# 基于视觉相似性的中文古籍内容检索方法*

施伯乐，张 亮，王 勇，陈智峰

(复旦大学 计算机科学与技术系，上海 200433)

E-mail：{bshi,zhangl,wangyyong}@fudan.edu.cn

http://www.fudan.edu.cn

**摘要**：人类文化遗产的数字化应用是数字图书馆计划的重要组成部分.目前,数字化手书中文古籍尚缺乏有效的内容检索手段.提出了一种基于视觉相似性的计算机古籍内容检索方法,研制出关键支撑技术.该方法提取视觉对象的形态特征、全局位置特征和页面特征,采用高维空间索引技术组织形态特征构成的特征空间,完成视觉相似对象的快速检索,定义精度控制参数,动态调整由形态到语义的映射,借助约束验证技术提高一组相关对象的检索精度.原型系统证实了新方法的可行性,获得了直接在数字化图像上自动完成古籍内容检索的技术效果.

**关键词**：基于内容的图像检索；手写体中文；特征提取；空间索引；古籍检索

**中图法分类号**：TP311 **文献标识码**：A

人类文化遗产具有极高的学术研究价值和艺术欣赏价值,其数字化应用已引起图书馆情报界和计算机界的共同关注[1~3].中国具有悠久的历史和灿烂的文化,开发这些珍贵资源是我国数字图书馆工程的主要目标之一.数字化古籍内容的有效检索是其中亟待解决的技术问题.

通行的计算机古籍检索方法可分为两类:标引方法和附带文本文件方法.标引方法(包括纯粹的页面图像浏览方法[3])广泛应用于图书馆领域.标引是指对古籍特定项目,如书名、著者、出版信息、批校者、题跋者、藏印等建立索引,用户根据这些预定检索点访问页面图像.该方法的主要缺点是检索点仅限于少量的特定标引项目,难以满足用户广泛的检索要求.附带文本文件方法首先借助OCR(optical character recognizer)建立与特定古籍对应的文本文件[1,2],对该文本文件应用全文检索技术,再由对应关系得到古籍页面图像.由于中文古籍的形式主要为毛笔手书,手写汉字笔划模糊、不规范,笔划倾角、位置和相对长度的变化、书写风格的差异、软笔笔划变形等诸多因素使得中文古籍的OCR自动识别异常困难;过早地在OCR阶段"冻结"形态至语义的映射,无端地限定了检索精度或准确度;不同年代和不同版本的古籍用字差别增加了建立所需字典的难度;无法自动判定附带文本与古籍原稿内容的同一性,降低了检索的可信度;通假字在古籍中的广泛使用也使全文检索技术难以有用武之地.因此,需要独辟蹊径,开发新的自动化古籍内容检索方法与技术.

我们认为,古籍内容检索的关键是汉字及符号对象(简称"对象")的自动匹配.这种匹配可以建立在对象的视觉特征上,脱离具体字、词的语言学语义而完成.因此,本文借助基于内容的图像检索技术,提出了一种直接在页面图像上自动完成的基于视觉相似性并支持任意检索点的计算机古籍内容检索新方法.

* 收稿日期：1999-11-24；修改日期：2000-05-26

基金项目：国家自然科学基金资助项目(69953010)；上海市自然科学基金资助项目(00ZD14006)

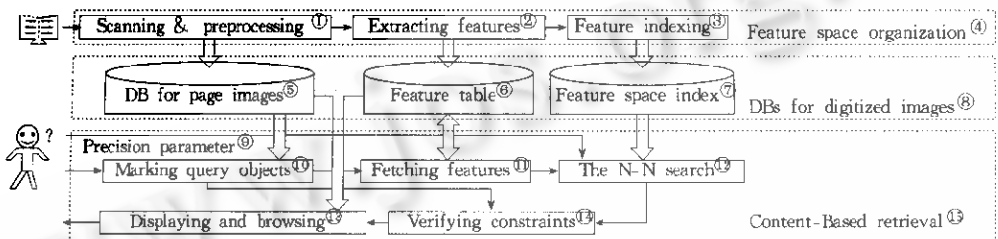
作者简介：施伯乐(1936－)，男，浙江吴兴人，教授，博士生导师，主要研究领域为数据库，知识库，面向对象数据库；张亮(1963－)，男，湖北武汉人，博士，副教授，主要研究领域为多媒体技术，支持多媒体应用的数据库技术；王勇(1971－)，男，江苏盐城人，博士生，主要研究领域为对象数据库，多媒体信息检索技术，电子商务；陈志峰(1978－)，男，上海人，主要研究领域为支持多媒体应用的数据库技术.

