面向排序学习的锦标赛排序特征选择方法

蒋宗礼,李涵昱
(北京工业大学 计算机学院,北京 100124)

摘 要:排序问题在信息检索领域是一个非常重要的课题。虽然排序学习模型的算法早已被深入研究,但针对排序学习算法中的特征选择的研究却很少。现实的情况是,许多用于分类的特征选择方法被直接应用到排序学习中。但由于排序和分类有着显著的差异,应研究出针对排序的特征选择算法。文中在介绍常用的排序学习的特征选择方法的基础上,提出了一种全新的、适用于 QA 问题的排序学习的特征选择方法-锦标赛排序特征选择方法。实验结果显示,这种新的特征选择方法在提高特征提取效率和降低特征向量维数方面都有显著改善。

关键词:特征选择;排序学习;信息检索
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2014)02-0050-05
doi:10.3969/j.issn.1673-629X.2014.02.012

Championships Sort Feature Selection Method of Oriented Learning to Rank

JIANG Zong-li,LI Han-yu
(College of Computer,Beijing University of Technology,Beijing 100124,China)

Abstract:Ranking is a very important topic in information retrieval. And algorithms for learning to ranking models have been intensively studied,this is not the case for feature selection,despite of its importance. The reality is that many feature selection methods used in classification are directly applied to ranking. Argue that because of the striking differences between ranking and classification,it is better to develop different feature selection methods for ranking. To this end,a new feature selection method in this paper is proposed. The feature selection based on common feature selection methods,a novel method adapting better to learning to rank of QA problems,the championships sort feature selection method is proposed in this paper. The experimental results show that,this method can improve the efficiency of feature selection and reduce the dimensions of the feature vector.

Key words:feature selection;learning to rank;information retrieval

0 引 言

最近几年,大量的 QA(Question Answer)[1] 问题频繁出现在各类网站上,比如 QA 社区或论坛。把问题答案对(QA,thread)作为知识资源的一个重要挑战是如何按质量的高低自动地对一个问题的若干个答案进行排序。因为一个(QA,thread)中答案的质量良莠不齐,而且几乎所有的 QA 社区和论坛对回答的答案不做任何处理。显然,这会对用户的体验造成一些负面影响。

目前,利用监督学习[2] 的方法构造排序模型是信息检索领域中对排序方法研究的热点。基于人工标注的数据,排序学习[3] 算法构造出排序模型并且将其用于预测新的未标注数据。

当今已经提出百余种用于构建排序函数的特征,而从信息检索排序函数的构建方式易知,构成排序函数的特征之间并不是完全独立的,如 TF(Term Frequency)和 IDF(Inverse Document Frequency)这两个特征本身就是 BM25[4] 特征的组成部分,因此排序函数性能的影响因素还涉及到构成排序函数的特征集合的组成。当前的排序学习领域,对特征进行分析的研究较少,因此如何进行特征重组和选择,从而在新的特征集合上构建更为有效的排序函数是文中的重点。

收稿日期:2013-03-24 **修回日期**:2013-07-06 **网络出版时间**:2013-11-12
基金项目:国家级教学团队建设项目(00700054J1901)
作者简介:蒋宗礼(1956-),男,河南安阳人,博导,研究方向为网络信息搜索与处理;李涵昱(1986-),女,河南周口人,硕士生,研究方向为搜索引擎、排序学习。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20131112.1636.025.html

1　排序学习现用特征选择

1.1　特征选择

特征选择(Feature Selection)[5]是监督学习中很重要的一部分,具体指从候选的特征中筛选一个特征子集从而能构建出更健壮的学习模型。特征选择对于排序学习来说有两大优点[6]:

首先,特征选择可以用来提高机器学习的精确度。比如,虽然支持向量机(Support Vector Machines,SVM)具有边缘优化功能,并且这种功能使得当无用的特征增加时效果并不会发生改变,但支持向量机是基于训练数据到边缘的距离,所以当特征增多时训练数据到边缘的距离也会相应的增大。而且随着特征空间的不断增大,随之产生的问题还有过度拟合。然而特征选择是防止过度拟合的有效方法之一。

其次,特征选择可以提高训练的效率。在信息检索中,尤其是网络搜索引擎中,特征的数量相当的庞大,所有训练模型都非常耗时。为了解决这一问题,在训练模型之前可以先进行特征选择,因为大多数学习算法的时间复杂度和特征的数量成正相关关系。

