层次维编码片段立方体生成算法应用研究

张子轩,万定生,朱 凯
(河海大学 计算机与信息学院,江苏 南京 210098)

摘 要:数据量大、数据多维是水利普查数据的重要特征。根据水利普查决策分析的需要,在对数据立方体技术研究的基础上,基于部分物化策略,提出了建立层次维编码片段立方体(HDEFC)。利用维度属性的概念分层特性,在层次维片段中采用混合索引(B-tree 和 Bit Code)技术对每个层次维的层次属性进行二进制编码,再利用生成的维度编码代替原表中关键字,非层次维片段中采用倒排索引技术对每个片段子立方体进行物化,减少了多表连接操作,从而提高 OLAP 查询效率。实验结果表明,生成的 HDEFC 占用较小的存储空间,查询方法在面对高维的复杂查询时具有优势。通过建立水利普查数据分析系统,说明了该方法能够有效地解决因数据量庞大、维度多导致的数据计算和查询效率低下等问题,降低了物化水利普查成果数据立方体的时间和空间成本。

关键词:水利普查;数据多维;数据立方体;数据分析系统;层次维编码片段
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2017)02-0134-05
doi:10.3969/j.issn.1673-629X.2017.02.030

Application Research on Hierarchical Dimension Encoding Fragment Cube Algorithm

ZHANG Zi-xuan, WAN Ding-sheng, ZHU Kai
(College of Computer and Information, Hohai University, Nanjing 210098, China)

Abstract: Large amount of and multidimensional data is an important feature of water census data. According to the need of water census decision analysis, on the basis of data cube technology and partial materialization strategy, the establishment of Hierarchical Dimension Encoding Fragment Cube (HDEFC) is put forward. By the concept hierarchy characteristics of the dimension attribute, hybrid index (B-tree and Bit Code) technology is used to execute binary coding for hierarchy properties of each dimension, and the generated dimension code is applied to replace the key in the original table. In addition, non hierarchical dimension fragment uses inverted index technology to materialize each sub cube, so as to reduce the multi table join operation and improve OLAP query efficiency. Experiments show that the generated HDEFC occupies less storage space, and the query method has advantages in the face of high dimensional complex query. Through the establishment of water census data analysis system show that the method can effectively solve the problem of low efficiency of data calculation and query because of the huge amount of and multi-dimensional data, which reduces the cost of time and space of the material of water census results data cube.

Key words: water census; multi-dimensional data; data cube; data analysis system; hierarchical dimension encoding fragment

0 引 言

随着全国水利普查[1]工作的开展,形成了迄今最为全面细致、完整系统的涉水基础数据资源和规范权威的水利普查成果数据,如何对这些普查成果数据进行有效的分析与利用成为了制定正确水利建设方案的关键问题。

数据仓库[2]作为一种新兴技术被越来越多的领域所重视,数据挖掘[3-4]和联机分析处理(OLAP)[5-6]都是基于数据仓库的分析工具。在 OLAP 中,数据通常以数据立方体(Data cube)[7]的方式存储,数据立方体以“维度+度量”的方式组织数据,提供给分析人员多维的数据视图,直观地支持了 OLAP 中所需要的复杂多维分析。为节省存储空间同时兼顾 OLAP 查询效率,对数据立方体进行预计算的有效方法可提高联机分析处理能力。

结合水利普查数据量大、多维度、分层次的特征,

收稿日期:2016-04-08 **修回日期**:2016-08-02 **网络出版时间**:2017-01-10
基金项目:国家科技支撑计划课题(2015BAB07B01);水利部公益性行业科研专项(201501022)
作者简介:张子轩(1994-),女,硕士研究生,研究方向为信息处理与信息系统;万定生,教授,CCF 会员,研究方向为信息处理与信息系统。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20170110.1019.058.html

文中以建立层次维编码片段立方体(HDEFC)[8-9]为重心,采取能够有效处理高维度的HDEFC生成算法和查询算法,将其运用于水利普查数据分析,并逐步实现OLAP、数据挖掘等功能。以水利普查中的水电站工程数据分析主题为例,利用层次维编码片段立方体作为底层数据结构,并展开了对水利普查数据分析与展示系统的建立与应用研究。

