第35卷 第18期 计 算 机 工 程 2009年9月
Vol.35 No.18 Computer Engineering September 2009
·博士论文· 文章编号：1000—3428(2009)18—0015—04 文献标识码：A 中图分类号：TP391

基于最大熵模型的汉语词义消歧与标注方法

张仰森[1,2]
(1. 北京信息科技大学智能信息处理研究所，北京 100192；2. 中国科学院自动化所模式识别国家重点实验室，北京 100080)

摘 要： 分析最大熵模型开源代码的原理和各参数的意义，采用频次和平均互信息相结合特征筛选和过滤方法，用Delphi语言编程实现汉语词义消歧的最大熵模型，运用GIS(Generalized Iterative Scaling)算法计算模型的参数。结合一些语言知识规则解决训练语料的数据稀疏问题，所实现的汉语词义消歧与标注系统，对800多个多义词进行词义标注，取得了较好的标注正确率。

关键词： 词义消歧与标注；最大熵模型；上下文特征；特征筛选

Approach to Chinese Word Sense Disambiguation and Tagging Based on Maximum Entropy Models

ZHANG Yang-sen[1,2]
(1. Institute of Intelligent Information Processing, Beijing Information Science & Technology University, Beijing 100192;
2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academic of Sciences, Beijing 100080)

【Abstract】 This paper analyzes the principle and every parameter meaning of open-source code of maximum entropy models, uses the method of the combination of feature frequency and average mutual information to select the features from the candidate feature set, realizes the maximum entropy models for Chinese Word Sense Disambiguation(WSD) by Delphi, and computes models parameters by GIS algorithm. It solves the data sparseness problem by combining the linguistic knowledge. The system for Chinese word sense automatic disambiguation and tagging is implemented. It uses the system to tag word sense of more than 800 multivocal words, and achieves the better correcte rate.

【Key words】 Word Sense Disambiguation(WSD) and tagging; maximum entropy models; contextual features; feature selecting

1 词义消歧问题的最大熵模型描述

词义消歧(Word Sense Disambiguation, WSD)问题的最大熵模型根据样本信息进行概率估计可分为2种：联合最大熵模型和条件最大熵模型。假设 a 是某个事件，b 是事件 a 发生的环境(或称上下文)，则 a 和 b 的联合概率记为 $p(a, b)$。一般地，设所有可能发生的事件组成的集合为 A，所有环境组成的集合是 B，则对任意给定的 $a\in A, b\in B$，求概率 $p(a, b)$ 须建立联合最大熵模型。若要计算在 b 的条件下，事件 a 发生的概率，即概率 $p(a \mid b)$，则须建立条件最大熵模型。

对于汉语词义消歧问题，若将 A 看作当前多义词的所有可能义项的有限集合，B 为其上下文信息组成的集合，则可将确定某个多义词(称为中心词)的某个义项 $a\in A$ 看成一个事件，中心词周围出现的词及其词性看成这个事件发生的环境 $b\in B$。建立语言模型的目的就是计算输出义项 a 的条件概率 $p(a \mid b)$，即条件最大熵模型。选择集合 A 中条件概率值 $p(a|b)$ 最大的义项作为多义词的意义[1]。

定义 1 设 $b\in B$，是决定当前多义词义项 $a(a\in A)$的上下文信息，若 b 对 a 具有表征作用，则称(a, b)为模型的一个特征。定义一个$\{0, 1\}$域上的二值函数来表示特征：

$$f(a, b)=\begin{cases}1 & \text{若}(a, b)\in(A, B),\text{且满足某种条件}\\ 0 & \text{其他}\end{cases} \tag{1}$$

词义消歧建模的目标是构造能够对实际文本进行准确描述的统计模型，它的概率分布与训练语料中的经验概率分布应该相符。而训练语料中上下文信息与输出的经验概率分布 $\tilde{p}(a, b)$ 可由下式计算：

