第 33 卷 第 2 期 计 算 机 工 程 2007 年 1 月
Vol.33 No.2 Computer Engineering January 2007
·软件技术与数据库· 文章编号：1000—3428(2007)02—0090—04 文献标识码：A 中图分类号：TP311

空间连接优化方法的研究

姜素芳，陈天滋
(江苏大学计算机科学与通信工程学院，镇江 212013)

摘 要：基于 MBR 及直接查询谓词，提出了能够优化多路 R 树连接筛选阶段的加权处理方法，扩展了 R 树结构及 MRJ 算法。使用该方法能够得到更加有效的候选集，减少磁盘访问次数，节省了 CPU 及 I/O 的时间开销，通过实例验证了其在空间数据库查询优化方面的优势。
关键词：空间连接；多路 R 树连接；派生谓词；查询图；加权处理

Research on Optimization Method of Spatial Join

JIANG Sufang, CHEN Tianzi
(School of Computer Science and Telecommunications Engineering, Jiangsu University, Zhenjiang 212013)

【Abstract】 Based on the MBR and the direct query predicate, this paper proposes a weighted processing method which mainly contributes to optimize the filter step of the multi-way R-tree join and extends the structure of R-tree and the MRJ algorithm. This method contributes to get a set of more effective candidates and then reduce the frequency of disk access and the overhead of CPU as well as I/O. Through the experiment of specific application by using this method, the paper shows the significant superiority in optimizing the spatial query of SDBMS.
【Key words】 Spatial join; Multi-way R-tree join; Derived predicate; Query graph; Weighted processing

1 概述

空间连接是指从两个空间数据集 R、S 中检索出所有满足空间谓词θ(如交、包含等)的空间对象对(O，O')∈R×S。它是空间查询中最常见、最重要的基本操作之一，基于 R 树的空间连接(RJ)是一种高效的处理机制。但由于空间对象的表达高度复杂且数据海量，需要大量的磁盘存取操作和复杂的几何计算，使得处理开销很大。因此为提高空间查询的效率，一般将该过程分 2 个阶段进行：(1)筛选（filter）——采用简单的数据结构近似表示空间对象（最常见为 MBR），而后进行空间连接运算，产生候选集进行下一步操作；(2)精化（refinement）——以候选集为输入，逐一检查每组真实的空间对象，输出查询结果。

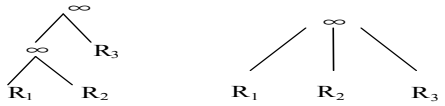
多路空间连接[1]是指从M(M>2)个空间数据集中检索出满足查询条件Q的空间对象对，用查询图可表示为

Q (V,E)

其中：$V=\{R_i \mid 2<i\le M\}$表示空间关系；

$E=\{Q_{i,j} \mid \forall i,j: r_i\in R_i, r_j\in R_j$且$r_iQ_{i,j}r_j=true\}$表示$R_i$与$R_j$之间的连接谓词。

Min J.K.等根据查询图的特征将其分为 4 类：完全图，环图，回路图，无圈(树形或星形)图。图 1 是处理多路空间连接的 2 种方法，其中图 1(b)是图 1(a)在多路连接中的扩展，它同步遍历具有相同连接谓词的关系，能够很快得到查询结果，在紧凑查询图中尤为高效。目前对MRJ优化算法的研究很多，但大多只是对关系或元组的连接顺序[4]及代价模型[5]进行改进。本文利用R树结构中的MBR和直接连接谓词，将原有查询图改造为加权完全查询图，扩展了R树结构及MRJ算法，使多路R树连接在筛选阶段得到范围更小的候选集，改善MRJ性能，从而提高空间连接效率，减少系统开销。



(a)顺序 RJ 连接 (b) MRJ(M-way R-tree join)连接

图 1 多路空间连接的解决方案

2 相关概念

2.1 常见 2 路空间连接

设R_A、R_B为参与连接的两个关系，根据R_A、R_B是否建有索引，可将空间连接分为 3 种情形：(1)R_A、R_B都有索引；(2)仅其中一个有索引；(3)都无索引。在多数据集连接时，2 个都有索引的数据集连接后产生的中间结果并无索引，这时可将情形(1)转为情形(2)或情形(3)处理。3 种情形的相关算法详见表 1。