文章首先介绍该方法的总体思想,讨论方法的特点,接着阐述其关键技术,并以原型系统和实验结果举证方法的可行性,最后给出研究结论.

## 1 工作原理

新方法的本质是,在古籍的数字化位图上将每个对象看作一幅图像,利用基于内容的图像检索思想,提取对象的内容特征*,并应用高维空间索引来组织特征,依照最近邻原则获得与查询样本视觉相似的所有对象.

方法由特征空间组织和内容检索两个相继阶段构成,处理过程如图1所示.其中,特征空间组织一次性完成,内容检索可根据检索者的要求多次重复.




①扫描及预处理,②提取特征,③索引特征,④特征空间组织,⑤页面图像库,⑥特征表,⑦特征空间索引,⑧数字化古籍库,⑨精度控制参数,⑩标定检索样本,⑪获取特征,⑫查询近似对象,⑬显示/浏览,⑭验证约束条件,⑮内容检索.

Fig. 1 Flowcharts of the contents retrieval by visual criteria
图1 基于视觉相似性的内容检索流程图

特征空间组织的目的是为古籍内容(对象及其序列关系)生成其形态特征聚类,建立易于根据视觉相似性快速查找近似对象的索引结构.在扫描及预处理模块中,一册古籍首先变换为页面图像的有序集合,存入页面图像库.对页面图像中的污损、错位、偏斜等噪声,可通过亮度/对比度调整、版面校正、平滑滤波、二值化等预处理手段予以消除.然后,对页面图像分列、切字,分离出独立对象,应用细化算法得到线宽为单像素的骨架对象.提取特征模块将每个骨架对象变换为一组特征,保存于特征表中.同时,部分特征传递到索引特征模块,利用高维空间索引技术,根据形态特征施行良好的组织,达到视觉相似对象彼此靠近,形态差异对象相互分离的目的.上述步骤循环往复,直到一册古籍的所有对象全部进入特征空间索引结构为止.页面图像库、特征表和特征空间索引共同构成数字化古籍库.这是开展内容检索的基础.

内容检索利用业已建成的特征空间索引,迅速获得与检索者给定样本视觉语义近似的所有对象.从本质上讲,内容检索是一个基于形态特征的近似查询过程,它采用QBE(query by example)的工作方式.检索者在所浏览的页面图像上随时标定任意对象,标定检索样本模块记录点击的位置和顺序.该顺序作为约束条件传给验证约束条件模块.所记录的坐标序列被获取特征模块用作查询特征表的条件,从中提取检索样本成员的形态特征向量.查询近似对象模块按照最近邻原则分别对各形态特征向量在特征空间索引结构中搜索视觉相似对象,返回这些对象的全局位置特征.所有样本成员的搜索结果形成一个集簇,经验证约束条件模块检验后,合理的组合作为最终检索结果由显示/浏览模块呈现给用户.

采用视觉相似匹配的技术路线,复杂的古籍内容检索问题可化解为特征提取、特征空间索引、最近邻搜索和验证搜索结果这4个易于驾驭的子问题.开发并合理组合相关技术,即可实现直接在

* 一册古籍的内容特征定义为对象的全局位置特征、形态特征和页面特征.具体定义请参见第2.1节.

数字化图像上自动完成古籍内容检索的技术效果.本方法的主要特征表现在以下几个方面:

• 检索者可以在页面图像上任意标定检索样本,不受检索点的限制;

• 检索样本直接出自页面图像,毋须顾及同一性判定、字符集规模、通假字、词库等问题,自动化程度较高,操作简便;

• 在技术路线上,以"视觉相似匹配"取代"自动识别",摆脱了"识别"所引入的额外困难;

• 检索者可以方便地调整形态特征到语义的映射,权衡查全率与查准率;

• 本方法可以作为标引方法的扩充.标引方法引导检索者发现候选的古籍卷宗,基于视觉相似性的内容检索方法为检索者在卷内发现目标提供帮助.