1.2　排序学习算法

Ranking SVM[7]是一种有效的排序学习算法,它采用偏序的文档对作为训练样例,学习的优化目标是在排序函数对文档的排序中,文档逆序对的个数最少,逆序对是指将更为相关的文档排在后面的文档对,优化函数为:

$$\min \frac{1}{2}\boldsymbol{\omega} \times \boldsymbol{\omega} + C\sum \varepsilon_{i,j,k} \tag{1}$$

$$\forall (d_k^i, d_k^j) \in d_k \times d_k: \quad \boldsymbol{\omega}\varphi(q_k, d_k^i) > \boldsymbol{\omega}\varphi(q_k, d_k^j) + 1 - \varepsilon_{i,j,k} \tag{2}$$

式中,$\boldsymbol{\omega}$ 是在学习过程中需要逐步调整的权重向量;参数 C 是模型复杂性与训练误差之间的一个折中参数;$\varepsilon_{i,j,k}$ 是非零的松弛变量;$\varphi(q_k, d_k^i)$ 是从查询 q_k 与文档 d_k^i 到其对应的特征向量的一个映射。

1.3　面向排序学习的特征选择方法

目前在分类上特征选择被研究的比较多。用于分类的特征选择[8]一般分为三类。

第一类被称为过滤器(filter)方法。在这种方法中,特征选择被定义为一个预处理步骤,并且可以独立于学习过程。过滤器方法为每一个特征计算一个得分,并且根据得分进行特征选择。

第二类被称为包装(wrapper)方法。这种方法把学习系统作为一个黑盒,并对每一个特征子集进行打分。

第三种方法被称为嵌入式(embedded)方法。这种方法把特征选择作为训练过程中一部分。

虽然特征选择在分类上已经被广泛应用,但是在排序学习中特征选择的研究却很少。大多数情况下,直接把用于分类的特征选择的方法应用到排序学习之中。而排序和分类是不相同的。把用于分类的特征选择方法直接用于排序学习应该注意以下问题[9]:

(1)排序和分类有着明显的差距。在排序中,序列标号被用来区分元素在序列中的位置;而在分类中,类别信息相对比较平缓,只有相关或者不相关。

(2) 排序和分类的评价指标也是不一样的。例如在排序中排在前 n 位的往往比排在其他位置的更重要一些,而分类并不关注这些。所以并不能把用于分类的特征选择方法直接用于排序学习。

假设目标是从特征集合 $\{v_1, v_2, \cdots, v_m\}$ 中挑选出 t 个特征。特征重要性和特征相似性是两个重要参数。

① 特征重要性。

首先会对各个特征计算重要性得分。一般情况下,认为用评价指标比如说 MAP/NDCG 或者损失函数来计算特征重要性得分。文中采用前者,其计算方法为:用当前特征的权重对序列进行排序,然后用评价指标公式计算排序后的序列,评价指标的得分即为该特征的重要性得分。

② 特征相似度。

任意两个特征的相似度的计算基于用它们对待排序序列的排序结果,即把一个特征看成排序模型,两个特征的相似度就是它们参数的模型产生的序列的距离。有许多表示两个序列距离的方法,Kendall 定义的 τ 就是其中一种[10],文中的特征相似度即用 τ 来做相关说明。对一查询词或者文中所说的问题,其任意两个特征的 τ 值为:

$$\tau_q(v_i, v_j) = \frac{\#\{(d_s, d_t) \in D_q \mid d_s \prec_{Vi} d_t \quad \text{and} \quad d_s \prec_{Vj} d_t\}}{\#\{(d_s, d_t) \in D_q\}} \tag{3}$$

其中,D_q 表示由查询或者问题返回的文档对的集合;$\#\{\cdot\}$ 表示集合中含有元素的数量;$d_s \prec d_t$ 表示在序列中 d_s 排在 d_t 的前面。对所有查询或者问题来说,所有查询的 $\tau_q(v_i, v_j)$ 的平均值即为特征 v_i 和特征 v_j 的相似度 $\tau(v_i, v_j)$。由以上阐述可以很明显地看出 $\tau(v_i, v_j) = \tau(v_j, v_i)$。

目前用于排序的特征选择方法主要基于以下两个原则[11]:

(1) 特征重要性之和尽量大;

