行测试运算,结果表明,改进后的索引算法的效率平均提高约 15%,而且随着空间数据集中数据量的增加效果越加明显。

3.4 改进聚类的 R-link 树空间索引算法

R-link 树是 R-树的一种变形,其中心思想是将聚类算法与 R-link 树索引方法相结合。而对于 K-Means 算法,若将初始聚类中心选取在几个分布密集区域的中心,其周围对象易分到最近的簇,聚类收敛越快,需要迭代的次数越少[15]。

赵伟等[16]在构建 R-link 树之前,先对数据的 MBR 进行

K-Means 聚类,然后根据聚类实际结果再进行建树。但这里的 K-Means 算法是经过改进的,主要改进了 K-Means 算法中确定初始聚类中心的方法,即采用均值-标准差法确定初始聚类中心,并以类间离散度最大、类内离散度最小的原则,使用准则函数优化并最终确定最优聚类数 K。该聚类算法的改进是针对减少迭代次数的原理提出的,提高了收敛速度。最后用此聚类结果构造 R-link 树,使得 R-link 树各子结点更紧凑,聚类性能更高,从而达到提高空间检索效率的目的。与改进前 R-link 树的算法相比,改进后的基于 K-Means 的 R-link 树,经空间查询运算对不同容量的空间数据集进行测试,当 R-link 树深为 3 层时改进后的查询效率平均提高约 20%,R-link 树深为 4 层时改进后的查询效率平均提高约 1 倍,这说明空间数据量增大趋于海量集时其效率也会逐步提高。

4　结论与展望

以上几种方法,均是以 K-Means 聚类算法为基础的 R-树空间索引技术,总体思路都是在对海量空间数据进行索引前先进行空间聚类,将待检索的空间范围或数据量有效缩小或减少,从而提高空间数据检索和分析的效率。当然,这些只是空间聚类和空间索引技术相结合的一个缩影,还有其他一些空间聚类算法(如基于随机搜索的聚类方法、平衡迭代消减聚类法、基于密度的聚类法和采用遗传算法的空间聚类法[1]等)和基于它们与空间索引算法结合而产生的索引技术;而空间索引技术也还有几类,如基于二叉树的索引、基于四叉树的索引以及基于空间目标排序的索引[1]等,且空间聚类和空间索引技术都还处在不断研究和发展之中,还存在很多问题和技术有待进一步的探讨,如高维空间数据的聚类和索引问题。随着计算机 CPU 速度的不断提升,处理海量空间数据效率的瓶颈主要表现在 I/O 上,因此空间聚类与索引算法的 I/O 代价问题也是未来研究的主要问题之一。随着空间数据库应用的不断扩大和深入,空间数据也将朝着更加庞大和复杂的方向发展,对大型高维空间数据集进行高效数据分析和数据挖掘也已成为许多领域的迫切需要。因此,聚类算法的研究及其在空间数据库中的应用,仍然有很多工作要做,也是未来一段时间研究提高空间数据存取效率的主要方向之一。

参考文献(References)

[1] 郭薇, 郭菁, 胡志勇. 空间数据库索引技术[M]. 上海: 上海交通大学出版社, 2006: 78-82.
Guo Wei, Guo Jing, Hu Zhiyong. Spatial database index technology[M]. Shanghai: Shanghai Jiaotong University Press, 2006: 78-82.

[2] 王礼常, 王志章, 陶果. 致密砂岩气藏储层分类新方法 [J]. 科技导报, 2011, 29(24): 47-49.
Wang Lichang, Wang Zhizhang, Tao Guo. *Science and Technology Review*, 2011, 29(24): 47-49.

[3] 杨悦. 面向空间数据复杂性特征的聚类分析方法研究 [D]. 哈尔滨: 哈尔滨工程大学, 2008: 40-43.
Yang Yue. Research on complexity-oriented spatial data clustering analysis methods [D]. Harbin: Harbin Engineering University, 2008: 40-43.