1　层次维编码片段立方体技术

1.1　HDEFC思想

数据仓库中,立方体可以用 $C=(D,M)$ 表示。其中,$D=\{D_1,D_2,\cdots,D_n\}$ ($n\geq 1$)是维度的集合;$M=\{M_1,M_2,\cdots,M_m\}$ ($m\geq 1$)是度量的集合。D_i 表示一个维度,D_i 维中各成员的层次维编码是由 D_i 维中所有层次维属性 $(L_1^i,\cdots,L_2^i,\cdots,L_h^i)$ 的二进制编码按照 $B^{D_i}=(\cdots((B^{L_1^i}\ll B_{L_2^i}\mid B^{L_2^i})\ll B_{L_3^i}\mid B^{L_3^i})\cdots)\ll B_{L_h^i}\mid B^{L_h^i}$,根据层次维由高到低依次进行组合而成的混合编码。为了详细说明HDEFC方法在水利普查数据分析中的可用性,以水利工程中的水电站工程数据为例,水电站工程数据如表1所示。

表1　水电站工程数据表

TID	ADDVCD	GCJB	JSQK	KFFS	SL	ZJRL
1	210201	G_1	J_0	K_1	1	1 000
2	210202	G_1	J_0	K_1	2	900
3	210304	G_1	J_1	K_1	1	500
4	210304	G_2	J_0	K_2	1	250
5	210305	G_2	J_1	K_2	1	80

例1:表1生成的原始数据立方体由ADDVCD(水电站所在行政区划代码)、GCJB(水电站的工程级别)、JSQK(水电站建设情况)和KFFS(水电站开发方式)4个维度(简称A,G,J,K)和SL(数量)、ZJRL(水电站装机容量)2个度量构成,SK_Cube中包含16个聚集cuboids:{(A,G,J,K),(*,G,J,K),(A,*,J,K),(A,G,*,K),(A,G,J,*),(*,*,J,K),(*,G,*,K),(*,G,J,*),(A,*,*,K),(A,*,J,*),(A,G,*,*),(*,*,*,K),(*,*,J,*),(*,G,*,*),(A,*,*,*),(*,*,*,*)}。维度ADDVCD中包含3个概念分层以及4个不同的属性值,维度GCJB包含2个不同的属性值,维度JSQK包含2个不同的属性值,维度KFFS包含2个不同的属性值,则将生成32((3+1)×2×2×2)个聚集cuboids和810((4+1)×(2+1)×(1+1)×(2+1)×(2+1)×(2+1))个聚集单元。

定义1:二进制编码。设 S 是一个字符串集合,且 $S=\{S_1,S_2,\cdots,S_n\}$,$N=|S|$,则 S 中任意字符串的二进制编码的位数为 $\lceil\log_2^N\rceil$,即 S 的二进制编码为 $\{e_1,e_2,\cdots,e_n\}$。其中,若 $S_i\leqslant S_j$,则 $e_i\leqslant e_j$。

定义2:层次维编码。对于具有 D_i 个维度的层次维,对维中的所有层次片段 $(L_1^i,\cdots,L_2^i,\cdots,L_h^i)$ 采用混合索引技术,对每个层次维的层次属性采用二进制编码,并按层次的高低排序进行组合。其中,每层二进制编码总位数为 $\lceil\log_2^m\rceil$,m 为 L_j^i 层中不同成员的最大数量。

以水利普查对象中的行政区划维ADDVCD为例,通过层次维编码方法创建Province(省份)、City(城市)和County(区县)这三个维度的层次编码,并将其与事实表中记录的TID进行关联。表1中ADDVCD维的编码与水电站工程数据简表中各记录的TID关联关系如表2所示。

表2　ADDVCD层次维编码表

Province	City	County	B^{ADDVCD}	TID列表	列表大小
20	02	01	101000001000001	{1}	1
20	02	02	101000001000010	{2}	1
20	03	04	101000001100100	{3,4}	2
20	03	05	101000001100101	{5}	1