## 2 关键技术

本节详细描述解决上述4个子问题的技术方案,它们构成了本方法的技术支撑.其他一些相关技术,如平滑、分割及符号细化等,可参照常规汉字OCR的预处理技术.

### 2.1 特征定义与提取

本方法针对单册古籍来定义和提取3类基本特征,即古籍的页面特征、对象的全局位置特征和形态特征.如果将同一人誊写的多卷古籍组合在一起进行处理,应添加书籍标识.

**定义1(全局位置特征).** 对象的全局位置特征是该对象在古籍中的线性序编号.

只要能保证对象与其全局位置特征是一一对应的,定义中的线性序就可以采取任意形式.一般情况下,线性序宜采用古籍的誊写与阅读习惯(页码从小到大,页内从右向左,各列自上而下).

**定义2(页面特征).** 古籍的页面特征由页面编号和页面内各对象的几何坐标构成.

页面特征描述了页面中对象的几何布局.扫描及预处理模块分离出的每个独立对象都包含于一个矩形区域之中.可以将该矩形区域的中点坐标作为对象的几何坐标.

汉字的书写决定了文字的语言学语义(多音字除外),所以通过对汉字形态的比较可以实现古籍的内容检索.在描述毛笔手书汉字形态特征的问题时,有许多可变因素会导致笔划或部件精确提取的失败.例如,笔划粗细不均匀、部分笔划模糊或欠落、笔划相对位置偏移、笔划倾角/相对长度变化等,都将影响对象在视觉意义上的匹配.因此,我们用笔划因素替代笔划描述对象基本成份,以多级分划区域中的笔划因素统计描述对象形态.注意到"方块汉字部件部位和比例的固定划一是长期以来汉字书法艺术的结晶"[4]这一事实,分划点选定在区域前景像素的质心.以上所谓笔划因素,是指可构成横、竖、撇、捺这4种笔划的基本元素,其点阵排列如图2(a)所示.基于笔划因素的特征构成对软笔手写汉字笔划不均匀、笔划模糊、倾角/相对长度缺乏规律等现象都具有较强的容错能力,而且还便于古籍中文字/符号对象的统一处理.

**定义3(形态特征).** 对象的形态特征是其图像在多级质心分划区域中笔划因素分量的累计值.

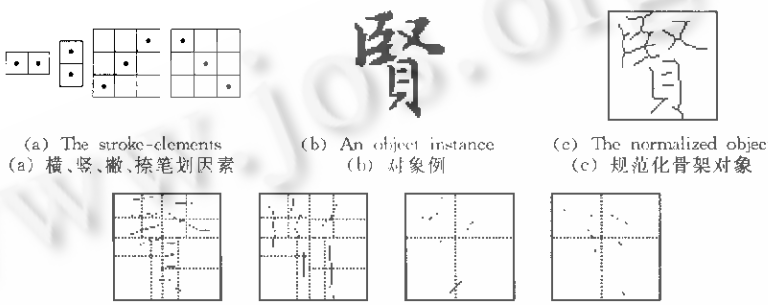
形态特征提取分3个步骤进行.首先,为排除古籍中对象尺寸变化的影响,规格化骨架图像.选对象高度和宽度的最大值为边长作一个正方形位图,使其落入位图的正中,如图2(c)所示.最小外接正方形规格化方法有助于保持对象的宽高比信息.然后,根据对象质心对区域作多级分划.每个区域的分划点定为该区域中前景像素点集的质心.最初的区域取规格化正方形,深级的分划在浅一级的基础上递归进行.依区域质心对位图作多级分划对少数笔划畸变有较好的容错性.最后,统计各区域中的笔划因素,分类累计后形成特征向量.分别以这4种笔划因素为结构元素,应用数学形态学方法对区域位图作腐蚀运算,得到4种笔划因素在各区域中的分布,再用区域中所有笔划的像素数除之,可近似得出该区域里各种笔划因素分量与笔划总数之间的比例关系.注意到汉字中横、

竖笔划的出现频度大大高于撇、捺笔划[4],同时为了降低特征空间的维数,提高索引及检索的效率,对撇、捺笔划因素的统计可以比横、竖笔划浅一个级次,即对横、竖笔划因素用二级区域分划,对撇、捺笔划因素用一级分划(如图 2(d)所示),构成 $16\times2+4\times2=40$ 维的形态特征向量 $f$.

