CAGD 中参数曲面的光滑拼接研究*)

黄俊英[1] 王相海[1,2]
(辽宁师范大学计算机与信息技术学院 大连 116029)[1]
(南京大学计算机软件新技术国家重点实验室 南京 210093)[2]

摘 要 参数曲面作为 CAGD 中形状数学描述的标准形式一直受到关注,而参数曲面的光滑拼接作为实现复杂客体几何造型的重要手段一直是该领域的一个热点和难点问题。以不同参数域的参数曲面为线索,对常用的矩形域和三角域参数曲面的 GC^1 光滑拼接的条件进行了分析,同时对这些条件在实际应用中的一些问题进行了讨论,最后对曲面光滑拼接中一些令人关注问题进行了展望。

关键词 参数曲面,几何连续,光滑拼接

Investigation of Smooth Joining of Parametric Surfaces in CAGD

HUANG Jun-ying[1] WANG Xiang-hai[1,2]
(College of Computer and Information Technology, Liaoning Normal University, Dalian 116029, China)[1]
(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)[2]

Abstract As a standard form to express figure math, people always pay attention to parametric surface, and as an important means of complex object geometric modeling, smooth joining of parametric surfaces is always a hotspot and difficulty of this area. Through the thread of different domains parametric surfaces, this paper analyses the GC^1 smooth joining condition of common rectangular and triangular parametric surface, and discusses some questions about the practical application of these conditions. In the end takes a long view of some interested problems on smooth joining of surfaces.

Keywords Parametric surface, Geometric continuity, Smooth joining

1 引言

在实际几何造型应用中,由于客体表面的复杂性,一般用一个单一的曲面很难实现对其表面的描述,通常采用带有光滑约束的多片曲面拼接来实现。曲面的参数表示自从 60 年代被美国波音飞机公司的 Ferguson 采用以来作为形状数学描述的标准形式一直受到人们的关注,同时参数曲面的光滑拼接作为计算机辅助几何设计(CAGD)的一个热点和难点问题而受到重视。

在曲面的光滑拼接中,有两种不同的光滑度(smoothness)度量[1,2]:一种是函数曲线的可微性,即把组合曲面沿公共连接线处处具有直到 n 阶的连续偏导矢,这类光滑度被称为沿着拼接线具有 n 阶参数连续性,记为 C^n;另一种度量为几何连续性,记为 GC^n 或 G^n,即沿着正则公共拼接线的两个参数曲面,其中之一可以通过重新参数化来达到使二者在公共拼接连线处达到 C^n。参数连续性总是与曲面的参数选取或具体参数化有关,并且对于具有非正则特性的两曲面拼接曲线,尽管沿连接线是 C^1 的,却有可能不是都处处存在公共的切平面,因而是不光滑的。而几何连续性与参数选取及具体的参数无关,这就排除了由参数选取引起的非正则情况。同时,几何连续性为形状的定义和控制提供了额外的自由度,人们可以通过人机交互在保证几何连续约束条件下对形状进行设计和调整。

多年来,随着曲面设计实际应用的需求,曲面的几何连续拼接得到了不断的发展,人们提出了许多几何连续的拼接条件,以满足实际应用的需求。本文以不同参数域的参数曲面为线索,对常用的矩形域和三角域的参数曲面的 GC^1 光滑拼接条件进行了分析和讨论,同时对实际应用中的一些问题进行了讨论,最后对该领域值得关注的一些问题进行了展望。

2 基于矩形域的参数曲面的拼接

2.1 Bézier 曲面的 GC^1 拼接

两个参数曲面在拼接边界上达到 GC^1 连续的最具普遍性条件是 Bézier 给出的[3]:设两个相邻的参数曲面 $p(u,v)$ 和 $q(u,v)$,如果它们有公共的边界 $p(u,0)=q(u,1)$,且存在关于 u 的正数量函数 $h(u)$,$g(u)$ 使得式(1)成立,则两个曲面在拼接边界处达到 GC^1 连续。

$$p_v(u,0)=h(u)q_v(u,1)+g(u)q_u(u,1) \tag{1}$$

在此基础上,Bézier 具体给出了双 3 次 Bézier 曲面的 GC^1 充分条件:将式(1)中的数量函数取为 $h(u)=\alpha>0$,$g(u)=\beta+\gamma u$,推出一个具体计算待定参数的关系式。然而,由于这种推导方法缺乏规律性、运算量大且易于出错,所以在实际应用中受到一定的限制。

kahmann 在文献[4]中对参数曲面 GC^1 的拼接条件进行了更为深入的研究,给出了任意两个 $m\times n$ 次 Bézier 曲面的 GC^1 连续条件,即两个任意的 $m\times n$ 曲面要在拼接边界上达