$$\tilde{P}(a, b)\approx\frac{C(a, b)}{\sum\limits_{A, B}C(a, b)} \tag{2}$$

其中，$C(a, b)$为(a, b)在训练语料中出现的次数。

如果有特征 f_j，它在训练样本中关于经验概率分布 $\tilde{p}(a, b)$ 的数学期望为

$$E_{\tilde{p}}(f_j)=\sum_{A,B}\tilde{p}(a, b)f(a, b) \tag{3}$$

而特征 f_j 关于所建模型 $p(a, b)$的数学期望为

$$E_p(f_j)=\sum_{A,B}p(a, b)f(a, b) \tag{4}$$

因为 $p(a, b)=p(a)p(b \mid a)$，且所建模型应符合训练语料中的概率分布，所以若 $\tilde{p}(a)$ 表示 a 在训练样本中的经验分布，令 $p(a)=\tilde{p}(a)$，则式(4)变为

$$E_p(f_j)=\sum_{A,B}\tilde{p}(a)p(b \mid a)f(a, b) \tag{5}$$

若特征 f_j 对模型有用，则应要求式(5)所表示的特征 f_j 之数学期望与它在训练样本中的相同，即

基金项目： 国家自然科学基金资助项目(60873013)；北京市自然科学基金B类资助重点项目(KZ200811232019)；中科院自动化所模式识别国家重点实验室开放专项经费基金资助项目；北京市属市管高校人才强教计划基金资助项目(PXM2008_014215_055942)

作者简介： 张仰森(1962－)，男，教授、博士，主研方向：中文信息处理，人工智能

收稿日期： 2009-01-05 **E-mail：** zys@bistu.edu.cn

定义 2 建立语言模型应满足条件：

$$E_p(f_j)=E_{\tilde{p}}(f_j) \tag{6}$$

称该条件为语言建模的约束条件。

2 基于最大熵的模型遴选及参数计算

2.1 基于最大熵的模型遴选

设存在 n 个特征 $f_i(i=1, 2,\cdots, n)$，是建模过程中对输出有影响的统计单元，所建立的模型 p 应属于这 n 个特征约束所产生的模型集合 C：

$$C=\{p\in\Gamma \mid E_p(f_i)=E_{\tilde{p}}(f_i)\ \ i\in\{1,\ 2,\ \cdots,\ n\}\} \tag{7}$$

其中，Γ 为所有的(无条件或无约束)概率分布模型空间；C 为加入特征约束条件后得到的 Γ 的一个子集。满足约束条件的模型集 C 中有许多模型，所要求取的是分布最均匀的模型，而条件概率 $p(y \mid x)$均匀性的一种数学测量方法为条件熵，定义为

$$H(p)=-\sum_{A,\ B}\tilde{p}(\mathrm{a})p(a\mid b)\mathrm{lb}p(a\mid b) \tag{8}$$

其中，$0\leqslant H(p)\leqslant \mathrm{lb}|b|$。

定义 3 在满足 n 个约束条件的前提下，具有使 $H(p)$值最大的模型即为具有最均匀分布的模型。即

$$p^*=\arg\max_{p\in C}H(p) \tag{9}$$

可以证明[2]，满足式(9)的解具有如下 Gibbs 形式：

$$p^*(a\mid b)=\frac{1}{Z(b)}\exp(\sum_{j=1}^{1}\lambda_j\cdot f_j(a,\ b)) \tag{10}$$

其中，$Z(b)=\sum_{\mathrm{A}}\exp(\sum_{j=1}^{1}\lambda_j\cdot f_j(a,\ b))$ (11)

为归一化因子。

在式(10)和式(11)中，λ_j 为特征 f_j 的权重参数。当 j=l 时，$f_{\mathrm{l}}(a, b)$为修正特征函数，λ_{l} 和 $f_{\mathrm{l}}(a, b)$对应的参数在后面的模型参数计算部分说明。可以通过在训练集上进行学习，得出 λ_j 的值，即可得到 $P^*(a|b)$，完成条件最大熵模型的构造。

2.2 模型参数计算及说明

本文采用 GIS 算法计算最大熵模型参数值 λ_j，GIS 算法要求对训练集中的每个实例中的任何$(a,\ b)\in A\times B$，特征函数之和为常数，即对每个实例均满足

$$\sum_{j=1}^{k}f_j(a,\ b)=C\ (C\ \text{为常数}) \tag{12}$$

如果上述条件不能满足，则根据训练集选择 C，C 为在训练集所有实例中根据式(12)等号左边算得的最大值。还须增加一个修正特征(correction feature)f_l，其中，l=k+1。

$$f_l(a,\ b)=C-\sum_{j=1}^{k}f_j(a,\ b) \tag{13}$$

GIS 算法详细描述请参见文献[3]。本文参阅 OpenNLP MaxEnt 提供的 Java 程序[4]，用 Delphi 语言实现了 GIS 算法并做了适当的改进。用 OpenNLP MaxEnt 中的一个例子说明生成的模型参数文件中各项的意义。训练语料样本如下：