表 1 空间连接算法的分类

双索引	单索引	无索引
1z-值变换(transformation to z-values) 2 空间连接索引（Spatial Join Index） 3 基于 R 树的空间连接算法 4 基于广度优先搜索的改进的 R 树连接算法	1 嵌套索引循环（Index Nested Loops） 2 种子树连接（Seeded Tree Join） 3 分类归并（Sort and Match） 4 槽索引空间连接（Slot Index Spatial Join）	1 空间哈希连接（Spatial Hash Join） 2 基于划分的空间归并连接（Partition Based Spatial Merge Join） 3 大小分割空间连接（Size Separation Spatial Join） 4 基于扫描的可扩展空间连接（Scalable Sweeping-based Spatial Join）

基金项目：信息产业部基金资助项目(2003xk320014)
作者简介：姜素芳(1981－)，女，硕士生，主研方向：计算机图形学，GIS；陈天滋，副教授
收稿日期：2006-01-24 **E-mail：**jsfvbx@126.com

2.2 多路 R 树连接(MRJ)

MRJ是二路RJ在多路空间连接上的扩展[2]，它同步遍历M个R树索引，检测结点的项元组是否满足查询谓词。如满足条件的元组由叶结点组成则输出，并继续处理下一项元组，若由中间结点组成则同样处理其孩子结点；如检测的项元组不满足任一查询谓词，则处理下一项元组直至处理完所有项。

2.3 空间连接的基本优化方法

设两关系R_A、R_B，如果项$E_A(E_A \in N_A \in R_A)$的MBR和$N_B(N_B \in R_B)$的MBR不相交，则E_A与N_B所包含的任何空间对象都不可能相交。图 2 中，$a_1.mbr \cap B.mbr=\Phi$，则滤去a_1；继续判断得$a_2.mbr \cap B.mbr \neq \Phi$，则将$a_2$纳入搜索域，以检测$a_2.mbr$是否与B内各项的MBR相交。搜索空间紧缩排序算法[2]就是在此基础上提出的。

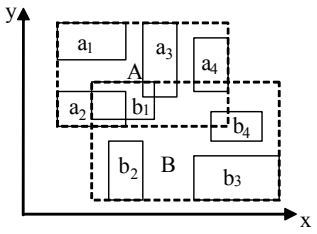


图 2 矩形相交的 R 树结点

另一优化算法是平面扫描正向检查[2]，基本思想是：设(r_ix_l, r_iy_l)和(r_ix_r, r_iy_r)分别为矩形r_i的左下角和右上角点坐标，$|r_ix_l|$和$|r_iy_l|$分别为r_i在x轴和y轴的投影。将两结点所包含的矩形按照r_ix_l从小到大进行排序，得到顺序矩形序列$R=\{r_1,r_2,\ldots,r_n\}$，沿y轴方向判断矩形是否相交。以图 2 为例，$A=\{a_1,a_2,a_3,a_4\}$、$B=\{b_1,b_2,b_3,b_4\}$，矩形序列为$\{a_1,a_2,b_1,b_2,a_3,b_3,a_4,b_4\}$，从左向右扫描，对矩形$a_1$需判断在y轴方向是否与$b_1,b_2,b_3,b_4$相交；对$a_2$判断$b_1,b_2,b_3,b_4$；对$b_1$判断$a_3,a_4$等。

本文对MRJ算法的扩展及试验使用了搜索空间紧缩算法和平面扫描正向检测算法[2]。

3 MRJ 的加权优化

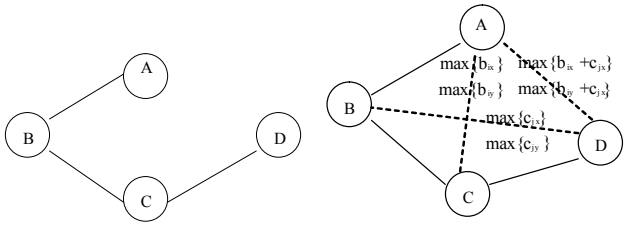
在判断项元组是否相交时存在情形[3]：(1)父结点的MBR相交，而子结点的MBR不相交；(2)结点的MBR相交，而结点项的MBR不相交，本文暂且称之为“ 误检索 ”(false retrieval)。直接连接谓词越少，此类情形出现的几率越大。对M-路空间连接，连接谓词最多可有C_M^2个，即两两关系之间都存在连接，对应于完全查询图。本文通过构造加权完全查询图尽量避免误检索发生，进一步缩小候选集范围，减少磁盘访问次数，从而优化MRJ的处理过程。