*)本文受辽宁省高等学校优秀人才支持计划(RC-04-11),辽宁省自然科学基金(20072156),辽宁省教育厅科学技术研究项目(20060486)和南京邮电学院图像处理与图像通信江苏省重点实验室开放基金(ZK207006)资助。**黄俊英** 硕士研究生,研究方向为 CG&CAGD;**王相海** 博士,教授,主要研究领域为 CG、CAGD、多媒体信息处理。

到GC^1拼接，除了满足式(1)外，为了使两个拼接曲面在公共边界处不形成“尖锥”，两个曲面与公共边界相对的另两条边应位于公共边界的两侧，同时为了保证$p(u,v)$的次数不变，$h(u)$，$g(u)$应为：

$$h(u)=\alpha>0, g(u)=(1-u)\beta+u\gamma$$

其中α,β,γ为实常数。将式(1)中各偏导用控制顶点给出，整理后比较两边同次项参数u的系数，得到GC^1连续条件：

$$\Delta^{0,1}P_{k,0}=\alpha\Delta^{0,1}Q_{k,n-1}+\beta\frac{m-k}{n}\Delta^{1,0}Q_{k,n}+\gamma\frac{k}{n}\Delta^{1,0}Q_{k-1,n},\quad k=0,1,\cdots,m \tag{2}$$

其中$\Delta^{0,1}$，$\Delta^{1,0}$为一阶差分。该条件表明，当曲面$p(u,v)$给定后，与之沿公共边界GC^1连接的曲面$q(u,v)$的第二排控制顶点$Q_{k,1}$就可由式(2)确定下来，其中$\alpha>0$。β,γ可用来作为对曲面$q(u,v)$进行形状控制的参数。该条件便于控制曲面的形状，同时可适用于确定已知曲面的光滑拼接曲面。

刘鼎元等在文献[5]中提出了一种双3次Bézier曲面的拼接条件：设曲面$p(u,v)$和$q(u,v)$的控制点分别为$P_{i,j}$和$Q_{i,j}(i,j=0,1,2,3)$，如果存在实数$a(>0)$，b,c，使得

$$T_2=-aT_0+bT_1, V_2=-aV_1+(1+a-2/3b-c/3)T_1+2/3b(E+T_1')$$

$$T_2'=-aT_0'+cT_1', V_2'=-aV_1'+(1+a-2c/3-b/3)T'_1+2/3c(E'+T_1)$$

其中

$$\begin{cases}T_0=P_{10}-P_{00}, T_1=P_{01}-P_{00}, T_2=Q_{10}-Q_{00}, V_1=P_{11}-P_{00}, V_2=Q_{11}-Q_{00}, E=Q_{03}-Q_{00}\\ T_0'=P_{13}-P_{03}, T_1'=P_{02}-P_{03}, T_2'=Q_{13}'-Q'_{03}, V_1'=P_{12}-P_{03}, V_2'=Q_{12}-Q_{03}, E'=Q_{00}-Q_{03}\end{cases}$$

那么，曲面$p(u,v)$和$q(u,v)$就是GC^1连续的，如图1所示。

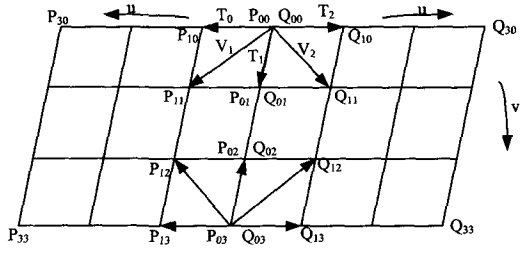

图1 两个双3次Bézier曲面拼接

此外，文献[6]取$h(u)=h, g(u)=u_0+u_1u$，其中h,u_0,u_1均为常数，利用Bernstein基函数所具有的关系式$tB_{n,j}(t)=\frac{j+1}{n+1}B_{n+1,j+1}(t)$将式(1)改写为：