$$f(i)=\sum_{1\leqslant k\leqslant i}\frac{h(k)}{p_2(k)},\quad f(16+i)=\sum_{1\leqslant k\leqslant i}\frac{s(k)}{p_2(k)},\quad i=1,2,\ldots,16;$$

$$f(32+j)=\sum_{1\leqslant k\leqslant j}\frac{p(k)}{p_1(k)},\quad f(36+j)=\sum_{1\leqslant k\leqslant j}\frac{n(k)}{p_1(k)},\quad j=1,2,3,4.$$

其中,$p_1(k)$和 $p_2(k)$分别为腐蚀运算前位图一级和二级划分区域 $k$ 中的前景像素点数;$h(k)$,$s(k)$,$p(k)$,$n(k)$分别为腐蚀运算后横、竖、撇、捺笔划因素在区域 $k$ 中的前景像素点数分别为腐蚀运算后横、竖、撇、捺笔划因素在区域 $k$ 中的前景像素点数.



(a) The stroke-elements (b) An object instance (c) The normalized object
(a) 横、竖、撇、捺笔划因素 (b) 对象例 (c) 规范化骨架对象

(d) Stroke distribution in partition areas
(d) 区域中笔划因素分布

Fig. 2
图 2

### 2.2 特征空间索引

以欧氏距离为度量,古籍中所有对象的形态特征向量构成 40 维的向量空间.视觉相似对象表现为具有较小距离的点集.为加速近似对象的搜索,可以设计空间索引结构,合理组织所有的特征点,以较小的存储开销换取快速的信息查询.考虑到存储效率和检索速度两方面的优势,原型系统采用 PK 树[5]作为特征空间索引结构.索引构造算法如下,详细内容请参考文献[5].

/*PK-树的每个内部节点维持着一个数组 $R[40]$,$R[i]$记录了该子树第 i 维的覆盖范围*/

1. 若 PK-树为空,则生成一个内部节点,标记为根,覆盖范围为论域;退出.
2. 生成叶节点 *child*,节点坐标为对象的形态特征向量 $f$,节点属性为对象的全局位置特征;
3. 沿根节点向叶节点的路径搜索,考察各内部节点的数组 $R$,寻找覆盖 $f$ 的最小内部节点;
4. 若这样的节点存在,标记它为 *father*,否则将路径中最后的内部节点记为 *father*;
5. 将 *child* 连为 *father* 的子节点;
6. 沿 *father* 向树根的路径逐层检验各内部节点是否满足"实例化条件"[5].相应地调整路径上内部节点的数组 $R$.

### 2.3 近似对象查询

如上所述,在所形成的向量空间中,视觉相似对象表现为具有较小距离的点集.近似对象查询是在特征空间索引中依照最近邻原则,搜索与样本成员距离小于阈值的对象集合.需重点解决两个关键问题:根据精度控制参数设定近似范围以及搜索近似范围内包含的对象.

精度控制参数 $r$ 由检索者给出,它反映了搜索的工作方式:严格或宽松.参数取值分为 $s+1$

级.第 0 级 $r=0$,表示严格匹配;第 $s$ 级 $r=s$,表示最宽松的匹配;其间按步长 1 逐步增大.下面的公式设定了近似范围宽度 $w_i$.

$$w_i=\begin{cases}\varepsilon, & r=0\\ W_i\times r/s, & 0<r\leqslant s\end{cases},\quad i=1,2,\ldots,40.$$

其中 $\varepsilon$ 是一个十分小的数,对应于严格搜索的情况;$W_i$ 是特征空间第 $i$ 维的变动范围.$w_i$ 的起止点需按照如图 3 所示的方式加以调整,以使其包含于 $W_i$ 内且 $v[i]$ 尽可能位于 $w_i$ 的中点.所有 $w_i$ 的交集构成一个超多面体 $\Omega$.