(2) 两两特征的相似度之和尽量小。

公式化后为:

$$\max \sum_i \omega_i x_i$$

$$\text{s.t. } x_i \in \{0,1\} \quad i = 1, \cdots, m \tag{4}$$

$$\sum_i x_i = t$$

$$\min \sum_i \sum_{j \neq i} e_{i,j} x_i x_j$$
$$\text{s.t. } x_i \in \{0,1\}\ i = 1, \cdots, m \quad (5)$$
$$\sum_i x_i = t$$

其中，t 表示选择的特征的数量；$x_i = 1$ 或者 $x_i = 0$ 表示特征 v_i 被选择或未被选择；ω_i 表示特征 v_i 的重要度；e_{ij} 表示特征 v_i 和特征 v_j 之间的相似度即 $e_{i,j} = \tau(v_i, v_j)$。

把公式合并后转换为：

$$\max \sum_i \omega_i x_i - c \sum_i \sum_{j \neq i} e_{i,j} x_i x_j$$
$$\text{s.t. } x_i \in \{0,1\}\ i = 1, \cdots, m \quad (6)$$
$$\sum_i x_i = t$$

相应的贪婪搜索算法（Greedy Search Algorithm of Feature Selection）基于构建一个无向图 G_0，图中的每一个节点表示一个特征 v_i，节点 v_j 的权重为 ω_i。连接节点 v_i 和节点 v_j 的边表示特征 v_i 和特征 v_j 的相似度，其权重即为特征 v_i 和特征 v_j 的相似度 $e_{i,j}$。该算法的主要步骤如下。

1）初始化存储被选择特征的集合 S 为 $\varnothing$。

2）for $i = 1, \cdots, t$

（1）挑选权重最大的节点，假设这个节点为 v_k；

（2）依据与节点 v_i 的相似度对其他节点做惩罚。同时其他节点的权重更新为：$\omega_j \rightarrow \omega_j - e_{k,j} * 2c, j \neq k$；

（3）添加 v_k 到集合 S，并从图 G 中剔除。

3）输出 S。

2 适应排序学习特征选择的锦标赛排序算法

从上述贪婪搜索算法的具体求解过程可以发现，当特征数量特别大的时候，构建的无向图将会特别庞大，而且单纯考虑两特征的相似度并不是特别合适。基于上述考虑，提出了循环特征选择方法-锦标赛排序方法，基本步骤如下。图 1 是其流程图。

