面向分类的网页主题特征提取 *

刘 建[1,2], 孙 鹏[2], 倪 宏[2]
(1. 中国科学院研究生院,北京 100049; 2. 中国科学院 声学研究所 国家网络新媒体工程技术研究中心,北京 100190)

摘 要: 提出一种基于页面空间特征、视觉特征和内容特征的主题相关性判别方法,通过主题相关度大小量化描述不同内容的重要性,并采用混合加权方法从主题相关节点中提取网页的主题特征。分类实验结果表明,相比传统的 FullDoc 全文分类,基于此方法提取的主题特征具有更好的分类效果。

关键词: 网页分类; 主题特征; 主题相关性
中图分类号: TP301 **文献标志码**: A **文章编号**: 1001-3695(2010)09-3399-04
doi:10.3969/j.issn.1001-3695.2010.09.053

Web-page topical feature extraction for Web-page classification

LIU Jian[1,2], SUN Peng[2], NI Hong[2]
(1. *Graduate University of Chinese Academy of Sciences, Beijing* 100049, *China*; 2. *National Network New Media Engineering Research Center, Institute of Acoustics, Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract: This paper presented a method that identifies the topical correlativity of one node based on the spatial features, visual features and content features of the page, quantitatively described the different degree of importance of the content, and extracted the topical features through the hybrid weighting method. Experimental results show that Web-page classification based on the extracted page features has better effect compared to the traditional FullDoc text classification.
Key words: Webpage classification; topical features; topical correlativity

网页分类按照网页主题来自动划分其所属类别,是组织和管理网页信息的有力手段,是 Web 个性化服务的重要基础。当前 Web 页面通常含有很多与网页主题无关的噪声信息,如广告栏、导航条和版权信息等,它们分布于网页四周,甚至附着在正文旁边,一定程度上影响了网页分类的效果。

本文提出的基于混合特征的网页主题提取方法的主要思想是:依据网页中不同信息所处的位置、占据的空间大小、视觉显示效果和内容的不同而具有的不同重要度,以网页中的容器类节点为最小单位,从空间、视觉和内容三个特征计算每个节点的主题相关度大小,并据此加权生成网页的主题特征向量用于网页分类。实验结果表明,基于此方法提取的网页主题内容保证了主题信息的完整性,与传统的基于文本的全文分类方法相比,分类效果有一定的提高。

1 相关研究

在 Web 信息提取领域已经有大量的研究工作, 主要有以下几个研究思路:a)基于传统的文本处理算法,其研究对象是整个网页的文本序列[1],没有考虑网页的特殊性;b)基于学习的方式, YI L 等人[2]为 Web 文档构造了一个 style 树,用于抽取风格类似网页的主要内容,这取决于文档特征的学习和积累;c)基于网页的内容特征进行分析,Wang 等人[3]将 DOM 转换为语义树,将 HTML 文档转换为含有语义信息的 STU-DOM 树,并对其进行基于结构的过滤和基于语义的剪枝,最后得到包含主题信息的子树;Pasternack 等人[4]采用最大子序列分割的办法分析网页文本,获取新闻类页面的主要内容;Kohlschütter[5]利用网页文本的稠密程度进行页面分割,将密度最大的作为主题,仅利用页面的内容特征对文本的主题相关性予以判别;d)还有一种效果比较好的思路是基于网页的分块结构,Song 等人[6]基于 VIPS[7]页面分割算法将网页分块,从内容特征和视觉特征两方面计算每个块的权重大小,选择权重最大的块作为页面的主题;Lin 等人[8]通过网页标签〈table〉分块,根据特征词在每个块中出现的概率计算每个块的熵值,选择熵值最大块作为主要内容;Debnath 等人[9]以网页中块结构为单位,提取符合某一个特征的所有块结构,然后进行 K-means 聚类,最后选择聚类效果最好的集合为目标,并将集合中所有分块的文本作为主要内容。这种方法对于结构规范的网页有很好的主题提取效果,但是忽略了网页中其他部分的主题信息,对结构不规范的网页效果不是很好。从这些相关研究中可以看出,页面的内容特征、空间特征和视觉特征对于网页主题信息的判定均有一定的指导作用。本文从这些特征出发,研究页面中影响节点主题相关性的因素,提出一种计算节点主题相关度大小的方法来量化描述页面中不同内容的重要程度,最后基于混合加权的方式提取网页主题特征。