定义3:层次维编码片段立方体。对于一个具有 n 个维度的高维数据立方体 $C(D,M)$,将维度集合 $\{D_1,D_2,\cdots,D_n\}$ 按照互不相交的原则,划分为 k 个独立的片段,其中层次维单独划分为一个片段,而非层次维划分为大小为 m 的片段。对层次维中的层次属性利用混合索引技术进行二进制编码,然后用倒排索引技术[10]存储非层次维片段中的聚集cuboids。这样就将一个 n 维的立方体分割成 k 个低维的立方体,形成HDEFC。

例2:以表1水电站工程数据简表为例,采用HDEFC生成算法,立方体片段大小为3,则划分出的HDEFC为(ADDVCD),(GCJB,JSQK,KFFS)共2个片段,需要物化12((3+1)+2×2×2)个聚集cuboids和57((4+1)×(2+1)×(1+1)+(2+1)×(2+1)×(2+1))个聚集单元,所需存储空间明显减少。

定义4:非层次维倒排索引。对于HDEFC立方体中每个非层次维的每个属性值,列出具有该值的所有元组的元组标识符(TID),形成非层次维的倒排索引。

例3:在表1中,非层次维有GCJB、JSQK、KFFS,对全部非层次维建立倒排索引。例如,属性值 G_1 出现在元组1、2和3中,则 G_1 的TID列表包含3项,即1、2和3。对这3个维中的属性值建立倒排索引得到表3。

至此,HDEFC立方体中层次维自成一个片段立方体,而非层次维片段立方体需要计算生成。表1中的

非层次维恰好可以形成一个片段大小为 3 的立方体。

表 3 由表 1 生成的非层次维倒排索引表

TID	SL	ZJRL
G_1	{1,2,3}	3
G_2	{4,5}	2
J_0	{1,2,4}	3
J_1	{3,5}	2
K_1	{1,2,3}	3
K_2	{4,5}	2

1.2 HDEFC 生成算法

首先,将给定数据集(即基本方体)的所有维划分成独立的维群组,也可称作立方体片段。其中每个层次维单独为一个片段,数个非层次维组合成一个片段(关于非层次维的片段大小以及维分组将在下文进行讨论)。

扫描基本方体,并构造每个维属性的倒排索引。对于每个片段,计算完全局部(即基于片段的)数据立方体,而保留倒排索引。此外,对方体中每个单元保留倒排索引,即对于每个单元,记录它的关联 TID 列表。

算法伪代码如下所示:

输入:一个具有 n 个维度的立方体 C:(D_1,D_2,…,D_n)。

输出:片段划分集合{ F_1,F_2,…,F_k }和相对应的局部片段立方体{ HC_1, HC_2,…, HC_k };ID_measure 数组。

方法:

将 n 维立方体 C(D_1,D_2,…,D_n)分割成 k 个分段{ F_1,F_2,…,F_k };

for($i=1$; $i<=n$; $i++$){

将每个元祖的 TID 和 MEASURE 插入 ID_measure 数组;

if (维可分层){

利用二进制编码创建层次维编码表和 TID 关联表 <B,TID list >;

}

else{

创建 D_i 各属性 TID 关联表,即倒排索引<V,TID list >;

}

}

for(每个非层次维片段 F_i){

交叉计算同一 F_i 片段中的 D_i 维对应的 TID 列表值并计算其相应的度量值,创建局部片段立方体 HC_i ;

}

1.3 HDEFC 片段大小选择

HDEFC 生成算法中,立方体片段大小的选择显得至关重要。片段过大会导致需要预聚集的立方体过多,从而增加存储空间的负担;片段过小会导致预聚集的立方体过少,从而无法满足大部分的临时聚集操作,增加响应时间。必须在存储空间和响应时间中寻找最佳平衡点,即 HDEFC 片段最佳大小。

根据 HDEFC 生成算法的思想,假设某立方体共有 S 个维度,片段大小为 L,则每个片段将生成 2^L 个立方体。将立方体片段大小 L 设为自变量 x,因变量 y 为立方体总数量,得到函数($x>1$)。