（1）假设原始特征集合为 $S[\,]$。同时确定最大循环次数 signal。

（2）构建集合 V，V 用来存储被选择的特征，初始化 $V_0 = \varnothing$。同时初始化排序评价指标：value = 0。

（3）选取集合 $S[\,]$ 比较离散的若干个点作为聚类的初始中心点。

（4）对集合 $S[\,]$ 进行聚类划分。

（5）对各个类中的特征按重要性进行排序。

（6）遍历各个类中的特征，并将该特征添加入集合 V，用集合 V 训练排序模型。如果新模型的评价指标大于原有模型的评价指标，则处理下一个类。否则到（7）。

（7）如果该类的所有特征已经遍历完，则处理下一个分类。否则处理该分类中下一个元素。

（8）如果循环次数大于最大循环次数 signal，则算法结束。否则跳到（3）。

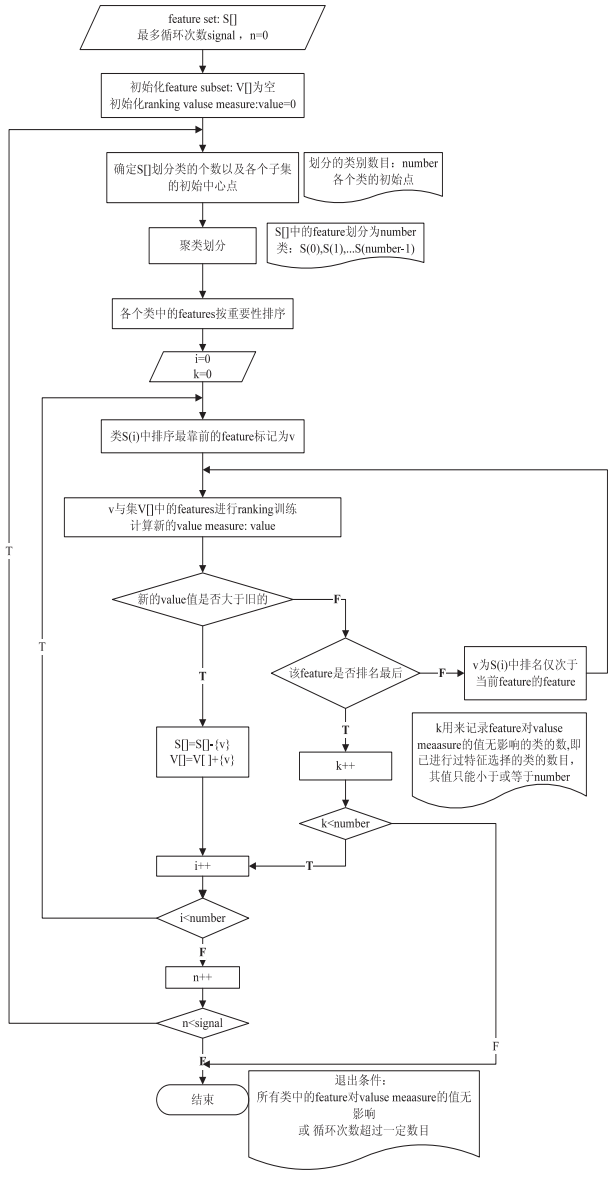

图 1 锦标赛排序算法流程图

在第一部分已经介绍，一般用评价指标 MAP 和 NDCG 来评价信息检索系统性能。由于 QA 问题和传统信息检索的不同之处：传统的检索系统需要根据查询词确定需要从文档库中召回哪些文档；而在 QA 问题中，问题的答案已经存在，不需要根据问题确定召回哪些答案。因此在 QA 问题中不存在召回相关的评价指标。在文中只用 NDCG 来评价排序，即用 NDCG 值作为 ranking values measure：value。

3 锦标赛排序算法用于 QA 问题的实验

3.1 数据集获取

需要从 QA 论坛上抓取一些问题答案对（QA，

thread)来提取用于获取排序模型的训练数据。文中选择的 QA 论坛为百度知道。在抓取到“百度知道”的问题答案对后,对问题和答案进行分析,抽象出一些特征,用这些特征组成的特征向量来表示问题和答案的相关度。具体数据集获取过程如图 2 所示。

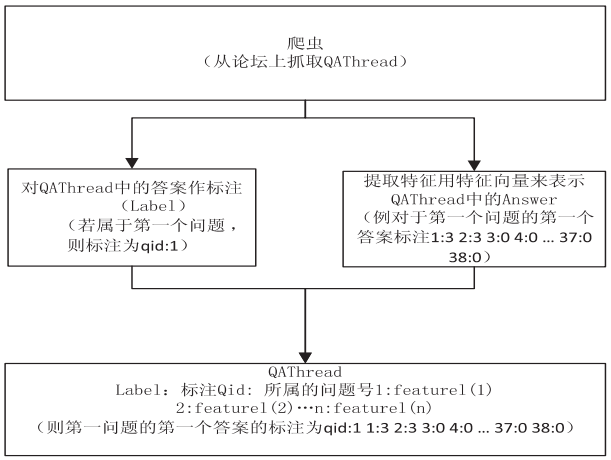

图 2　数据集合获取过程

“百度知道”是用户自己根据自身需求提出具有针对性的问题,通过积分奖励机制发动其他用户,来解决该问题的一个中文问答系统。同时,这些问题的答案会进一步地作为搜索结果,如果其他用户有类似疑问,则通过搜索就可以获取相关答案。

百度知道对所提问的问题进行了相关分类,具体有“电脑/网络”、“生活”、“医疗健康”、“体育/运动”、“电子数码”、“商业/理财”、“教育/科学”、“社会民生”等。该数据集取自“百度知道”的“电脑/网络”分类,一共从中抓取了 2 208 个 QA thread,其中包括问题 2 208 个,答案 11 342 个。

3.2　特征提取

用能表示答案和问题相关性的特征组成的特征集合来表示问题答案对中的答案。这些特征有些来自答案本身,例如答案中所包含的文字信息,从文字上和问题的匹配度,其回答时间,回答时间和问题提问时间的时间差,等等;有些特征来自问题回答者,例如回答者级别,回答者在该领域回答过多少答案,在该领域回答的答案都有多少答案被标注为“满意答案”,等等;另外一部分答案来自提问者对答案的反应,例如提问者对该问题是否有追加问题,有多少个追加问题等等。