$$Q_{10}-Q_{00}=h(P_{30}-P_{20})+u_0(P_{31}-P_{30})$$

$$Q_{11}-Q_{01}=h(P_{31}-P_{21})+\frac{u_0}{3}[(P_{31}-P_{30})+2(P_{32}-P_{31})]+\frac{u_1}{3}(P_{31}-P_{30})$$

$$Q_{12}-Q_{02}=h(P_{32}-P_{22})+\frac{u_0}{3}[2(P_{32}-P_{31})+(P_{33}-P_{32})]+\frac{2u_1}{3}(P_{32}-P_{31})$$

$$Q_{13}-Q_{03}=h(P_{33}-P_{23})+u_0(P_{33}-P_{32})+u_1(P_{33}-P_{32})$$

这样条件(1)便转化为一种关于控制顶点的拼接条件，从而避免了实际应用中繁琐的数值计算。

在几何造型中，有理曲面尤其是有理Bézier曲面被广泛使用，这主要是因为有理曲面能精确地描述在工程上经常使用的二次曲面。有理Bézier曲面GC^1拼接的充要条件早在1990年分别由刘鼎元、DeRose TD、Degen. W. L. F提出[7-9]，但是这些条件在实际应用中过于繁琐。

2.2 B样条曲面的GC^1拼接

B样条曲面间GC^1连续的实现方法主要分为两类，一类是在曲面公共边界处布置检查点，并计算这些点处两曲面的边界切矢与跨界切矢，再根据几何连续定义调整边界控制顶点，实现光滑拼接[10]。之所以提出这种拼接条件是因为在达到切平面连续的条件中参数因受曲面具体表达的限制而难以确定，该种充要条件具有普遍的适用性，不受曲面具体表达的限制，对形状不一、拼接部位不同的曲面也能达到良好的拼接效果。另外，应用这种方法进行GC^1拼接后，两曲面公共边界线不会发生任何改变，从而保证曲面通过具有严格约束要求的分界线。该条件的缺点是在对控制顶点进行调整的过程中通过某一调整量逐一检验是否达到光滑，在实际运用中计算量很大。另一类方法是在几何连续的基础上利用B样条控制向量的本征方程，求得与已知曲面光滑拼接的另一曲面控制顶点。文献[11]详细推导了具有内部单节点的双3次B样条曲面的几何连续条件，以及在公共边界线上控制向量的本征条件。利用节点插入公式将B样条曲面中的切向量曲线转化成分段的Bézier形式，多项式参数定义为常用的BézierGC^1光滑关系式中的系数[12]，从而推导出对于不同内部节点情况下的拼接条件及本征方程。这种方法的优点在于可以通过一个已知曲面构造它的光滑拼接曲面，并可以在任意的空间四边形剖分上构造GC^1光滑的NURBS曲面模型，缺点是计算量相当大。

上述两类实现方法均利用了B样条曲面和Bézier曲面间可以相互转换的性质，借助Bézier曲面在端点处良好的几何性质求解B样条曲面的拼接条件。

NURBS曲面的优点是对标准解析形状及自由曲线和曲面提供了统一的、精确的数学表达式，同时提供了额外的自由参数(权系数)，其能控制曲线或曲面形状，具有仿射不变性。但是由于NURBS解析表达式的复杂性，要得到其在边界上切平面连续的充要条件比较困难。文献[13]给出了具有二次公共边界曲线的NURBS曲面GC^1光滑拼接的一个实用算法。该算法首先将有理曲面转换为与权值函数有关的齐次坐标形式，再利用商数定理推出两个具有公共边界曲线的有理曲面$r(u,v)$，$\bar{r}(\bar{u},\bar{v})GC^1$连续的充要条件为：

$$\begin{cases}\bar{Q}(0,\bar{v})=c_0(v)Q(0,v)\\ \bar{Q}_{\bar{u}}(0,\bar{v})=c_1(v)Q(0,v)+c_0(v)p_1(v)\\ Q_u(0,v)+c_0(v)q_1(v)Q_v(0,v)\end{cases}$$