Sunny Happy Outdoor
Sunny Happy Dry Outdoor
Sunny Happy Humid Outdoor
Sunny Sad Dry Outdoor
Sunny Sad Humid Outdoor
Cloudy Happy Humid Outdoor
Cloudy Happy Humid Outdoor
Cloudy Sad Humid Outdoor
Cloudy Sad Humid Outdoor
Rainy Happy Humid Indoor
Rainy Happy Dry Indoor
Rainy Sad Dry Indoor
Rainy Sad Humid Indoor
Cloudy Sad Humid Indoor
Cloudy Sad Humid Indoor

用该训练语料可学习一项运动是在 Indoor 或在 Outdoor 举行的知识。能获得的特征信息与天气和情绪有关(如 Sunny、Happy)。在 OpenNLP MaxEnt 提供的 Java 程序中，以下几个参数可以选择：

(1)*cutoff*：确定使用特征的最小数量。当 *cutoff*=0 时，表示不使用此参数，将出现的所有上下文都选为特征。

(2)*USE_SMOOTHING*：是否使用模型平滑处理技术。

(3)*_useSlack Parameter*：是否使用修正特征函数。

就上述语料，当 *cutoff*=0, *USE_SMOOTHING*=false 时，可构造出 12 个特征函数。例如：由于在训练语料样本文件中有(*Sunny*, *Outdoor*), (*Happy*, *Indoor*)出现，所以有下列的特征函数 $f_1(a, b), f_2(a, b)$。

$$f_1(a,\ b)=\begin{cases}1 & \text{当}a=Outdoor, b=Sunny\text{时}\\ 0 & \text{其他}\end{cases}$$

$$f_2(a,\ b)=\begin{cases}1 & \text{当}a=Indoor, b=Happy\text{时}\\ 0 & \text{其他}\end{cases}$$

由于在训练语料样本文件中没有(*Sunny*, *Indoor*)的出现，故不会有以下特征函数

$$f(a,\ b)=\begin{cases}1 & \text{当}a=Indoor,\ b=Sunny\text{时}\\ 0 & \text{其他}\end{cases}$$

根据这样的原则，共可构造出 12 个特征函数 f_1~f_{12}。修正特征函数如下：

$$f_{13}(a,\ b)=C-\sum_{j=1}^{12}f_j(a,\ b) \tag{14}$$

当不使用修正特征函数时，运用 GIS 算法求得的模型参数文件的内容及意义如下：

(1) GIS　　//表示所用的算法为 GIS 算法
(2) 3　　//为常数 C
(3) 0.0　　//为修正特征函数对应的参数
(4) 2　　//为输出结果数目，在义项标注中为多义词的标注义//项的数量
(5) Outdoor　　//为输出结果，在义项标注中为多义词对应的义//项标注符号，对“斗争 vn”的标注记号为“!1$”、“!2$”
(6) Indoor　　//与(5)意义相同
(7) 3　　//按预测条件对应的结果情况所做的分类数
(8) 1　0　　//表示输出结果为 0(表示结果为 Outdoor)的上下文//特征数为 1
(9) 5　0　1　　//表示输出为 0(Outdoor)和 1(Indoor)2 个结果的//上下文特征数为 5
(10)　1　1　　//表示输出结果为 1(表示结果为 Indoor)的上下文//特征数为 1
(11)7　　//用于预测输出结果的特征或条件数量；
//(12)~(18)为上述语料中的 7 个显性特征或条件
(12)Sunny
(13)Happy
(14)Dry
(15)Humid
(16)Sad
(17)Cloudy
(18)Rainy
//(19)~(30)为上述特征函数 f_1~f_{12} 所对应的特征参数
(19)9.941 938 146 233 399
(20)1.641 830 606 189 509 6
(21)-3.588 776 157 349 490 5
(22)-0.172 110 742 671 619 4

(23)0.181 611 872 340 052 83
(24)-0.059 378 170 227 835 98
(25)0.143 460 399 621 254 9
(26)-0.936 063 747 483 156 3
(27)0.963 653 758 189 680 2
(28)0.914 501 976 053 200 5
(29)-1.614 369 689 370 09
(30)12.573 539 229 046 837