以水利普查中水库工程为例,维度总数为 19(行政区划、流域水系、水资源区划、水库类型、挡水主坝类型按材料分、挡水主坝类型按结构分、主要泄洪建筑物形式、工程建设情况、水库调节性能、工程等别、主坝级别、重要保护对象、供水对象、水库归口管理部门、是否完成划界、是否完成确权、工程任务、2011 年供水量数据来源、主要挡水建筑物类型),3 个区域层次维自成片段,对剩下 16 个维度划分立方体片段。

经过研究得出,立方体数量随着立方体片段大小的增加呈指数级增长,即 HDEFC 存储空间与立方体片段大小呈指数关系。考虑到立方体片段过小造成的临时聚集时间增加这一可预见的事实,所以对于最佳立方体片段大小的选择应该是 3 或 4 为宜。

1.4 HDEFC 查询算法

通常的查询类型包括点查询[11]、范围查询[6]和冰山查询[12]。由于 HDEFC 定位于水利普查高维数据查询,现在基于 HDEFC 结构模型,采用范围查询方法进行有效查询处理。

以范围查询为例,HDEFC 范围查询算法可以被分解成多个点查询,在范围查询中,立方体至少存在一个维度 d_i ,其取值不唯一。该算法核心在于将查询条件中的维度条件分解整理后,找出同一 HDEFC 片段中的维度,分别从不同的 HDEFC 片段中查询出符合查询条件的 TID 列表,最终再对各 TID 列表进行求交。

详细的算法伪代码如下:

输入:立方体片段集合{ HC_1,HC_2,…,HC_k };ID_measure 数组;< a_1,a_2,…,a_n : M >形式的范围查询。

输出:基于范围查询的结果度量。

方法:

for($i=1$; $i<=k$; $i++$){

B_{qi} ←运用点查询算法得到每个点查询的相交集合;

}

if(存在至少一个非空 B_{qi}){

B_q ←{ B_{q1},B_{q2},…,B_{qk} }集合求交得到最终集合;

}

$M \leftarrow B_q \cap$ ID_Measure

return M

2　实验结果与分析

本节在不同大小的数据集上对文中提出的 HDEFC 生成算法进行性能比较。实验硬件环境:处理器为 Intel(R) Core(TM) i5-4750 CPU @ 3.40 GHz,内存为 4 GB,硬盘为 640 GB、7 200 rpm。软件环境:操作系统为 Win 7,数据库为 SQL Server 2005,编程语言为 Java,平台为 Eclipse。

实验采用水利基本情况普查中的水电站工程对象数据,记录数为 98 000,对基本表进行整理后选取 18 个维度属性和 10 个度量属性,并在实验前对各数据进行了规范化处理。各实验中每个片段大小为 3。

图 1 是层次维编码相对简单位图索引的压缩比(坐标经过处理)。

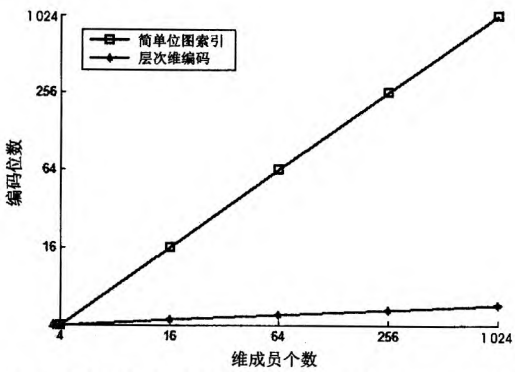

图 1　层次维编码相对简单位图索引的压缩比

从图中可以看出,当维成员个数为 256 时,简单位图索引需要 256 bits,而层次维编码位数只需 8 bits,数据压缩比为 256/8 = 32。随着维成员数量的持续增加,数据压缩比将进一步增大。

