

基于 ListNet 排序学习的特征处理方法

李伟宁¹, 王磊²

(1. 南京邮电大学 计算机学院, 江苏 南京 210003;
2. 南京邮电大学 电子科学与工程学院, 江苏 南京 210003)

摘要: 排序学习 (learning to rank) 是一种机器学习与信息检索的交叉学科, 可以从大量的包含标记的训练集中自动学习排序模型。特征选取对于排序模型的预测结果有很大的影响, 而排序学习对其特征领域的研究却很少。针对这一问题, 提出一种特征处理方法: 利用基于主成分分析 (PCA) 的特征重组方法扩展数据集, 然后在扩展后的数据集上进行排序算法隐含的特征选择。在 LETOR4.0 数据集 (MQ2007, MQ2008) 上基于排序评测函数对 ListNet 排序算法进行验证。通过对比特征处理前后的排序性能差异, 以及添加新特征的个数对排序结果的影响, 实验结果表明, 经过特征处理的利用排序学习算法构建的排序函数一般要优于原始的排序函数。

关键词: 信息检索; 排序学习; 特征处理; ListNet

中图分类号: TP301

文献标识码: A

文章编号: 1673-629X(2018)09-0030-04

doi: 10.3969/j.issn.1673-629X.2018.09.007

A Feature Processing Method Based on Ranking Algorithm ListNet

LI Wei-ning¹, WANG Lei²

(1. School of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. School of Electronic Science and Engineering, Nanjing University of Posts and Telecommunications,
Nanjing 210003, China)

Abstract: Learning to rank is an interdisciplinary of machine learning and information retrieval and learns ranking model automatically from given training data set. The feature space has a great influence on the performance of learning to rank approach, however, there are a little research in terms of feature generation. For this, we propose one feature analysis method which extends data set by feature recombination based on PCA, and then performs feature selection implied by learning to rank methods on the extended data set. We evaluate ranking algorithm ListNet on the LETOR4.0 (MQ2007, MQ2008) data set based on ranking evaluation index, and experimentally compare the performance of ListNet using the data set with new feature vectors and not, as well as the impact of the number of the new features added to the result of sort. The experiment shows that ranking functions learned through learning to rank method based on the feature analysis methods outperform the original ones.

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

0 引言

排序是信息检索系统的一个重要组成部分, 为了提高用户检索的准确性, 将更为相关有效的查询页面返回给用户, 如何提高搜索的排序质量就显得尤为重要。随着搜索引擎的发展, 对于某个网页进行排序需要考虑的因素越来越多, Google 目前的网页排序公式考虑 200 多种因子, 如 TFIDF、BM25、PageRank、HITS 以及基于点击数据的用户行为的特征等, 此时机器学习的作用即可发挥出来^[1]。

特征以及特征的组合方式对于任何以机器学习算法为依托的任务来说都是至关重要的, 并与机器学习算法结合形成了复杂的算法体系^[2]。当前对排序学习的特征处理的研究较少, 文中的研究重点是对特征进行有效的处理, 从而在新的特征集合上构建更加有效的排序函数。

1 相关研究

排序学习尝试用机器学习的方法解决排序问题,

收稿日期: 2017-08-17

修回日期: 2017-12-27

网络出版时间: 2018-04-28

基金项目: 国家“863”高技术发展计划项目 (2006AA01Z201)

作者简介: 李伟宁 (1993-), 女, 硕士研究生, 研究方向为机器学习、数据挖掘。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20180427.1626.008.html>

已得到深入研究并广泛应用于不同的领域,如信息检索、个性化推荐、文本挖掘、生物医学等^[3-5]。排序学习算法根据其数据形式的不同主要分为三类:Pointwise 型,如基于感知器的 Prank 算法^[6],将问题转换为单个文档的分类和回归问题;Pairwise 型,如用神经网络模型结合梯度下降算法去优化排序损失函数的 RankNet 算法^[7]和使用支持向量机的 RankingSVM 算法^[8-9];Listwise 型,如基于神经网络和梯度下降优化方法,应用概率模型来构造损失函数的 ListNet 算法^[10]和基于直接优化信息检索评价方法的 LambdaMART 算法^[11]。研究表明^[12],Listwise 方法的排序效果一般要优于 Pointwise 及 Pairwise 方法。因此,文中采用 Listwise 型中经典的 ListNet 算法进行比较研究。