2 主题提取算法

2.1 相关定义

定义 1 内容节点集合 $N_{\mathrm{DTree}} = \{d \mid d \in N_{\mathrm{tags}}, N_{\mathrm{tags}} \subseteq \mathrm{html\text{-}tags}\}$, HTML 文档被解析后被转换为文档对象模型(DOM)[10],N_{DTree}包含页面中的所有内容节点,DOM 树为一棵多叉树,节点类型主要有元素、属性、文本、文档、评论等。每个

收稿日期: 2010-01-30; **修回日期**: 2010-03-09 **基金项目**: 国家科技支撑计划课题(2008BAH28B04)
作者简介: 刘建(1982-)男,湖南人,博士,主要研究方向为网络通信、网络新媒体技术(liuj@dsp.ac.cn);孙鹏(1976-),男,副研究员,硕导,主要研究方向为网络新媒体技术、嵌入式系统等;倪宏(1964-),男,研究员,博导,主要研究方向为网络通信、网络新媒体技术等.

节点可以拥有任意多个孩子,其中根节点为文档对象,其他非叶子节点对应网页中的一个标签,叶子节点对应标签之间的文本。

定义 2 可视化节点集合 $N_{\mathrm{RTree}}=\{r\mid f(d)\rightarrow r,\mathrm{isVisible}(d)\wedge d\in N_{\mathrm{DTree}}\}$。$N_{\mathrm{RTree}}$是 N_{DTree}的一个映射,N_{RTree}中的每个节点 r 都与 N_{DTree}的一个可视化节点 d 对应。节点 d 经过渲染操作生成节点 r,因此 r 拥有 d 的排版属性,即显示区域、背景颜色、对齐方式、字体相关属性等。

定义 3 主题节点集合 $N_{\mathrm{MTree}}=\{m\mid m=r,\mathrm{isCoreNode}(r)\wedge r\in N_{\mathrm{RTree}}\}$。$N_{\mathrm{MTree}}$由 N_{RTree}的主题相关节点组成,$N_{\mathrm{MTree}}\subseteq N_{\mathrm{RTree}}$,定义节点 m 主题相关度为 V,表示该节点的内容与主题内容的相关程度和主题表达的重要程度。

2.2 算法流程

N_{DTree}中的节点包含了网页内容特征,而 N_{RTree}中节点包含了对应节点的视觉特征和空间特征。本文从 N_{RTree}出发,结合这三方面的特征,采用一定的判别规则和剪枝策略得到 N_{MTree},最后提取页面的主题内容和主题特征,大致流程如下:

a)读取 HTML 文档,通过词法分析和语法分析,生成文档对象节点集合树 N_{DTree}。

b)根据网页的样式表单(CSS),生成网页样式规则,对 N_{DTree}满足可视化条件的节点并进行排版和布局,生成 N_{RTree}。

c)根据主题节点的判别算法,遍历 N_{RTree}中的容器类节点进行主题相关性判别,删除主题不相关的节点,最后得到主题内容节点集合 N_{MTree}。

d)根据节点的标签等级策略和主题相关度大小,从 N_{MTree}中输出页面的主题内容,并通过混合加权的方式提取主题特征。

本文通过网页解析器完成前两步:生成 N_{DTree}和 N_{RTree},这两部分工作是主题节点判别和主题提取的前提,因为大部分浏览器内核都可以直接完成这两步,其具体细节不作描述,下面重点描述后两步。