图 2 展示了基本事实表记录数对数据立方体存储空间的影响。

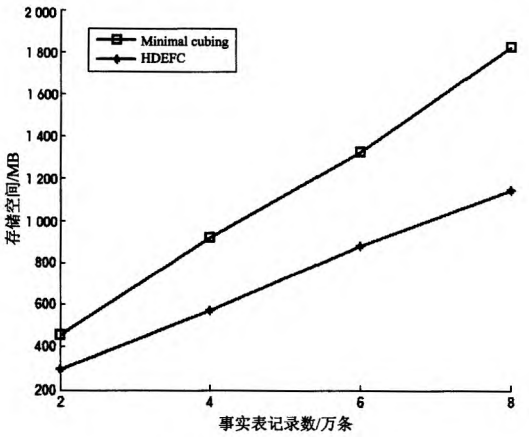

图 2　存储性能比较

总的来说,HDEFC 方法所需的存储空间整体小于简单原始的外壳片段立方体(Minimal cubing)[13-14]方法。随着记录数的增加,HDEFC 方法中层次维编码高压缩性的特点就表现得更加明显,且所需存储空间的增长趋势相对缓慢。实验结果表明,HDEFC 方法能够有效减少立方体存储空间,相对 Minimal cubing 方法具有一定优势。

3　系统设计与实现

依据文中所述的层次维编码片段立方体技术,系统进行了相应的数据库逻辑结构和物理结构设计。开发与部署主要运用了 JSP 和 GROOVY 技术,其中前台使用 JSP 技术进行界面展示,后台与 SQL Server 数据库的连接使用 GROOVY 语言,再结合 ArcGIS 平台对水利普查数据进行地图可视化展示。文中所构建的水利普查层次维编码片段数据立方体在系统运行过程中表现出显著的优势,比如对水利普查数据维度的分析与查询响应时间在该系统中得到了高效的实现,给用户带来了极大方便。图 3 显示了水利普查数据分析与展示信息结果。

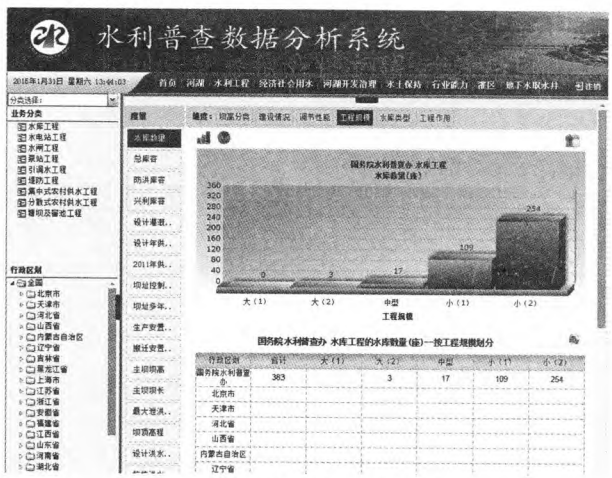

图 3　水利普查数据分析系统

该系统实现了水利普查成果展示、水利普查基础数据查询、水利普查主题查询、空间分布以及文档资料等功能,能够对水利普查数据进行查询与展示,可以有效地分析水利普查成果数据,进而进行决策分析。该系统运用信息可视化技术,建立了可交互的友好美观的用户界面,能有效地查询水利普查对象地理信息与具体信息。

水利普查数据分析系统还加入了“水利普查主题展示”的概念,在水利普查对象的大环境中以用户关注点为中心,进行单类对象多指标以及多类对象相关指标的主题定制展示。

总之,该系统的开发研究,为水利普查数据分析提供了有效的平台,以方便水利业务人员实现对于水利

普查数据的二次加工和深度分析。

4 结束语

在介绍水利普查数据特征的基础上，文中提出了适用于高维水利普查数据分析的部分物化立方体结构—HDEFC，以具有代表性的水电站工程数据为例，就 HDEFC 的生成算法、查询算法和立方体片段大小选择进行了探究，并构建了水利普查数据分析系统。此系统能够满足对水利普查数据分析的基本要求，且数据分析查询响应速度快于一般数据仓库的多维数据分析系统。通过实验进一步验证了提出的 HDEFC 方法在水利普查数据分析领域的适用性。相比于目前数据分析系统中普遍采用的低维度少度量的处理方式，得益于 HDEFC 的底层立方体结构和相应的查询方式，可以灵活地进行高维度多度量的查询和分析，为水利决策者提供了更加广阔的数据视野。

参考文献：

[1] 庞进武，程益联，罗志东. 水利普查与信息化[J]. 水利信息化，2012(1)：19-22.

[2] 占 军，万定生，李 宇. 基于 Oracle 数据仓库的水利普查数据展现系统[J]. 计算机与数字工程，2012，40(10)：55-57.