其中，$c_0(v)$，$c_1(v)$，$p_1(v)$，$q_1(v)$是关于公共边界参数$v=\bar{v}$的函数，$Q(u,v)$，$\bar{Q}(\bar{u},\bar{v})$分别为$r(u,v)$，$\bar{r}(\bar{u},\bar{v})$的齐次坐标形式。文献[14]中取$c_0(v)=\bar{\omega}(0,\bar{v})/\omega(0,v)$，$c_1(v)$，$p_1(v)$为任意常数，$q_1(v)=\beta_0+\beta v$，最终推导出一系列显式表示条件。文献[15]在上述方法的基础上进一步考虑具有q次公共边界的NURBS曲面的情况，得到了实现NURBS连续拼接的充分条件，该条件可以通过调整一个曲面的靠近边界的两排控制顶点和权因子而实现两曲面片的GC^1拼接。

3 基于三角区域的曲面拼接技术

3.1 三角Bézier曲面的光滑拼接

基于三角形域上的三角参数曲面以其具有拓扑解析结构和适应于无规律、复杂散乱数据几何造型以及有限元分析中

三边形元素的需求[16,17]。

给定两个 n 次相邻三角 Bézier 曲面：$p(\tau)=\sum_{i=0}^{n}\sum_{j=0}^{n-i}P_{ijk}B_{ijk}^{n}(\tau)$，$p'(\tau')=\sum_{i=0}^{n}\sum_{j=0}^{n-i}Q_{ijk}B_{ijk}^{n}(\tau')$，其中 P_{ijk}，Q_{ijk} $(i=0,1,\cdots,n;j=0,1,\cdots,n-i;i+j+k=n)$ 为两个 n 次相邻曲面的控制点，其中 $\tau=(u,v,w)$，定义 $\tau_c=(0,v,w)=(0,v',w')$，则公共边界线为 $p(\tau_c)=p'(\tau_c)$，如图 2 所示。

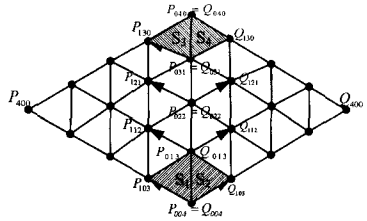

图 2　两个 4 次三角 Bézier 曲面片的拼接

Farin 等人在文献[18]中获得了三角 Bézier 曲面沿边界 GC^1 连续的充分条件，即对公共边界上的任何点均有：

①$P_{001}^{n-1}(\tau_c)$，$P_{010}^{n-1}(\tau_c)$，$P_{100}^{n-1}(\tau_c)$，$Q_{100}^{n-1}(\tau_c')$ 四点共面；

②$P_{001}^{n-1}(\tau_c)$，$Q_{100}^{n-1}(\tau_c')$ 两点位于直线 $\overline{P_{001}^{n-1}(\tau_c),P_{010}^{n-1}(\tau_c)}$ 的两边。

其中 $P_{001}^{n-1}(\tau_c)$，$P_{010}^{n-1}(\tau_c)$，$P_{100}^{n-1}(\tau_c)$，$Q_{100}^{n-1}(\tau_c')$ 四点是两曲面相对于边界某点执行 $n-1$ 次 deCasteljau 算法后得到的点。条件①保证了在拼接边界上存在公共切平面，条件②保证了曲面在公共边界处不出现奇异情况。该拼接条件可以表达为如下形式：

$$\begin{cases} Q_{1,0,n-1}=k_1P_{0,0,n}+k_2P_{0,1,n-1}+k_3P_{1,0,n-1} \\ Q_{1,i,n-1-i}=\dfrac{n-i}{n}(k_1P_{0,i,n-i}+k_2P_{0,i+1,n-1-i}+k_3P_{1,i,n-1-i})+ \\ \quad\dfrac{i}{n}(k_4P_{0,i,n-i}+k_5P_{0,i+1,n-1-i}+k_6P_{1,i,n-1-i}) \\ Q_{1,n-1,0}=k_4P_{0,n-1,1}+k_5P_{0,n,0}+k_6P_{1,n-1,0} \end{cases}$$