2.3 主题判定规则

结合网页的内容特征、空间特征和视觉特征,根据以下规则判定节点主题相关性并确定主题内容。

规则 1 主题节点类型为容器类节点,主题内容由主题节点内部的可视化文本组成。

在 HTML 标记语言中,有一类标签用于网页布局,被称为容器类标签,如〈TABLE〉〈TD〉〈TR〉〈DIV〉〈P〉等。容器类标签将页面划分为不同的区域,页面设计者通常利用大量的容器类标签来设计网页,每个分块里面放置逻辑上意义相近的内容,以这类标签节点为最小的处理单位进行主题提取。

一般来说,网页设计者通过图片和文字来表达主题信息,鉴于图片包含的信息无法直接获取,目前只分析网页中可视化的正文文本以及图片和链接的描述文本信息,对网页之间的上下文链接等其他信息暂不考虑。

规则 2 主题内容位于页面区域的正文部分或与正文部分的交界处。

通过研究大量网页的布局,页面一般可以划分为正文区域和边界区域(页头、页尾、左导航、右导航)两大部分,如图 1 所示。因为对绝大部分网页而言,页面的宽度为固定宽度,页面的高度不是固定长度,所以设置各区域的范围:W_1 占页面宽度的 15%,W_2 占页面宽度的 20%,$H_1=150$ 像素,$H_2=200$ 像素。

结合上文页面划分结构,计算主题相关度大小并判定节点 n 是否主题相关:

a)如果节点 n 的显示区域完全位于正文部分,其主题相关度大小 $V=1$,该节点是主题相关。

b)如果节点 n 的显示区域完全位于边界区域,其主题相关度大小 $V=0$,该节点是主题不相关。

b)如果节点 n 的显示区域位于正文与区域的交界,需要根据正文区域重叠率(V_1)和文本相关度(V_2)两个参数进一步判定:

$$V_1=\frac{\text{与正文区域相交的面积}(S_i)}{\text{节点显示区域的面积}(S_r)} \tag{1}$$

$$V_2=\frac{\text{节点内部可视化文本的长度(text length)}}{\text{节点内部链接的数目(link count)}} \tag{2}$$

文本相关度的定义是基于这样一个事实:与主题无关的节点总是含有大量无关链接和极少描述文本。也就是说,文本相关度越大,表明该节点与主题越相关;反之,表明该节点与主题越不相关。这种方法对于移除链接从而保留主题文字,在 Gupat 等人[11]的方法中已经被证明是行之有效的。

根据经验规则,如果节点显示区域大部分位于正文部分,少部分位于边界区域,并且文本较多、链接数目较少,则可以认为该节点是主题相关。推出结论:V 与 V_1 成正比,与 V_2 也成正比,且 V_1 与 V_2 不相关。得到 V 的近似关系可以表示为 $V=\alpha V_1+\beta(1-\exp(-\lambda V_2))$,并且 $\alpha+\beta=1$。其中 λ 与每条链接的平均文本长度有关。

最后得到计算节点 n 主题相关度 V 大小的公式:

$$V=\begin{cases}1 & (n\text{ 在正文区域})\\ 0 & (n\text{ 在边界区域})\\ \alpha V_1+\beta(1-\exp(-\lambda V_2)),\alpha+\beta=1 & (n\text{ 在交界区域})\end{cases} \tag{3}$$

所以,只要定义合适的阈值,即可很大程度上区分节点是否是主题相关。在本文实验中,主题相关度的阈值取为 $V_c=0.5$,主题相关度大小在$[V_c,1]$范围内,判定为主题相关,反之主题不相关;另外为了简化实验,正文重叠率 V_1 与文本相关度 V_2 的权重系数取值 $\alpha=\beta=0.5$;λ 参数的大小与 V_c 相关,通过观察大量网页,得知有效链接平均文本长度大约 12 个字符,此时主题相关度贡献值 $1-\exp(-\lambda\times 12)\approx V_c$,也就是说 $V_2<12$ 时,$1-\exp(-\lambda V_2)<V_c$,$V_2>12$ 时,$1-\exp(-\lambda V_2)>V_c$ 并且很快逼近于 1,因此得到 λ 参数的大小,$\lambda=0.06$,这也说明了当文本较多、链接数目较少时,主题相关度越大。