3.3　答案标注

已知现在的问题系统一般通过问题提问者标注“最佳答案”(百度知道中的“满意答案”)来确定最好答案,其他答案被认为是不相关的答案。但是这样存在很大弊端。下面将介绍此次实验答案标注方法。

实验安排 3 个计算机专业的学生作为标注者对抓取到的 QA thread 中的答案同时进行标注。标注者依据以下准则对答案进行相应标注。

1)答案完全满足问题信息需求。

2)答案部分满足问题信息需求。

3)答案完全不满足问题信息需求。

把标注级别映射为权重(judgment weight),在文中的研究中,简单的让 A:B:C=2:1:0。然后根据其权重确定答案级别。因此可以得到映射关系如表 1。

表 1　标注、权重与分级

标注	权重	分级
AAA	6	L_6
AAB	5	L_5
ABB	4	L_4
BBB	3	L_3
BBC	2	L_2
BCC	1	L_1
CCC	0	L_0

例如,对 QA thread 中一个答案,从标注者处获取的标注组合为“AAA”,那么该答案的权重为 3 * 2=6,同时该答案的级别为 L_6。

然后对 2 028 个问题答案标注为“满意答案”进行统计分析,统计出其所处的级别,具体情况如表 2。

表 2　统计数据

分级	L_0	L_1	L_2	L_3	L_4	L_5	L_6
实测数据	3	7	23	38	97	147	803

由表 2 可以看出,在“百度知道”的“电脑/网络”分类中,被标注为“满意答案”的答案只有不到一半的级别为 L_6,即提问者标注“最佳答案”在多数情况下并不是最佳答案。这也充分说明了对 QA 论坛中的答案进行排序的重要性。

3.4　数据集

前面已经介绍了怎样把答案抽象为一个特征向量,以及怎么对答案做相应的标注。得到的数据生成如图 3 所示。

```
0 qid:1 1:3 2:0 3:2 4:2 … 37:0 38:0
2 qid:1 1:3 2:3 3:0 4:0 … 37:0 38:0
```

图 3　生成的数据集

其中第一列表示对该答案所做的标注得分,第二列的数据表示问题的标号,其余列表示特征编号和其对应的特征值。以第一行数据来做相应说明:第一列“0”表示该答案的标注得分为 0,即该答案和问题完全不相关;“qid:1”表示该答案所属的问题标号为 1;“1:3 ”表示第一个特征的特征值为 3。

把生成的数据集分成两部分,其中 2/3 用作训练集合用锦标赛排序的特征选择方法来训练排序模型,另外 1/3 用作测试集来测试排序模型的质量。

4 实验结果分析

用“特征选择方法一”表示改进前的贪婪搜索算法特征选择方法,用“特征选择方法二”表示文中新提出的基于锦标赛排序的特征选择方法。在这一节中将两种不同的特征选择方法作用于原始数据集来生成新的数据集。经过特征选择方法后,把用特征选择方法一生成的新的数据集称为新数据集一,把用特征选择方法二生成的新的数据集称为新数据集二。其中新数据集一包含特征 28 个,新数据集二包含特征 25 个。

图 4 为两种特征选择方法在特征选择过程中 DNCG@1 的变化过程。

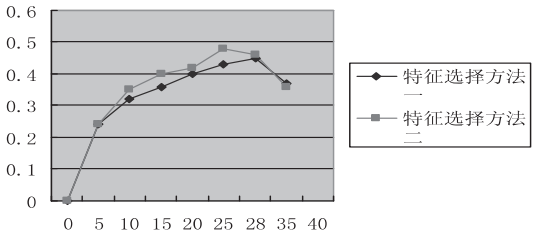

图 4 DNCG@1 结果图

从图中可以明显看出:

(1)无论是特征选择方法一还是特征选择方法二,都在挑选到 25 个左右的时候 DNCG 值达到最高,然后开始逐渐地下降。这充分说明用所有特征参与排序训练得到的模型并不一定是最好的模型。

(2)这两种不同的特征选择方法,选择出来的特征并不完全相同,但是却几乎都可以提升检索性能。所以面向排序学习方法的特征选择过程并不是对单独的特征的选择,而是要选择出一个特征子集,子集中的每个特征对最终的排序函数都有一定的贡献,特征的组合方式也对最终的排序函数有一定的贡献。