其中 $k_i(i=1,\cdots,6)$ 为实数，且 $k_1+k_2+k_3=1$，$k_4+k_5+k_6=1$。这一条件对于二次曲面来说是充分必要的。

而根据重心坐标定义，Farin 还给出了 GC^1 连续的必要条件：公共边界两端顶点处相邻的两对三角形面积比相等[2]（参见图 2 的阴影部分）。

此外，根据 Bernstein 基函数的独立性，充分条件还可以简化为：$\alpha A_i+\beta B_i=\lambda C_i$，其中 α,β,λ 分别为系数多项式，$A_i=P_{1,i,3-i}-P_{0,i,4-i}$，$B_i=Q_{1,i,3-i}-Q_{0,i,4-i}$，$C_i=P_{0,i+1,3-i}-P_{0,i,4-i}$ $(i=0,\cdots,4)$，这说明当两张三角 Bézier 曲面的 GC^1 光滑拼接条件仅与曲面的边界控制点和相邻一排控制点有关。

对于 α,β,λ 的取值，文献[17]中给出了一种确定方法，将重心坐标变换成直角坐标，将每个控制顶点表示为三角形域三个顶点的关系式，最后推导出 α,β 分别为 $\alpha=S'/\Delta$，$\beta=S/\Delta$，其中，Δ 表示三角形 $P_1P'_1P_2$ 的面积，S 为曲面 P 的面积，S' 为曲面 P' 的面积，其几何意义为将两三角曲面在 xoz 平面投影，如图 3 所示。

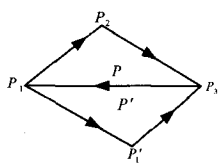

图 3　两个相邻三角域

此外，文献[19]给出了 α,β,λ 的另一种取法，将 α,β,λ 设为关于 τ^0 的函数：

$\alpha=\alpha_1\tau_2+\beta_1\tau_3$，$\beta=\alpha_2\tau_2+\beta_2\tau_3$，$\lambda=\alpha_3\tau_2+\beta_3\tau_3$

这种取法更加灵活、自由，但需要考虑的系数也相应增加。文献[20]利用曲面细分的局部设计方法构造光滑曲面，内部控制顶点的系数多项式取 $\alpha=\alpha_1$，$\beta=1$，$\lambda=\tau_2\lambda_1+\lambda_2\tau_3$。刘鼎元和 Hoschek 在文献[21]中讨论了三边 Bézier 曲面片与四边 Bézier 曲面片的 GC^1 连续问题。

3.2　三角 B 样条曲面的光滑拼接

三角 B 样条曲面既可以表达定义在平面上非规则三角形参数域中的任意复杂的几何形状，还可以被拓展到球面和具有任意拓扑的流形域上。对于给定的空间点 P_{ij} $(i=0,1,\cdots,m;j=0,1,\cdots,n)$，三角 B 样条曲面可生成一张由 $(m-1)\times(n-1)$ 个曲面元组成的曲面，其中曲面元 R_{ij} 定义为 $R_{ij}(u,v)=UAB_{ij}A^TV^T$，$U=(1,\cos u,\sin u,\cos 2u)$，$V=(1,\cos v,\sin v,\cos 2v)$，

$$B_{ij}=\begin{pmatrix} P_{i-1,j-1} & P_{i-1,j} & P_{i-1,j+1} \\ P_{i,j-1} & P_{i,j} & P_{i,j+1} \\ P_{i+1,j-1} & P_{i+1,j} & P_{i+1,j+1} \end{pmatrix}$$

$$A=\frac{1}{4}\begin{pmatrix} 3 & -2 & 3 \\ 0 & 4 & -4 \\ -4 & 4 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

从定义中可以看到，每个曲面元都由 9 个控制顶点来定义，文献[22]定义了两个参数：$\alpha=1$，$\beta=2$，证明了此时两个曲面元在边界处满足：$R''_{ij}(\pi/2,v)=\alpha^2R''_{ij+1}(0,v)+\beta R'_{ij+1}(0,v)$，其中 $R''_{ij}(\pi/2,v)$，$R'_{ij+1}(0,v)$ 为两个曲面元边界曲线的二阶导数，$R'_{ij+1}(0,v)$ 为一阶导数，即此时三角 B 样条曲面元在边界处可以达到 GC^2 连续。此外，在达到同样连续要求的情况下，三角 B 样条只需要张量积曲面的一半次数，所以，它在构造光滑混合曲面方面具有很大的潜力[23]。