3 汉语训练语料中词义消歧知识及其提取

3.1 汉语训练语料中的词义消歧知识

汉语训练语料文本中的显性特征信息包括词形信息、词性信息、词形+词性信息[1]。

定义 4 词形信息特征就是多义词所在的上下文中指定窗口内的所有词或部分有限定的词(如只要求实词)$Wi(-m \leqslant i \leqslant -1;\ 1 \leqslant i \leqslant n)$当作词义消歧的特征。

定义 5 词性信息特征是指多义词所在上下文中各个词所标注的词性信息，即 $P\text{-}m,\cdots, P\text{-}1, P0, P1,\cdots, Pn$，$P\text{-}k$ 表示多义词向左数第 k 个词 $W\text{-}k$ 的词性，Pk 表示多义词向右边数第 k 个词 Wk 的词性。

定义 6 词+词性信息特征是将目标多义词上下文中的词 Wi 及其词性 $Pi(-m \leqslant i \leqslant -1;1 \leqslant i \leqslant n)$同时作为词义消歧的特征。这 3 种特征信息又可根据是否考虑词在上下文中的位置分为词袋型显性特征和位置型显性特征 2 种。所谓词袋型显性特征是指目标词周围在一定窗口范围内的词、词性或词+词性组成的集合；而位置型显性特征则是在从上下文中提取词、词性或词+词性等特征时考虑了其与目标词的距离因素。

例如，在“叶利钦/nr　发表/vt4!1$　电视/na　讲话/na　后/ft，/wd”中，“发表”为多义词，若把其上下文中的所有词作为特征信息，则显性的词形特征向量为<*W*-1=叶利钦, *W*1=电视,*W*2=讲话,*W*3=后>，显性词性特征向量为<*W*-1=nr, *W*1=na,*W*2=na,*W*3=ft>,词+词性的特征向量为<*W*-1=叶利钦/nr, *W*1=电视/*na*, *W*2=讲话/na, *W*3=后/ft>。

显性特征信息的提取可以采用词袋方法和特征模板的方法：词袋的方法就是提取目标多义词周围在一定窗口范围内的词，组成一个特征信息集合，而特征模板的方法则是将距离和位置信息以及特定位置上的语法属性信息都考虑成影响多义词词义的因素，按设计的特征模板从上下文中提取特征，组成特征信息集合。

特征模板一般包括词语的位置参数和语法属性信息参数。本文特征模板考虑 3 个方面的选择特征：(1)特征类型，包括词形、词性、词形+词性；(2)窗口大小，包括语句内当前词左右 1 个词、2 个词、3 个词、当前词所在的整个句子(除当前词)；(3)是否考虑位置特征，包括是或否 2 个特征。若不考虑位置特征，则就退化为词袋方法。

考虑上述 3 个方面的特征，通过组合搭配，可得到的特征选择模板共有 3×4×2=24 种组合。在进行词义消歧时，选择一种模板进行获取上下文特征，并以这些特征进行模型参数的计算，在利用所构建的模型进行词义标注时，所选用的模板要和机器学习时的一致。

3.2 特征的提取与选择

利用特征模板所得到的候选特征集合中通常包含许多特征，从中选择对输出影响较大的特征时，常用的方法一般有 3 种：

(1)从候选特征集中选择那些在训练语料中出现一定频次的特征。

(2)利用互信息作为评价工具从候选特征集中选择满足一定互信息要求的特征。

(3)利用 Della Pietra 等人提出的增量式特征选择法从候选特征集中选择特征。

第 3 种方法计算比较复杂，本文采用将频次与平均互信息相结合相结合特征选择方法[5]，取得了较好的效果。

4 词义标注的编程实现

在应用最大熵算法对语料进行词义标注时，采用如下策略：

(1)若待标注的词为单义词，则通过查语义词典，直接为其标上语义代码。

(2)若待标注的词为多义词，且各义项的使用频度差别很大，在训练语料中表现为一个“词+词性”只有一种义项，在对测试语料进行标注时，可直接标注该义项。

(3)若待标注的词为多义词，且训练语料中的一个“词+词性”有 2 个及 2 个以上义项时，在对测试语料进行标注时，采用最大熵模型对该词的义项进行标注。

(4)若待标注的词为多义词，且在训练语料中不出现或出现频次很低，采用语法信息词典和语义词典中相关知识编制规则，由规则实现词义标注。

词义消歧与标注程序采用 Delphi 7.0 编写，包括机器学习、词义标注、结果评测几个模块。多义词词义消歧模型为最大熵模型，词义消歧与标注软件的实现流程如图 1 所示。