在排序学习领域,已经提出了很多方法来提高排序的准确率,例如提出一种新的排序学习算法^[13],构建一种新的优化函数,设计新的特征。然而,在排序技术较为成熟的基础上,如果仅靠改进排序学习算法来提高排序结果已经变得越来越难。那么添加由原始特征生成的新特征是否会提升准确率,一些学者进行了研究。Amini 等^[14]假设若找到一个与已标注文档最为相似的未标注的文档,那么就认为这个未标注的文档与已标注的文档的相关性条件也是相似的,使用这些文档实例来扩展训练数据集。Duh 和 Kirchoff^[15]应用基于核的主成分分析方法对测试集合抽取新的特征集合,利用测试集合的特征来扩展训练集合,获得了好的效果。林等^[16]将训练集与测试集进行合并后的集合进行奇异值分解,提取新的特征集合加入训练集,在 RankBoost 排序算法上提高了排序效果。

特征选择是从一组特征集中挑选出一部分最有效的特征以降低特征空间维数的过程^[17]。特征选择对排序学习有两大优点:第一,可以提高排序学习的精确度;第二,可以提高训练的效率。文中综合利用 PCA 方法扩展数据集,并在扩展后的数据集上进行排序算法隐含的特征选择。

2 基于排序学习的特征处理

2.1 基于 PCA 的数据集扩展

查询和文档集的特征组成的特征向量 $\mathbf{X} = (x_1, x_2, \dots, x_n)$, 求取特征向量 \mathbf{X} 的自相关矩阵 \mathbf{R}_x , 根据 PCA 原理对 \mathbf{R}_x 进行特征分解, 得到特征向量矩阵 \mathbf{A} 和特征矩阵 \mathbf{U} , 其中特征矩阵 \mathbf{U} 为:

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1} & u_{m2} & \cdots & u_{mn} \end{bmatrix}$$

令 $\mathbf{Y} = \mathbf{UX}$, \mathbf{Y} 即为 \mathbf{X} 的 K-L 变换。其中, \mathbf{U} 的每一列向量均包含了 \mathbf{X} 的相关信息, 将其作为 \mathbf{X} 的组合特征。选取最大特征值对应的特征向量作为 \mathbf{X} 的融合特征添加到原始数据集中, 在扩展训练数据集的同时达到特征融合的目的。算法思想如下:

输入: 训练集 D , 验证集 V , 测试集 T ;

输出: 新的训练集 D' , 新的验证集 V' , 新的测试集 T' 。

(1) 对训练集所对应的特征矩阵进行主成分分析, 选取前 k 个主元;

(2) 用选取的前 k 个主元对原训练集、验证集、测试集进行主成分分析, 分别得到它们的前 k 个特征向量集 $\mathbf{D}_k, \mathbf{V}_k, \mathbf{T}_k$;

(3) $D \cup \mathbf{D}_k \Rightarrow D', V \cup \mathbf{V}_k \Rightarrow V', T \cup \mathbf{T}_k \Rightarrow T'$, 添加后的特征向量维度比原始维度大 k 。

文中综合排序训练时间和排序效果选择加入合适的特征向量数目。

2.2 基于排序学习的特征选择

在基于 PCA 扩展之后的数据集上进行排序算法隐含的特征选择, 算法思想如下:

输入: 特征集合 $X = (x_1, x_2, \dots, x_{n+k})$;

输出: 特征子集 $S = (s_1, s_2, \dots, s_m), m \leq n + k$ 。

(1) 对基于 PCA 方法扩展之后得到的训练集和验证集使用排序学习算法得到排序函数;

(2) 将上述得到的排序函数的特征系数的绝对值作为对应特征的权重;

(3) 做五组实验, 对五组实验得到的特征权重求和取平均值, 权重从大到小排序;

(4) 取前 m 个特征构建新的训练集、验证集和测试集。

2.3 排序学习算法

Listwise 型排序学习方法是整个文档序列看作一个样本^[18]。其中, ListNet 方法是将排序问题建模为概率模型, 然后选取交叉熵衡量由模型训练出的文档排序和真正的文档排序之间的差异, 通过最小化这个差异值来完成排序。ListNet 定义了一个置换概率:

$$P_s(\pi) = \prod_{j=1}^{n(q_i)} \frac{\Phi(S_{\pi(j)})}{\sum_{k=j}^{n(q_i)} \Phi(S_{\pi(k)})} \quad (1)$$

其中, π 为作用于 $n(q_i)$ 个文档的置换; $\Phi(\cdot)$ 为一个递增、恒正的函数; S 为一个计算文档与查询相关度评分的函数; $S_{\pi(j)}$ 为在置换 π 中第 j 个位置的文档的评分值。