3.1 派生谓词

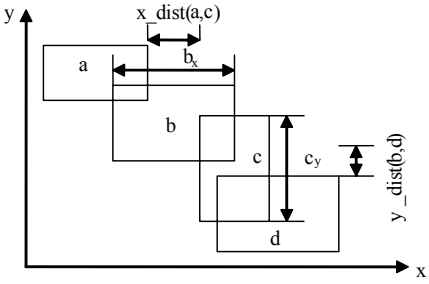
以图 3(a)中的 4 路空间连接为例，查询条件Q=A intersect B&B intersect &C intersect D，对应于边AB、BC、CD。令b_x、c_x分别为项b、c的MBR的宽度，x_dist(a,c)、y_dist(a,c)为项a、c的MBR在x轴和y轴上的距离（其余类似）。由图 3(c)可知，对叶结点符合条件的输出元组满足式(1)；对于非叶结点，本文取其包含项的最大x、y值作比较，对符合条件的元组应满足式(2)。

$$\begin{cases} x\_dist(a,c) \le b_x、\ x\_dist(b,d) \le c_x、\ x\_dist(a,d) \le b_x + c_x \\ y\_dist(a,c) \le b_y、\ y\_dist(b,d) \le c_y、\ y\_dist(a,d) \le b_y + c_y \end{cases} \quad (1)$$

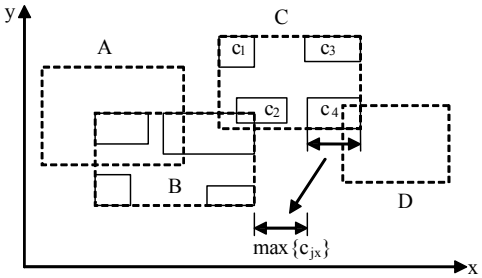
$$\begin{cases} x\_dist(a,c) \le \max\{b_{ix} \mid b_i \in B\}, x\_dist(b,d) \le \max\{c_{jx} \mid c_j \in C\} \\ x\_dist(a,d) \le \max\{b_{ix}\} + \max\{c_{jx}\} \\ y\_dist(a,c) \le \max\{b_{iy} \mid b_i \in B\}, y\_dist(b,d) \le \max\{c_{jy} \mid c_j \in C\} \\ y\_dist(a,d) \le \max\{b_{iy}\} + \max\{c_{jy}\} \end{cases} \quad (2)$$



(a)查询图　(b)派生连接谓词查询



(c)查询结果即 MBR 相交的元组



(d)MBR 相交的 R-树非叶结点

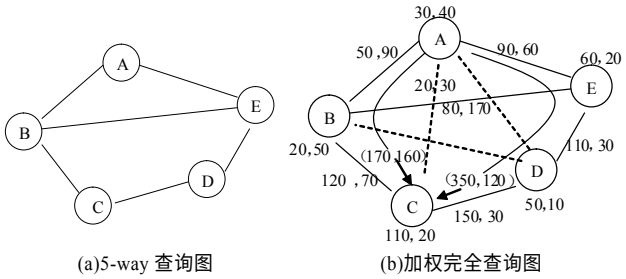
图 3 派生谓词查询图

图 3(d)中x_dist(B,D)>max{$c_{jx}|c_j \in C$}(即c_{4x})，表明元组<A,B,C,D>不满足查询谓词，则跳过该元组及其子项继续筛选其他元组，节省了系统开销。由于x_dist(B,D)≤max{$c_{jx}|c_j \in C$}等由给定查询谓词推导而得，本文称之为派生谓词，用它构造的完全查询图如图 3(b)所示。