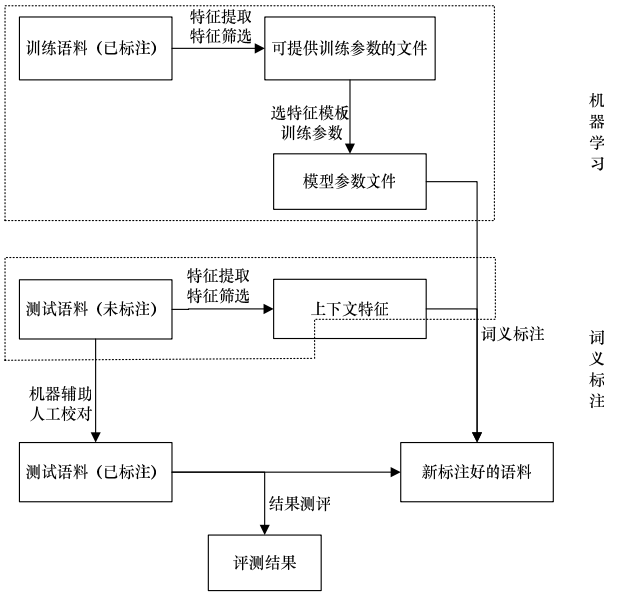


图 1　最大熵词义标注软件实现流程

图 1 中机器学习部分是生成模型参数文件，包括：(1)特征提取：根据选定的特征模板进行特征提取，生成用于训练参数的文件。(2)训练参数：根据选定的特征模板进行参数训练，生成参数 λ_i 的值，存放在文件中。词义标注部分根据选定的特征模板，读取参数 λ_i 的值，对特定的上下文 b，计算属于各类 a(即义项)的概率 $p(a|b)$，选择概率最大的类，用相应的标记进行词义标注。

5 实验结果及分析

5.1 封闭测试

对 2000 年 1 月的《人民日报》词义标注语料去除义项标注后，进行义项标注的测试，对几种特征模板及特征筛选方法的最大熵模型算法的实验结果如表 1 所示。

表 1　词义标注封闭测试评测表

序号	所用模型	特征筛选方法	特征模板			标注结果		
			特征类型	窗口大小	是否考虑位置	不正确数	总标注数	正确率
1	最大熵	频率与互信息结合	词形	整句	否	275	61 539	0.995 5
2	最大熵	频率与互信息结合	词形	3	否	5 477	61 539	0.911 0
3	最大熵	频率与互信息结合	词形	2	否	5 620	61 539	0.908 7
4	最大熵	筛选频率≥2	词形	整句	否	1 740	61 539	0.971 7
5	最大熵	筛选频率≥2	词形	3	否	8 161	61 539	0.867 4
6	最大熵	筛选互信息>0.6	词形	3	否	7 844	61 539	0.872 5

5.2　相对开放测试

从 2000 年 1 月的《人民日报》词义标注语料选择 28 天(2000 年 1 月 1 日—2000 年 1 月 28 日)语料作为训练语料，剩余 3 天(2000 年 1 月 29 日—2000 年 1 月 31 日)作为测试语料，在去除测试语料中相关多义词的义项标注后，利用所建立的最大熵模型算法对其进行义项标注实验。

特征模板的选择为：特征类型=词形、窗口大小=全句、不考虑位置特征，标注的多义词个数为 4 913，不正确的标注数为 278 个，正确率为 94.34%。采用其他模板和特征筛选方法的实验结果如表 2 所示。

表 2　词义标注相对开放测试评测表

序号	所用模型	特征筛选及方法	特征模板			标注结果	
			特征类型	窗口大小	是否考虑位置	不正确数	正确率
1	最大熵	频率与互信息结合	词形	整句	否	278	0.943 4
2	最大熵	频率与互信息结合	词形	3	否	508	0.896 6
3	最大熵	频率与互信息结合	词形	2	否	515	0.895 2
4	最大熵	筛选频率≥2	词形	整句	否	296	0.939 8
5	最大熵	筛选频率≥2	词形	3	否	700	0.857 5