规则 3 不同标签支持的内容具有不同的权重。

HTML 网页通过具体的元素来引入网页内容,通过不同元素和元素带有的不同属性来控制网页内容表现属性以及在版面中的呈现。不同的标签对网页布局和显示来说具有不同的作用,暗示出了不同文本在网页内容中的不同重要程度。在这些标签中出现的词,其表达文档内容的能力是有差别的。例如,〈TITLE〉标签表示其包括的文字是网页的标题;〈H1〉标签包括的文字是一级标题,显示在用户面前是大号加粗的字体和〈B〉对文本加粗显示等。显而易见,这样的文字对概括和强调网页的整体和局部内容起关键作用。按照在网页版面构成中的作用,对 HTML 标签元素进行分类,如表 1 所示。

表 1 HTML 标签权重等级(L_{level})划分

等级	标签举例	权重
1	〈TABLE〉〈TR〉〈TD〉〈P〉〈DIV〉〈BR〉〈HR〉〈A〉〈DL〉〈OL〉〈UL〉等	1
2	〈B〉〈I〉〈U〉〈EM〉〈FONT〉〈CENTER〉	2
3	〈H_2〉~〈H_6〉〈BIG〉〈STRONG〉	3
4	〈H_1〉〈META〉	4
5	〈TITLE〉	6

根据主题内容满足的规则,通过遍历 N_{RTree} 容器类节点的方式得到网页的主题内容 N_{MTree}。遍历过程从根节点出发,直到不能再划分的容器类节点为止,遍历采用宽度优先方式,复杂度最多为 $O(n)$。

2.4　主题特征提取和加权

在获取主题相关节点集合 N_{MTree} 中,每个节点中包含的所有叶子文本节点包含的就是主题内容文本,遍历抽取复杂度不超过 $O(n)$。

用二元集合序列 $\{s_k, c_k\}$ 来代表抽取的主题文本内容。其中定义集合 $\{s_k\}$ 为抽取的主题文本序列,s_k 表示每个叶子文本节点对应的文本。在抽取过程中,合并一个主题节点内的具有相同标签权重等级的不同叶子节点,以减少集合 $\{s_k\}$ 的大小;定义集合 $\{c_k\}$ 为集合 $\{s_k\}$ 中对应主题文本的权重大小序列,c_k 的大小描述了不同的主题文本对描述全文主题信息的贡献和重要程度,计算公式可以表示为

$$c_k = L(s_k) \times V(s_k) \tag{4}$$

其中:$L(s_k)$ 表示 s_k 对应的文档标签权重等级,$L(s_k) \in \{L_{\mathrm{level}}\}$;$V(s_k)$ 表示 s_k 对应的主题相关度大小,其范围在 $[V_c, 1]$。

网页主题特征描述采用向量空间模型:词条/权重序列 $\{t_i, w_i\}$。首先对集合 $\{s_k, c_k\}$ 中的文本序列 s_k 进行分词处理,定义词条 t_i 在文本 s_k 的频次是 tf_{ik};那么 t_i 的词频权重为混合加权的结果:$tf_i = \sum_k tf_{ik} \times c_k$,同时根据经验可知,中文词语长度和其语义表达能力之间存在联系,长度较长的词条相对较短的词条在语义上往往具有较强的表达区分能力,对每个词条 t_i 的权重 w_i 赋予一定的修正系数 μ,区分不同的词长:$\mu = 1 + \log(\mathrm{length}(t_i))$。最后基于传统的 TF-IDF 公式,$w_i$ 的归一化表达式为

$$w_i = \frac{\mu(\sum_k tf_{ik} \times c_k) \times \log(\frac{N}{df_i})}{\sqrt{\sum_{j=1}^{N} (\mu(\sum_k tf_{jk} \times c_k) \times \log(\frac{N}{df_j}))^2}} \tag{5}$$