3.2 加权完全查询

5-way空间连接(图 4(a))加权完全查询图的构造过程：(1)点权值：取R树索引中某结点所有项的MBR分别在x轴、y轴上的最大投影<x_maxlen,y_maxlen>；(2)直接谓词连接边的权值：连接边两端点分别在x、y轴上的权值之和，即 <$N_{ix\_maxlen}+N_{jx\_maxlen}$,$N_{iy\_maxlen}+N_{jy\_maxlen}$>。由图 4(b)可知，派生谓词的生成路径可有几种选择，如dist(A,C)可以是ABC或AEDC，本文中取最短派生路径，并计算得该路径长度（SPL－shortest path length）；(3)派生谓词连接边的权值：最短派生路径的长度减去路径两端点权值后再除以 2，如式(3)所示。

$$\begin{cases} x\_SPL = \left(SPL\left(N_{ix}, N_{jx}\right) - \max\{N_{ix}\} - \max\{N_{jx}\}\right)/2 \\ y\_SPL = \left(SPL\left(N_{iy}, N_{jy}\right) - \max\{N_{iy}\} - \max\{N_{jy}\}\right)/2 \end{cases} \quad (3)$$



(a)5-way 查询图　(b)加权完全查询图

图 4 加权完全查询图

下面描述图 4(b)中派生谓词边 AC 权值的求取过程。路径 ABC、AEDC 的长度分别为(170，160)、(350，120)，它们在 x 轴和 y 轴上取得的最短长度分别为 170、120，由派生边权值计算式(3)可得 AC 边的权值为（20，30)。图 4(b)中所有派生边的权值见表 2。

表 2 派生谓词连接边权值

派生边	x 轴最短路径	x 轴最短路径长度	y 轴最短路径	y 轴最短路径长度
AC	ABC	20	AEDC	30
AD	AED	60	AED	20
D	BED	60	BCD	20
CE	CBE	20	CDE	10

图 5 中x_SPL(A,D)=70>max{e_{ix}|e_i∈E}=60，则可知含有 A，D项的组合将不会产生符合查询条件的最终结果，这样就可避免误检索操作。

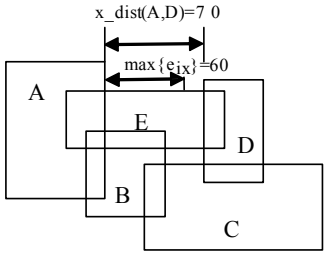


图 5 5-路空间连接查询

4 R 树结构的调整

由于在构造加权完全查询图时需要知道 MBR 在 x 轴和 y 轴的最大投影，本文对 R 树结构作了调整，引入了最大 x，y 信息量。对于叶结点，x_maxlen，y_maxlen 为结点中所有项的 MBR 在 x 轴和 y 轴上投影的最大长度；对于非叶结点，x_maxlen，y_maxlen 为子结点的 MBR 在 x 轴和 y 轴上投影的最大长度；根结点则拥有 R 树所有结点的 MBR 在 x 轴和 y 轴上投影的最大长度。设 n 为结点 N 所容纳的项数，式(4)、式(5)提供了各结点最大 x、y 的取值方法：

$$x\_\max(N)=\begin{cases}\max\{N_1.mbr_x,\ldots,N_n.mbr_x\}\\ \text{N为叶子结点}\\ \max\{x\_\max len(N_1.mbr),\ldots,x\_\max len(N_n.mbr)\}\\ \text{N为非叶子结点}\end{cases}\quad(4)$$

$$y\_\max(N)=\begin{cases}\max\{N_1.mbr_y,\ldots,N_n.mbr_y\}\\ \text{N为叶子结点}\\ \max\{y\_\max len(N_1.mbr),\ldots,y\_\max len(N_n.mbr)\}\\ \text{N为非叶子结点}\end{cases}\quad(5)$$

相应的 R 树结构调整如图 6。

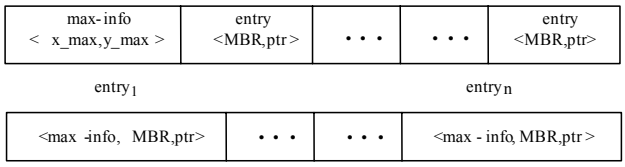


图 6 带有最大x、y值的R树非叶结点和结点

```
Typedef struct RTreeNode
{ Int x_max,y_max;
  Entry e;
  ….
}RTreeNode;
Typedef struct Entry
{ Int x_max,y_max;
  Rect MBR;
  Point ptr;
  …
}Entry;
```

结点的最大 x、y 值是其子结点的 MBR 在 x、y 轴上投影的最大值，也就是说由中间结点构成的最短路径的长度与由根结点构成的最短路径的长度近似相等，因此一般情况下只需计算根结点的最短路径的长度即可，这样可简化最短路径的求取过程。