[3] 杨嘉杰. 水量水费数据立方体的 OLAP 和数据挖掘技术研究[D]. 广州：中山大学，2012.

[4] 尹 涛，关兴中，万定生. 数据挖掘技术在水文数据分析中的应用[J]. 计算机工程与设计，2012，33(12)：4721-4725.

[5] Chaudhuri S，Dayal U. An overview of data warehousing and OLAP technology[J]. ACM SIGMOD Record，1997，26(1)：65-74.

[6] Ho C T，Agrawal R，Megiddo N，et al. Range queries in OLAP data cubes[J]. ACM SIGMOD Record，1970，26(2)：73-88.

[7] Gray J，Chaudhuri S，Bosworth A，et al. Data cube：a relational aggregation operator generalizing group-by，cross-tab，and sub-totals[J]. Data Mining and Knowledge Discovery，1997，1：29-53.

[8] Markl V，Ramsak F，Bayer R. Improving OLAP performance by multidimensional hierarchical clustering[C]//Proc of IDEAS'99. [s. l.]：[s. n.]，1999：165-177.

[9] 胡孔法，董逸生，徐立臻，等. 一种基于维层次编码的 OLAP 聚集查询算法[J]. 计算机研究与发展，2004，41(4)：608-614.

[10] 林俊鸿，姜 琨，杨岳湘. 倒排索引查询处理技术[J]. 计算机工程与设计，2015，36(3)：572-575.

[11] 朱 凯，万定生，程习锋. 水利普查成果分析中数据立方体计算研究[J]. 计算机与数字工程，2014，42(9)：1591-1594.

[12] Fang M，Shivakumar N，Garcia-Molina H，et al. Computing iceberg queries efficiently[C]//International conference on very large databases. New York：[s. n.]，1999.

[13] Li X，Han J，Gonzalez H. High-dimensional OLAP：a minimal cubing approach[C]//Proceedings of the thirtieth international conference on very large data bases. [s. l.]：[s. n.]，2004：528-539.

[14] Li C，Cong G，Tung A K H，et al. Incremental maintenance of quotient cube for median[C]//Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. Washington：ACM，2004：226-235.

(上接第 133 页)

断方法研究[J]. 航空动力学报，2009，24(7)：1649-1653.

[5] Blackwell R，Millott T. Dynamic design characteristics of the Sikorsky X2 technology demonstrator aircraft[C]//American helicopter society 65th annual forum. Montreal，Canada：[s. n.]，2008.

[6] 周 虹，左洪福，蔡 景，等. 基于 TMSDG 的民用飞机故障诊断隔离策略[J]. 航空学报，2012，33(3)：479-486.

[7] 陈 晓，马建仓. 基于 MEL 倒谱的某型飞机发动机振动故障的模式识别[J]. 计算机测量与控制，2012，20(8)：2028-2030.

[8] 马建仓，刘小龙，陈 静. 二代小波降噪与盲分离结合应用于航空发动机振动信号分析[J]. 机械科学与技术，2010，29(1)：7-11.

[9] 戴 敏，谢 椿. 基于模糊加权有色网和 BP 神经网络的飞机发动机故障诊断[J]. 科学技术与工程，2012，12(35)：9552-9556.

[10] 赵 鹏，蔡忠春，李晓明，等. 某型飞机发动机故障诊断专家系统设计[J]. 计算机测量与控制，2014，22(12)：3850-3852.

[11] 于 霞，张卫民，邱忠超，等. 飞机发动机叶片缺陷的差激励涡流传感器检测[J]. 北京航空航天大学学报，2015，41(9)：1582-1588.

[12] 黄 强，王 健，张桂刚. 一种航空发动机传感器故障诊断方法[J]. 传感技术学报，2014，27(10)：1315-1320.

[13] 李业波，李秋红，黄向华，等. 航空发动机气路部件故障融合诊断方法研究[J]. 航空学报，2014，35(6)：1612-1622.

[14] Błachnio J. Capabilities to assess health/maintenance status of gas turbine blades with non-destructive methods[J]. Polish Maritime Research，2015，21(4)：41-47.