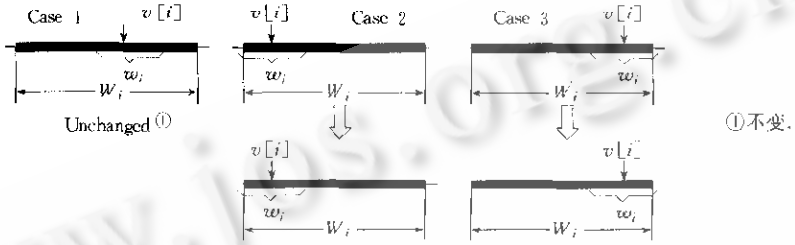



Fig. 3 Adjust $w_i$ to be contained in $W_i$ and centered with $v[i]$
图3 调整 $W_i$ 使之包含于 $W_i$ 且 $v[i]$ 尽可能位于 $w_i$ 的中点

精度控制参数 $r$ 影响近似范围 $\Omega$,而 $\Omega$ 又决定着对象是否属于查询样本成员 $v$ 的相似集合,从而 $r$ 调整了对象形态特征到语义的映射.由于内容检索可多次执行,检索者可参照上次检索结果动态地调整精度控制参数 $r$,作出查全率与查准率的新的权衡,以满足检索的需要.

空间索引结构中相似对象搜索算法如下,其中以节点名称指示以它为根的子树.

(1) 置表 $L$ 为空,将 PK-树的根节点压入栈 $S$

(2) 若栈 $S$ 非空,循环

(2.1) 弹出栈顶元素 $e$

(2.2) 对 $e$ 的所有子节点 $n$,判断 $n$ 与 $\Omega$ 的空间关系

i) 如果 $\Omega$ 完全覆盖了 $n$,则将子树 $n$ 所有叶节点的属性值并入 $L$

ii) 如果 $\Omega$ 仅部分覆盖 $n$,则将 $n$ 压入栈 $S$

(3) 返回 $L$

搜索结果 $L$ 记录了当前精度范围内某个样本成员的所有相似对象,具体是这些对象的全局位置特征.所有样本成员的搜索结果形成全局位置特征集簇.

### 2.4 验证约束条件

以样本成员的相对顺序作为约束条件(在标定检索样本模块取得),检验全局位置特征集簇中集合元素组合的合理性.设检索样本包含 $M$ 个样本成员,按其相对顺序依次记为 $e_1,e_2,\ldots,e_M$,相应的搜索结果记为 $L_1,L_2,\ldots,L_M$,具体验证过程如下:

(1) $L_1$ 赋予 $L$

(2) 循环,下标 $i$ 从 2 以增量 1 至 $M$

(2.1) 对 $L$ 中的每个元素 $e$,设其(全局位置特征)取值 $j$,如果 $L_i$ 中不存在全局位置特征为 $j+i-1$ 的元素,则将 $e$ 从 $L$ 中删去

(3) 返回 $L$

验证结束后,$L$ 保留了与第一样本成员相似的所有对象的全局位置特征.以其中的每一个元素为索引查找特征表,可以确定检索结果首元素的页面编号和页面内坐标,显示/浏览模块在页面图像上标示由此开始的连续 $M$ 个对象.

## 3 原型系统及实验

为检验本方法的可行性和评价关键技术的效果,我们开发了基于浏览器/服务器计算环境的原型系统.系统由客户端用户接口和服务器端服务进程构成,其基本结构如图4所示.


①Web浏览器,②Web服务器,③文件/数据库检索守护进程,④数字化古籍库.

Fig. 4 The basic structure of the prototype system

图4 原型系统基本结构

服务器端为3层结构:Web服务器、静态HTML文件与Java Servelet、文件/数据库检索守护进程.Java Servelet分析检索者的浏览、查询请求,调用检索守护进程获取数据,操纵页面图像,产生动态HTML文件.检索守护进程是整个运行系统的核心.它根据Servelet传达的要求访问数字化古籍库,得到结果数据.其中,Servelet与守护进程之间采用了IPC(inter-progress communication)实现数据交流.客户端界面采用支持Java通用Web浏览器,为检索者提供一个可视化的古籍浏览、检索环境.利用Java Applet技术实现了一些界面功能,如页面缩放、精度控制、检索样本标记与检索提交、检索结果标记、按标记和/或页面浏览等功能.

原型系统对《红楼梦》第3章页面图像进行检索实验.参考对象是《红楼梦》第3章正文(共7336个汉字)文件加字符串搜索.原型系统测试结果见表1.表中"/"前后分别对应检索结果的正确匹配数和检索结果集合中的元素数.所有测试的检索反馈时间均少于1秒.由表1可以看出,通过调整精度控制参数$r$,检索者可以实现查全率和查准率的折衷,获得直接在页面图像上完成内容检索的效果.