5 MRJ 算法的扩展

引入派生谓词后，通过对派生谓词连接边的点对的检测，可减少误检索率，进一步缩小搜索空间，从而优化 MRJ。扩展后的 MRJ 算法如下：

```
Ex_Space-Restriction(Query Q[][], RTNode N[], int i)
{
read Ni;
Di = Φ;          //搜索空间初始为空
for(对连接关系R内的所有结点Nj)     //构造加权完全查询图的//同时滤去不会产生有效结果的元组
{
If(Qij = FALSE)     //处理没有直接连接谓词的边
{
  x_Dist(Ni,Nj) = Nj.mbr_xl - Ni.mbr_xu;   //设Ni,Nj符合沿x轴//扫描线的正向检查
          y_Dist(Ni,Nj) = Nj.mbr_yl - Ni.mbr_yu;
//计算(Ni,Nj)最短路径 x、y 方向上的长度
x_SPL=(SPL(Nix,Njx)-max{Nix}-max{Njx})/2;
y_SPL=(SPL(Niy,Njy)-max{Niy}-max{Njy})/2;
          if((x_Dist(Ni,Nj) ≤ x_SPL(Ni,Nj))  or  (y_Dist(Ni,Nj) ≤ y_SPL(Ni,Nj)))
          Qij = TRUE;     //判断符合则派生的谓词连接边有效
}
 }
for (结点Ni内的各项Ei)
{
valid = true;          //标记项 Ei 需进一步处理
for(对连接关系R内的所有结点Nj)
{
if (Qij = TRUE)
{
If (Ei,x∩Nj.MBR =Φ
{
valid = false;
break;
                    }
}
if (valid=true)
Di = Di∪Ei;
          }
}
return Di;
}
```

6 试验

本文以实际开发的某应用系统中的空间连接查询（统计卢湾区内与建国中路相交的所有路口上的标杆）为例，验证了本文提出的方法在空间数据库查询优化方面的优势。

表 3 列出了 4-way 空间连接查询的数据集属性及查询结果，其中‖R‖为空间关系的 R 树叶节点数。该查询条件为：(1)找出卢湾区内所包含的建国中路；(2)找出与建国中路相交的所有路口；(3)找出与路口相交的所有标杆。图 7 为该空间

连接查询对应的查询图，图 8 给出了使用加权处理方法前后的时间花费比较，图 9 为在电子地图上定位的实际查询结果。

表 3 实验数据集的特征及查询结果

空间关系 R	‖R‖	属性
R_A	19	行政区 Administrative_region
R_R	6 528	道路 Road_Class
R_I	14 617	路口面 RoadIntersectionPanel_Class
R_P	112 753	标杆 Pole_Class
查询结果	33	

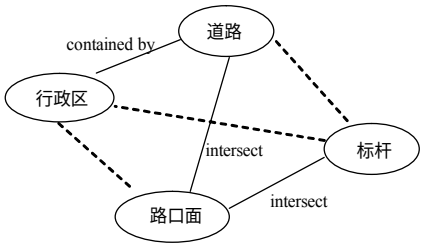


图 7 空间连接查询图

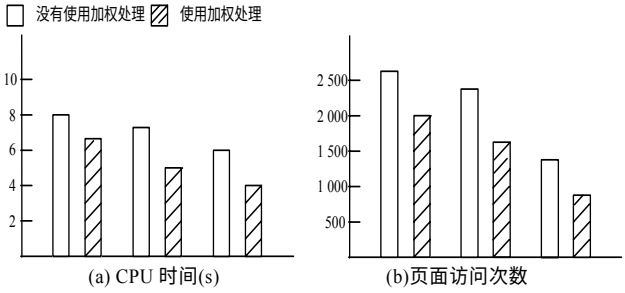


图 8 加权处理前后比较

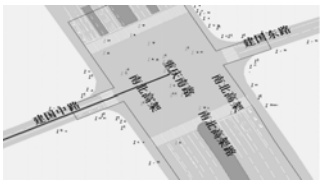


图 9 实验效果

7 总结

本文基于空间对象的 MBR 及直接查询谓词，提出了优化 MRJ 筛选阶段的加权处理方法，描述了 MRJ 算法及其基本优化措施，并给出了调整后的 R 树结构和扩展的 MRJ 算法。该方法能得到更加有效的候选集，减少磁盘访问，节省了 CPU 及 I/O 的时间开销，从而改善 MRJ 的性能。最后本文通过具体应用验证了其在空间数据库查询优化方面的优势。

参考文献

1 Mamoulis N, Papadias D. Multiway Spatial Joins[J]. ACM Trans. on Database Syst., 2001, 26(4): 424-475.

2 Park H H. Early Separated Filter/Refinement Strategies and Multiway Spatial Joins for Spatial Query Optimization[D]. Dept. of Electrical Engineering & Computer Science, Division of Computer Science, 2001.