其中:N 表示训练集页面总数,df_i 表示训练集中出现词条 t_i 的页面数目。

为了降低特征向量的维度,对词条的权重采取适当的剪枝策略,在 tf_i 归一化操作之前,根据阈值(本文实验中设定 tf_i 的阈值 $tf_{ic} = 1$)剪枝,不满足条件的词条(即 $tf_i < tf_{ic}$)不能作为最后的主题特征。

3　分类实验与结果分析

最简单的网页分类是全文 FullDoc 分类。该方法首先滤掉网页中的 HTML 标签,将网页转换为纯文本;其次通过滤除停用词和求词根(stemming)将网页表示成一个词袋(bag-of-words),每一个单词的权重用它们在网页出现的次数表示。在很多实验中,研究人员把这种方法作为底线,在本文的实验中,同样把这种方法作为底线实验。

为了检验所提出的基于网页主题特征的网页分类方法的有效性,本文作了一系列对比实验:采用传统的 FullDoc 全文和基于无加权主题特征(无加权 $c_k = 1$)两种方法进行网页分类,验证网页主题特征确实有助于网页分类;同时与基于加权网页主题特征的分类方法进行比较,验证加权特征有利于改进分类效果。实验中还研究了不同的参数阈值设置对于分类效果的影响,限于篇幅,未及详述。

3.1　实验工具及数据

支持向量机(SVM)在大规模数据处理中具有一定的优势,所以本文的实验选择了 SVM 分类器,分类工具采用 LibSVM[12],核函数采用径向基函数。

实验用的网页数据采用北京大学网络与分布式实验室提供的中文网页数据集[13](CCT2002-v1.1),它是 2002 年秋天北京大学网络与分布式实验室天网小组通过动员不同专业的几十个学生,人工选取形成了一个全新的基于层次模型的大规模中文网页样本集。它包括 11 678 个训练网页实例和 3 630 个测试网页实例,分布在 11 个大类别中。

文本分词软件采用中国科学院开源的中文分词工具 ICTCLAS[14],分词词典采用搜狗实验室[15]提供的互联网词库,本实验中只选择了对文本内容特征表示较好的名词、动词、形容词和常用词等共 77 448 条。

3.2　实验结果

本实验中特征选择的方法采用 χ^2 统计量(Chi-square, CHI),在训练集文档中选择 Top$N$ 的特征词条作为特征项。评价指标采用准确率、召回率和 F_1 值。原始网页的平均长度为 14 545 Byte, FullDoc 方法抽取的纯文本平均长度为 6 124 Byte, 抽取的主题文本为 3 256 Byte,训练集的特征项一共 39 677 词条。

实验结果如下面的图表所示。图 2 是三种分类方法在不同的特征维度条件下的分类效果 macro-F_1 值比较;表 2 和 3 分别描述了在特征词维度 $D = 1\,000$ 和 $D = 5\,000$ 条件下,三个评价指标的大小。

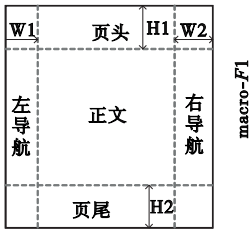

图1　页面区域结构划分

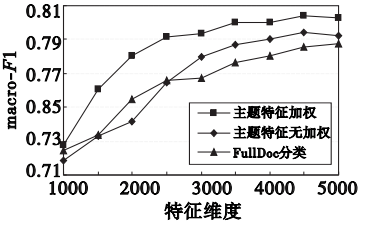

图2　分类结果macro-$F1$值比较

表 2　分类评价指标实验结果举例($D = 1\,000$)