结束语　本文对基于矩形域和三角域的参数曲面的 GC^1 拼接条件进行了分析和讨论。在实际 CAGD 的造型过程中，客体表面的表示比较复杂，通常很难直接用两张参数曲面来完全予以表示，比如通常先经过细分将曲面分为多个矩形域或三角域曲面片，然后再进行多个面片间的光滑处理。目前在这一领域以下几方面的研究工作值得关注：(1) 矩形域或三角域曲面片有着不同的几何拓扑结构，具有任意拓扑结构适应能力的光滑拼接技术一直是该领域的热点和难点问题，同时也是实际应用中亟待解决的问题；(2) 基于混合函数的多片曲面光滑拼接技术是实现曲面光滑拼接的重要手段，如何构造适应光滑曲面拼接，特别是高阶连续拼接的混合函数，也是一个值得重视的研究内容；(3) 三角 B 样条曲面以其优越的连续性质而受关注，有关三角 B 样条曲面的多片光滑拼接技术的研究也将是一个研究的重要内容。

参 考 文 献

[1] 施法中. 计算机辅助几何设计与非均匀有理 B 样条[M]. 北京：北京航空航天出版社，1994：306-351

[2] 朱心雄. 自由曲线曲面造型技术[M]. 北京：科学出版社，2000：206-211

[3] Bézier P. Numurical Control：Mathematics and Applications. Wiley，translated by Forrest R，1972

[4] Boehm W，Farin G，Kahmann J. A survey of curve and surface methods in CAGD. CAGD，1984(1)：1-60

[5] Liu D, Hosechek J. GC^1 contunuity conditions between adjacent rectangular Bézier surface patches. Computer Aided Geometric Design, 1989, 21(4): 194-200
[6] 白鸿武. 双三次 Bézier 曲面片光滑拼接条件的一个推导[J]. 咸阳师范学院学报, 2004, 12(6): 6-7
[7] DeRose T D. Necessary and sufficient conditions for tangent plane continuity of Bézier surfaces[J]. CAGD, 1990, 7(1/4): 165-180
[8] Liu D. GC^1 continuity conditions between two adjacent rational Bézier surface patches[J]. Computer Aided Geometric Design, 1990, (7): 151-163
[9] Degen W L F. Explicit continuity conditions for adjacent Bézier surface patches[J]. Computer Aided Geometric Design, 1990, 7(20): 181-189
[10] 曲学军, 宁涛, 席平. B 样条曲面的光滑拼接[J]. 计算机辅助几何设计与图形学学报, 2004, 16(1): 138-141
[11] 施锡泉, 赵岩. 双三次 B 样条曲面的连续条件[J]. 计算机辅助几何设计与图形学学报, 2002, 14(7): 676-682
[12] Du W H, Francis J M. On the G^1 continuity of piecewise Bézier surface: A review with new results[J]. Computer-Aided Geometric Design, 1990, 22(9): 556-573
[13] 周西军, 杨海成. NURBS 曲面 G^1 光滑拼接算法[J]. 计算机辅助几何设计与图形学学报, 1996, 8(3): 227-233
[14] Konno K, Tokuyama Y, Chiyokura H. A G^1 connection around complicated curve neshes using C^1 NURBS Boundary Gregory Patches[J]. Computer Aided Geometric Design, 2001: 293-306
[15] 赵庶丰. NURBS 曲面 G^1/G^2 光滑拼接方法[J]. 工程图学学报, 2003(2): 105-115
[16] Lai M J. Geometric interpretation of smoothness conditions of triangular polynomial patches[J]. Computer Aided Geometric Design, 1997, 14(2): 191-199
[17] 丁金扣. 三角域上 B-B 插值曲面片的拼接条件[J]. 北京邮电大学学报, 1994, 17(1): 71-78
[18] Farin G. Triangular Bernstein-Bézier Patches, CAGD, 1986, 3(2): 83-127
[19] 王相海. 三角 Bézier 曲面的一种 GC^1、GC^2 混合及一类隐式代数曲面参数话研究. 博士论文. 长春: 吉林大学, 1999
[20] 赵东福. Bézier 三角组合曲面的局域设计[J]. 工程设计学报, 2002, 12(5): 261-264
[21] Liu D Y, Hoschek J. GC^1 continuity condition between adjacent rectangular and triangular Bézier surface patches[J]. Computer-Aided Geometric Design, 1989, 21(4): 194-200
[22] 吴晓勤, 唐运海. 曲率连续的三角 B 样条曲线与曲面[J]. 计算机应用与软件, 2005, 22(1): 118-120
[23] Greiner G, Seidel H P. Modeling with triangular B-splines. IEEE Computer Graphics and Applications, 1994, 14(2): 56-60