3 Parka H H, Minb J K, Chungb C W, et al. Multi-way R-tree Joins Using Indirect Predicates[J]. Information and Software Technology, 2004, 46(11): 739-751.

4 李立言，秦小麟. 空间数据库中连接运算的处理与优化[J]. 中国图象图形学报, 2003, 8(7).

5 蒋苏蓉. 空间查询优化[J]. 计算机工程与应用，2004，40(9): 188-190.

(上接第 89 页)

在油田中，有些信息是需要保密的，国家的资产动辄上千万，内网数据库如受到攻击，后果不堪设想，所以在内外网之间架设 GAP 是完全必要的，网络结构如图 2 所示。该方案特点是：

(1)支持异构数据库，支持数据库间单向、双向两种交换模式。支持 Oracle、SQL Server、SYBASE 等绝大多数关系型数据库，能很好地适应油田企业应用环境复杂多变的特点。

(2)只允许外网数据库与核心数据库部分指定表的数据交换，内部网络不接受任何来自外网的其他服务请求。

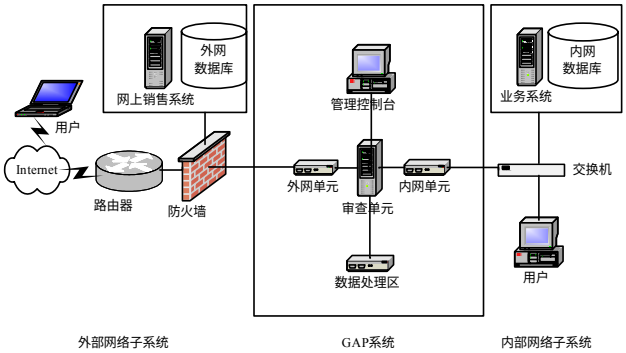


图 2 项目结构

4 系统开发的实现与结论

吐哈油田子公司外网是 SQL Server 数据库，内网是 ASA 数据库。基于以上的异构数据源信息安全交换方案，为其开发了销售考核综合系统。该系统包括油气销售、油气储运、物资、财务、辅助决策和人事管理子系统，将各加油站、员工和物资等情况进行及时的统计和分析，自动生成各种报表，供公司管理层及时掌握公司销售运营情况，对油品、人员进行合理调配，实现企业资源最优配置和效益最大化。并且可以根据各加油站销售业绩等情况，自动对加油站、加油工进行多方位业绩考核。

该系统从投入运行后，运行效果良好。开发实例表明，基于 XML、数字签名和网络隔离技术的安全信息交换方案与传统方法相比，具有通用性、安全性等优点，有效地降低了异构数据交换的难度，提高了信息集成的效率。

参考文献

1 Zhou E Z. XML and Data Exchange for Power System Analysis [J]. IEEE Power Engineering Review, 2000, 20(4): 66-68.

2 W3C. Extensible Markup Language (XML)1.0 (Second Edition) W3C Recommendation[EB/OL]. 2000-10. http://www.w3. org/TR/2000/ REC-xml -20001006.

3 李遵朝，徐颖强，绕　元，等. 基于 XML 的异构数据库间信息安全交换[J]. 计算机工程与应用, 2005, 41(13): 163-165.

4 张　勇，冯玉才. XML 数字签名技术及其在 Java 中的具体实现[J]. 计算机应用, 2003, 23(9): 93-95.

5 陈赫贝，阮　飞. XML 数字签名及其应用研究[J]. 微机发展, 2005, 26(5): 53-55.

6 朱　勤，陆建新，陈继红. 基于 XML 的异构数据交换技术及其 Java 实现[J]. 计算机应用与软件, 2004, 21(11): 52-53.