分类	macro$P$	macro$R$	macroF_1
FullDoc 分类	0.743	0.707	0.725
主题特征无加权	0.729	0.709	0.719
主题特征加权	0.739	0.717	0.728

表 3　分类评价指标实验结果举例($D = 5\,000$)

分类	macro$P$	macro$R$	macroF_1
FullDoc 分类	0.806	0.770	0.788
主题特征无加权	0.802	0.782	0.792
主题特征加权	0.810	0.795	0.803

从上面的图表中可以看出:

a)三种分类方法的 macro-F_1 值特征值向量的维度 D 在 3 000 以后结果趋于稳定,说明基于主题特征的分类方法是相对稳定的。

b)基于无加权主题特征的分类结果与全文分类结果比较接近,各项指标的值都非常接近,而且随着特征维度的增加,差别越来越小。可以证明,本方法抽取的主题文本大小虽然只有全文的一半左右,但是信息量没有减少,基本上包括了文本的主要信息,不会降低分类效果。

c)基于加权主题特征的分类结果最好,这充分说明基于多重特征的混合加权方式抽取的文本特征向量可以较好地表达文本的主题信息特征,能够提供比全文更好的分类效果。

d)加权的主题特征对于分类效果有一定的改进,但是幅度不是特别大。这是因为在主题文本以外还存在接近一半的文本信息影响着分类的效果,虽然是主题无关信息,其文本特征在分类过程中也能起到一定的作用。

4　结束语

本文的主题特征提取是把每一个网页作为独立的文档来处理,然后 Web 网页之间存在着丰富的关联关系,如果充分利用网页上下文信息,能够进一步提高主题特征的准确性,进而提高分类的效果。本文提出的基于多重特征的主题相关性判别方法涉及到多种影响因子。实际应用中,参数的选择需要经过大量的实验确定最优值。另外,网页主题特征可以应用于 Web 个性化服务系统中,在用户历史网页分类的基础上,结合用户浏览行为去分析用户的网页兴趣,具有很好的应用价值。

参考文献:

[1] KO Y, PARK J, SEO J. Automatic text categorization using the importance of sentences[C]//Proc of the 19th International Conference on Computational Linguistics. Morristown, NJ: Association for Computational Linguistics, 2002:1-7.

[2] YI Lan, LIU Bing, LI Xiao-li. Eliminating noisy information in Web pages for data mining[C]//Proc of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2003:296-305.

[3] WANG Qi, TANG Shi-wei, YANG Dong-qing, *et al*. DOM based automatic extraction of topical information from Web pages[J]. Journal of Computer Research and Development, 2004, 41(10):1786-1792.

[4] PASTERNACK J, ROTH D. Extracting article text from the Web with maximum subsequence segmentation[C]//Proc of the 18th International Conference on World Wide Web. New York: ACM Press, 2009: 971-980.

[5] KOHLSCHÜTTER C, NEJDL W. A densitometric approach to Web page segmentation[C]//Proc of the 17th ACM Conference on Information and Knowledge Management. New York: ACM Press, 2008: 1173-1182.

[6] SONG R, LIU H, WEN J, *et al*. Learning block importance models for Web pages [C]// Proc of the 13th International Conference on World Wide Web. New York: ACM Press, 2004:203-211.

[7] CAI Deng, YU Shi-peng, WEN Ji-yong, *et al*. VIPS: a vision-based page segmentation algorithm[R]. Microsoft Technical Report, MSR-TR-2003-79, 2003.

[8] LIN Shian-hua, HO J. Discovering informative content blocks from Web documents[C]//Proc of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2002:588-593.

[9] DEBNATH S, MITRA P, GILES C L. Identifying content blocks from Web documents[M]//Foundations of Intelligent Systems. Berlin: Springer, 2005:285-293.

[10] World Wide Web Consortium (W3C), DOM Specification [EB/OL]. http://www.w3.org/DOM/.