**Table 1** The experimental results for Chapter 3 in 'dream of the red chamber'

**表1** 《红楼梦》第3章页面图像检索实验结果

| Retrieved samples① | Actual occurrences② | Precision parameter③($s$=10) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 黛玉 | 77 | 4/4 | 4/4 | 22/22 | 22/22 | 35/35 | 50/50 | 63/63 | 77/77 |
| 宝玉 | 32 | 10/10 | 10/10 | 10/10 | 10/10 | 11/12 | 11/27 | 12/141 | 12/568 |
| 贾母 | 33 | 2/2 | 2/2 | 2/2 | 2/2 | 16/16 | 16/16 | 21/31 | 28/68 |

①检索样本,②库中出现次数,③精度控制参数.

## 4 结 论

本文提出了一种全新的基于视觉相似性的中文古籍内容检索方法.与传统的标引方法和附带文本文件方法相比,该方法不受检索点的限制,无须顾及同一性判定、字符集规模、通假字、词库等问题,自动化程度较高,操作简便.其另一显著特色是形态特征到语义的映射可由检索者方便地动态加以调整.原型系统实现了直接在数字化图像上完成古籍内容自动检索的技术效果,证实了以视觉相似性作为计算机古籍内容匹配的可行性,取得了较为令人满意的实验结果.

本方法不仅有助于数字图书馆古籍数字化应用,还适用于现代图书的数字化应用.通过定义新的特征描述,它也可以推广到外文书籍内容检索和商标检索等应用领域.

进一步的工作将开发更强的汉字特征提取手段,添加特征空间降维技术,引进相关反馈机制,以改善查全率/查准率综合性能,降低空间索引的复杂度.

**References:**

[1] Zhu, Yan. Experiences of electronic version of Si Ku Quan Shu complete library of the four branches of literature. The Journal of the Library Science in China, 1999,25(125):82~84 (in Chinese).

[2] Gladney, H., Mintzer, F., Schiattarella, F. Safeguarding digital library contents and users: digital images of treasured antiquities. D-Lib Magazine, 1997. http://www.dlib.org/dlib.html.

[3] Thibadeau, R., Benoit, F. Antique books. D-Lib Magazine, 1997. http://www.dlib.org/dlib.html.

[4] Zhang, Xin-zhong. Chinese Character Recognizing Techniques. Beijing: Tsinghua University Press, 1992 (in Chinese).

[5] Wang, W., Yang, J., Muntz, R. PK-Tree: a spatial index structure for high dimensional point data. In: Tanaka, K., Ghandeharizadeh, S., Kambayashi, Y., eds. Information Organization and Database. Boston: Kluwer Academic Publishers, 2000. 281~293.

**附中文参考文献:**

[1] 朱岩.《四库全书》电子版问世的启迪.中国图书馆学报,1999,25(125):82~84.

[4] 张炘中.汉字识别技术.北京:清华大学出版社,1992.

# Content-Based Chinese Antique Books Retrieval Through Visual Similarity Criteria*

SHI Bai-le, ZHANG Liang, WANG Yong, CHEN Zhi-feng

(*Department of Computer Science, Fudan University, Shanghai* 200433, *China*)

E-mail: {bshi,zhangl,wangyyong}@fudan.edu.cn

http://www.fudan.edu.cn

**Abstract:** The application of digitized civilization legacy plays an important role in the digital library project. Due to the intrinsic handwritten nature, it lacks effective mechanisms to perform content retrieval on digitized Chinese antique books. In this paper, an original method for content retrieval based on visual similarity is proposed and some key techniques are studied. By extracting morphological, positional and page features from images, the method makes up a feature space and applies spatial indexing to it. A range searching strategy is then employed to get all analogs to the query sample. In addition, a precision parameter is defined to dynamically adjust the mapping from morphological feature to semantics, and a constraint verifying technique is developed to improve the overall precision. The operational prototypical system demonstrates its feasibility and gets the effectiveness of automatic content-based retrieval directly on page images.

**Key words:** words content-based image retrieval; Chinese script; feature extraction; spatial indexing; antique-book retrieval

* Received November 24, 1999; accepted May 26, 2000

Supported by the National Natural Science Foundation of China under Grant No. 69933010; the Natural Sciene Fundation of Shanghai of China under Grant No. 00ZD14006