(上接第 180 页)

本主题的基础之上，这从一定程度上降低了描述文档所需的词汇量，起到了文档内容压缩和降维的作用。本文采用文档压缩率(R)表示文档内容被压缩的程度，其计算方法如下：

$$R=n_f/l_d \tag{22}$$

其中，n_f 表示文档特征词的数量；l_d 表示文档中不同词项的数目。在 VSM 中，$n_f=l_d$；在 TPDC 中，$n_f=n_{\text{topic}} * l_{\text{topic}}$，显而易见，$R$ 越小，用于表示一篇文档的特征词越少，文档被压缩的程度越大。

在表 1 所列的参数条件下，对 20 Newsgroups 中所有 20,000条新闻记录按照方程(22)计算压缩率，按组平均结果如表 2 所示。

表 2 压缩率比较

20 Newsgroups	R_{VSM}	R_{PDC}
alt. atheism	1.00	0.50
comp. graphics	1.00	0.59
comp. os. ms-windows. misc	1.00	0.70
comp. sys. ibm. pc. hardware	1.00	0.68
comp. sys. mac. hardware	1.00	0.73
comp. windows. x	1.00	0.58
misc. forsale	1.00	0.92
rec. autos	1.00	0.64
rec. motorcycles	1.00	0.73
rec. sport. baseball	1.00	0.67
rec. sport. hockey	1.00	0.66
sci. crypt	1.00	0.42
sci. electronics	1.00	0.64
sci. med	1.00	0.54
sci. space	1.00	0.50
soc. religion. christian	1.00	0.44
talk. politics. guns	1.00	0.44
talk. politics. mideast	1.00	0.37
talk. politics. misc	1.00	0.40
talk. religion. misc	1.00	0.48
Average	1.00	0.58

实验表明，与 VSM 模型相比，TPDC 模型有着更小的压缩率。如果忽略特征词长度的区别，假定存储每个特征词所需的空间是相同的，则由表 2 可以看出，在文档的特征词存储方面，TPDC 模型的空间开销远小于 VSM，平均约为 VSM 的 58%。因此，TPDC 模型对文档内容的压缩能力明显优于 VSM。

结束语 本文提出一个 TPDC 模型，将文档的相关性建立在文档后验概率的基础之上，并通过概率推理和合理近似，把求解文档之间的相关性转化为计算主题向量之间的相似性，使问题得以简化和解决。实验结果显示，与 VSM 模型相比，TPDC 模型主要有两方面的优点：1)TPDC 模型有较高的检索精度；2)TPDC 模型存储文档特征词的空间开销较少。因此，TPDC 模型在文档检索中有更好的应用前景。下一步工作将重点研究如何根据具体文档自动选取最佳的主题个数和主题长度，并通过更多的语料库实验检测 TPDC 模型的性能。

参 考 文 献

[1] Salton G, McGill M J. Introduction to modern information retrieval. New York: McGraw-Hill, 1983
[2] Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval. New York: ACM Press and Addison Wesley, 1999
[3] van Rijsbergen C J. Information retrieval. London: Butterworths, 1979
[4] Becker J, Kuropka D. Topic-based vector space model// Proceedings of Sixth International Conference on Business Information System. Colorado Springs, 2003: 7-12
[5] Wan Xiao-jun, Peng Yu-xin. A new retrieval model based on TextTiling for document similarity search. Journal of Computer Science and Technology, 2005, 20(4): 552-558
[6] Hearst M A. Multi-paragraph segmentation of expository text// Proceedings of 32nd Meeting of the Association for Computational Linguistics. Los Cruces, 1994: 9-16
[7] Lovasz L, Plummer M D. Matching Theory. Amsterdam: Elsevier Science Publishers B V, 1986
[8] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993-1022
[9] Griffiths T L, Steyvers M. Finding Scientific Topics// Proceedings of the National Academy of Sciences. 2004: 5228-5235