[11] GUPTA S, KAISER G, NEISTADT D, *et al*. DOM-based content extraction of HTML documents[C]//Proc of the 12th International Conference on World Wide Web. New York: ACM Press, 2003:207-214.

[12] CHANG C, LIN C. LIBSVM: a library for support vector machines [EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[13] 中文 Web 信息检索论坛(CWIRF)[EB/OL]. http://www.cwirf.org/.

[14] 中国科学院中文分词工具(ICTCLAS)[EB/OL]. http://ictclas.org/.

[15] 搜狗实验室(SogouLabs)[EB/OL]. http://www.sogou.com/labs/.

(上接第 3398 页)

图6　上海市三维WebGIS

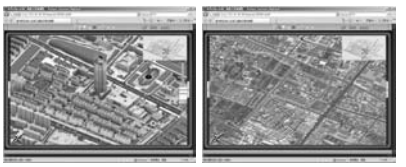

图7　上海市奉贤区部分WebGIS

参考文献:

[1] 龚健雅,杜道生,李清泉,等. 当代地理信息技术[M]. 北京:科学出版社,2004:87-88.

[2] 方裕,周成虎,景贵飞,等. 第四代 GIS 软件研究[J]. 中国图象图形学报,2001,6A(9):817-823.

[3] 邓红艳,武芳,翟仁健,等. 一种用于空间数据多尺度表达的 R 树索引结构[J]. 计算机学报,2009,32(1):177-184.

[4] CHRISTELLE V. Multi-representation inspatial database using the MADS conceptual model[C]//Proc of International Cartographic Association Workshopon Generalization and Multi-Scale Representation. Leicester:[s.n.],2004:337-342.

[5] MARKDAVID M, CHRISTIANETAL F. Cognitive models of geographical space[J]. Geographical Information Science, 1999, 13(8):747-774.

[6] PENG Hu, QI Qing-wen, LIU Zhao-li. Progress in studies on automated generalization of spatial point cluster[J]. IEEE International Geoscience and Remote Sensing Symposium, 2004, 13(8): 2841-2844.

[7] 李爱勤. 无缝空间数据组织及其多比例尺表达与处理研究[D]. 武汉:武汉大学,2001.

[8] GUTTMAN A. R-tree: a dynamic index structure for spatial search [C]//Proc of ACM SIGMOD International Conference on Management of Data. Boston:[s.n.],1984:47-57.

[9] SAYAR A, PIERCE M. FOX G. Integrating Ajax approach into GIS visualization Web services[J]. IEEE Computer Society, 2006(2): 169-170.

[10] OGC. OpenGIS Simple Features Specification 06-103r3_Candidate_Implementation_Specification_for_Geographic_Information_-_Simple_feature_access_-_Part_1_Common_Architecture_v1.2.0[S].

[11] 董鹏,杨崇俊,芮小平,等. 一种基于改进四叉树的 GIS 空间选择查询算法——以 ESRI SHAPE 格式文件为例[J]. 计算机工程与应用,2003,39(13):58-61.

[12] 宋关福,钟耳顺. 组件式地理信息系统研究与开发[J]. 中国图象图形学报:A 辑,1998,3(4):313-317.

[13] BANNAI N, FISHER R B, AGATHOS A. Multiple color texture map fusion for 3D models[J]. Pattern Recognition Letters, 2007(28): 748-758.

[14] 朱庆,高玉荣,危拥军,等. GIS 中三维模型的设计[J]. 武汉大学学报:信息科学版,2003,28(3):283-287.

[15] 李清泉,李德仁. 三维空间数据模型集成的概念框架研究[J]. 测绘学报,1998,27(4):325-330.

[16] 虞强源,刘大有,谢琦. 空间区域拓扑关系分析方法综述[J]. 软件学报,2003,14(4):777-782.

[17] BRUN L, KROPATSCH W. Introduction to combinatorial pyramids. [C]//Proc of Digital and image geometry LNCS vol2243. Berlin: Springer, 2001:108-127.