5.3　实验结果分析

封闭测试使用的训练语料和测试语料相同，即先使用训练语料求取模型参数后，对训练语料中的义项标注删除，再利用语言模型对其进行机器标注，将标注结果和原来的语料进行比较。可见，封闭实验 1~实验 3 除窗口大小不同外，实验条件均是相同的。特征选择窗口取整句的效果最好，这是因为有些决定多义词义项的关键特征位于语句前端或末端，对于较长的句子来说，距离较远，若窗口选择的小，则会丢失有效特征。频率与互信息结合的特征选择策略较之单独使用频率或互信息的正确性要高，主要是由于获得了更多的特征，约束限制更加准确。窗口长度约达，使用的特征多，则预测的词语义项正确率就高，与实际相符。

相对开放测试所用的训练语料和测试语料不同，但都是人民日报同 1 个月的内容。由于训练语料与测试语料不是相同的文本，因此实验结果的正确率稍低一些，与实际相符。

6　结束语

本文基于最大熵原理的语言建模方法构建一个符合多信息特征约束的语言模型，实现对大规模人民日报基本标注语料库的消歧与标注。所实现的系统由于结合了语法信息词典和语义词典的知识规则，对统计语言建模的数据稀疏问题做了较好的处理，可实现对语法信息词典中 800 多个多义词的词义消歧与标注，对综合语言知识库建设具有重要的作用。

参考文献

[1] 张仰森. 面向语言资源建设的汉语词义消歧与标注方法研究[D]. 北京: 北京大学, 2006.

[2] Adwait R. A Simple Introduction to Maximum Entropy Models for Natural Language Processing[R]. Philadelphia, PA, USA: University of Pennsylvania, Tech. Rep.: IRCS-97-08, 1997.

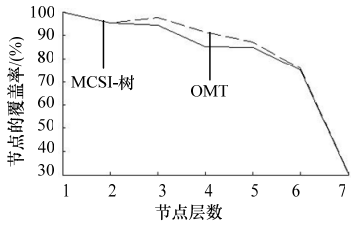
[3] Rosenfeld R. A Maximum Entropy to Adaptive Statistical Language Learning[J]. Computer Speech and Language, 1996, 10(3): 187-228.

[4] The OpenNLP Maximum Entropy Package[Z]. (2000-05-22). http://maxent.sourceforge.net.

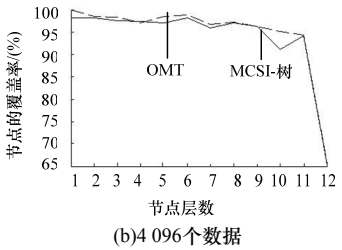
[5] 张仰森, 曹元大, 俞士汶. 最大熵方法中特征选择算法的改进与纠错排歧[J]. 北京理工大学学报: 自然科学版, 2006, 26(1): 36-40.

编辑　金胡考

(上接第 14 页)



(a)44个数据



(b)4 096个数据

图 2　树上各层节点覆盖率比较

5　结束语

R-树是目前最流行的空间数据索引方法之一，是一种完全动态的空间索引结构，其插入、删除及查询等操作可同时进行。它的致命弱点之一就是同一层节点覆盖大小难以有效缩小。因此，对于精确匹配查询，它会使查询的范围增大，从而增加查询的时间，降低查询的效率。本文针对这个问题，以降低兄弟节点间相互覆盖为目标，运用空间对象分割技术、结合二叉树 R-树思想，提出了一种空间数据索引结构——MCSI-树。实验表明，MCSI-树中 2 层节点之间的覆盖减少，在一定程度上达到了提高数据查询效率的目的。

参考文献

[1] Guttmann A. R-trees: A Dynamic Index Structure for Spatial Searching[C]//Proc. of ACM SIGMOD International Conference on Management of Data. New York, USA: ACM Press, 1984.

[2] Hung Powhei, Lin Hungyi. Optimizing Storage Utilization in R-tree Dynamic Index Structure for Spatial Databases[J]. Journal of Systems and Software, 2001, 55(3): 291-299.

[3] Lee Taewon, Sukho A. OMT: Overlap Minimizing Top-down Bulk Loading Algorithm for R-tree[J]. Advanced Information Systems Engineering, 2003, 11(7): 69-72.

编辑　张正兴