(3)相比于特征选择方法一,特征方法二能更高效地从所有特征组成的特征集合中挑选出适合的特征子集。

5 结束语

文中基于针对 QA 问题的排序学习算法中特征选择方法的现状,提出了其在实际应用中的一些弊端,并由此进行了改进。针对具体的 QA 问题,把锦标赛排序的思想应用于排序学习的特征选择中,提出了锦标赛排序的特征选择算法,并从理论和实际两个不同的方面展开研究,取得了一定的成果。

参考文献:

[1] Tetsuya S, Daisuke I, Noriko K. Using graded-relevance metrics for evaluating community QA answer selection[C]//Proc of WSDM'11. [s. l.]:[s. n.],2011.

[2] Suzan V, Hans H, Daphne T. Learning to rank QA data[C]//Proc of SIGIR. [s. l.]:[s. n.],2009.

[3] Tao Qin, Liu Tieyan, Zhang Xudong, et al. Learning to rank relational objects and its application to Web search[C]//Proc of WWW 2008. [s. l.]:[s. n.],2008.

[4] 郑实福,刘 挺,秦 兵,等. 自动问答综述[J]. 中文信息学报,2002,16(6):46-52.

[5] 陈 彬,洪家荣,王亚东. 最优特征子集选择问题[J]. 计算机学报,1997,20(2):17-22.

[6] Feng P, David A, Franco S. Feature selection for ranking using boosted trees[C]//Proc of CIKM'09. [s. l.]:[s. n.],2009.

[7] 代六玲,黄河燕,陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报,2004,18(1):26-32.

[8] 刘丽珍,宋瀚涛. 文本分类中的特征选取[J]. 计算机工程,2004,30(4):14-15.

[9] 章 兰. 一种基于 VSM 模型的动态文本分类器的设计[D]. 苏州:苏州大学,2004.

[10] 万 忠,张燕平,张 铃,等. 基于覆盖算法决策界的特征选择算法[J]. 计算机技术与发展,2006,16(4):84-87.

[11] Hua Guichun, Zhang Min, Liu Yigun, et al. Hierarchical feature selection for ranking[C]//Proc of WWW 2010. [s. l.]:[s. n.],2010.

(上接第 49 页)

detail control[C]//Proc of annual conference series on computer graphics. San Antonio, Texas: ACM Press, 2002.

[3] Losasso F, Hoppe H. Geometry clipmaps: Terrain rendering using nested regular grids[C]//Proc of annual conference series on computer graphics. New York: ACM Press, 2004: 769-776.

[4] 宫晓辉,温慧明,于 卓. 基于 CLipmap 的海量地形及纹理实时绘制方法[J]. 计算机技术与发展,2012,22(10):22-26.

[5] 邹 海,徐 军,褚维翠. 基于 OpenGL 的三维地形的模拟[J]. 计算机技术与发展,2011,21(6):239-241.

[6] 吴 颖,张新家,茹 芬. 基于四叉树分割的连续 LOD 漫游地形绘制[J]. 计算机技术与发展,2011,21(4):5-8.

[7] 申闫春,朱幼虹,温转萍,等. Chunklod 地形的过程化细节度增强算法[J]. 系统仿真学报,2008,20(21):5763-5766.

[8] 张立民,闫文君. 基于 GPU 的大规模地形数据绘制算法[J]. 计算机与现代化,2012(1):145-150.

[9] 刘慧媛. 三维海底地形绘制方法研究与实现[D]. 哈尔滨:哈尔滨工程大学,2009.

[10] 马淑芳. 基于 GPU 加速的分形地形生成方法[D]. 大连:大连理工大学,2008.

[11] 廖学军,王荣峰. 数字战场可视化与应用[M]. 北京:国防工业出版社,2010:39-42.

[12] Strugar F. Continuous distance-dependent level of detail for rendering heightmaps[J]. Journal of graphics GPU and game tools, 2009, 14(4):57-74.

面向排序学习的锦标赛排序特征选择方法

作者： 蒋宗礼， 李涵昱， JIANG Zong-li， LI Han-yu
作者单位： 北京工业大学 计算机学院,北京,100124
刊名： 计算机技术与发展
ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(2)