[4] 胡伟. 改进的层次 K 均值聚类算法[J/OL]. 计算机工程与应用, [2011-10-24]. http://www.cnki.net/kcms/detail/11.2127.TP.20111024.1013.064.html.
Hu Wei. Computer Engineering and Applications, [2011-10-24]. http://www.cnki.net / kcms /detail/11.2127.TP.20111024.1013.064.html.

[5] 王会青, 陈俊杰, 郭凯. 启发式初始化独立的 k-均值算法研究[J/OL]. 计算机工程与应用, [2011-07-20]. http://www.cnki.net/kcms/detail/11.2127.TP.20110720.1517.077.html.
Wang Huiqing, Chen Junjie, Guo Kai. Computer Engineering and Applications, [2011-07-20]. http://www.cnki.net/kcms/detail/11.2127.TP.20110720.1517.077.html.

[6] 雷小锋, 谢昆青, 林帆, 等. 一种基于 K-Means 局部最优性的高效聚类算法[J]. 软件学报, 2008, 19(7): 1684 -1688.
Lei Xiaofeng, Xie Kunqing, Lin Fan, *et al*. *Journal of Software*, 2008, 19(7): 1684-1688.

[7] 黄哲学, 徐军, 景丽萍, 等. K-均值算法家族和子空间聚类 [J]. 计算机研究与发展, 2005, 42(S): 55-60.
Huang Zhexue, Xu Jun, Jing Linping, *et al*. *Journal of Computer Research and Development*, 2005, 42(S): 55-60.

[8] 栾丽华. 聚类算法研究[D]. 南京: 南京师范大学, 2004: 11.
Luan Lihua. Study on clustering algorithms [D]. Nanjing: Nanjing Normal University, 2004: 11.

[9] Guttman A. R-tree: A dynamic index structure for spatial searching[J]. *ACM SIGMOD Record*, 1984, 41(2): 47-57.

[10] 王锡钢, 任伟, 李青元, 等. 基于 K-means 聚类距离准则的 R 树结点分配算法研究[J]. 测绘科学, 2006, 31(5): 117-118.
Wang Xigang, Ren Wei, Li Qingyuan, *et al*. *Science of Surveying and Mapping*, 2006, 31(5): 117-118.

[11] 何小苑, 闵华清. 基于聚类的 Hilbert R-树空间索引算法 [J]. 计算机工程, 2009, 35(9): 40-42.
He Xiaoyuan, Min Huaqing. *Computer Engineering*, 2009, 35 (9): 40-42.

[12] 韩秋英. 基于混合聚类的空间索引算法研究及其应用[D]. 开封: 河南大学, 2010: 34-36.
Han Qiuying. Research and application of spatial index algorithm based on hybrid cluster analysis[D]. Kaifeng: Henan University, 2010: 34-36.

[13] 王宝祥. 基于改进聚类的 Hilbert R-树空间索引算法研究 [D]. 开封: 河南大学, 2011: 34-45.
Wang Baoxiang. Research of Hilbert R-tree spatial index algorithm based on improved clustering analysis [D]. Kaifeng: Henan University, 2011: 34-45.

[14] 刘彦宾. 基于聚类分析的 R-树空间索引研究[J]. 廊坊师范学院学报: 自然科学版, 2009, 9(3): 27-29.
Liu Yanbin. *Journal of Langfang Teachers College: Natural Science Edition*, 2009, 9(3): 27-29.

[15] Ester M, Kriegel H P, Sander J. A density-based algorithm for discovering clusters in large spatial databases with noise[M]//Simoudis E, Han J, Fayyad U M, *et al*. Proceeding of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland: AAAI Press, 1996: 226-231.

[16] 赵伟, 张姝, 李文辉. 基于 K-means 算法的高性能空间索引方法[J]. 计算机工程, 2008, 34(10): 5-6.
Zhao Wei, Zhang Shu, Li Wenhui. *Computer Engineering*, 2008, 34(10): 5-6.

(责任编辑　马宇红,代丽)