以交叉熵的形式定义损失, 优化目标的公式如下:

$$L = - \sum_{i=1}^{N_q} \log \prod_{t=1}^{n(q_i)} \frac{\exp(f_{\omega}(x_{y(t)}^{(i)}))}{\sum_{k=t}^{n(q_i)} \exp(f_{\omega}(x_{y(k)}^{(i)}))} \quad (2)$$

3 实验

3.1 实验数据

实验采用的是微软亚洲研究院 (MSRA) 提供的 LETOR4.0^[19] 数据集中的 MQ2007 和 MQ2008 两个数据集。LETOR4.0 采用了 Gov2 的网页集作为原始数据,使用从百万查询跟踪 TERC2007 和 TERC2008 的两个查询集,每个查询有许多相关联的文档。其中,每个查询-文档对的相关度标注为三个等级 (2,1,0),分别对应非常相关、部分相关和不相关。对数据集抽取了基于链接、内容等的 46 个特征。

在实验过程中,采用了交叉验证的方法。数据被平分成五份,用 S_1 、 S_2 、 S_3 、 S_4 、 S_5 表示。做五次训练,再取五次结果的平均数,这样做可以避免数据过拟合的情况,保证结果的可信度。数据集划分见表 1。

表 1 五折交叉验证数据集划分

Folds	训练集	验证集	测试集
Fold ₁	{ S_1, S_2, S_3 }	S_4	S_5
Fold ₂	{ S_2, S_3, S_4 }	S_5	S_1
Fold ₃	{ S_3, S_4, S_5 }	S_1	S_2
Fold ₄	{ S_1, S_4, S_5 }	S_2	S_3
Fold ₅	{ S_1, S_2, S_5 }	S_3	S_4

3.2 评测函数

为了衡量排序函数性能的优劣,应用两种常用的排序性能评测函数:MAP 和 NDCG@n。

MAP (mean average precision) 的衡量标准单一, q 与 d 的关系非 0 即 1,核心是利用 q 对应的相关的 d 出现的位置来进行排序算法准确性的评估。系统检索出来的相关文档越靠前,MAP 值就可能越高。对于一个查询 q_i ,其平均查准率的计算公式为:

$$AvgP_i = \sum_{j=1}^M \frac{precision(j) \times pos(j)}{R_j} \tag{3}$$

其中, j 表示排序序列的位置; M 表示返回的文档总数; R_j 表示第 j 个查询的相关文档数目; $precision(j)$ 表示第 j 个检索到的文档的查准率; $pos(j)$ 定义为:

$$pos(j) = \begin{cases} 1, & \text{第 } j \text{ 位上的文档相关} \\ 0, & \text{第 } j \text{ 位上的文档不相关} \end{cases} \tag{4}$$

用 N_q 表示查询数量,平均查准率的均值 MAP 为:

$$MAP = \frac{\sum_{i=1}^{N_q} AvgP_i}{N_q} \tag{5}$$

NDCG (normalized discount cumulative gain) 是用来评估排序结果中顶部序列的准确性的评价指标, NDCG 的值越高,表示排序结果中的顶部序列的贡献越大,排序的结果也就越好,可以接受多种相关度的评

分。给定查询 q_i 以及对应的返回文档序列,其折扣累积增益 (DCG) 定义为:

$$DCG@n = \sum_{j=1}^n \frac{2^{r(j)} - 1}{\log(1 + j)} \tag{6}$$

其中, $r(j)$ 是第 j 个文档的等级,文中 $n = 1, 3, 5, 7, 10$ 。

为使不同等级上的搜索结果的得分值容易比较,需要对 DCG 做归一化处理,归一化的 DCG 值叫做 NDCG。计算公式定义为:

$$NDCG@n = \frac{DCG@n}{IDCG@n} \tag{7}$$

其中, IDCG 为理想的 DCG,即通过人工对查询的搜索结果进行排序,排到最好的状态后,计算得出的这个排列下本查询的 DCG。

3.3 实验结果及分析

随着向原始数据集添加 PCA 处理得到的主特征个数的增加,两个数据集的 MAP 值的变化如图 1 所示。考虑到模型的训练时间,这里最多只添加 15 个主特征。从图中可以看出,虽然随着添加个数的增加,MAP 值出现了下降的趋势,但最终的排序效果仍要优于原始的排序结果。

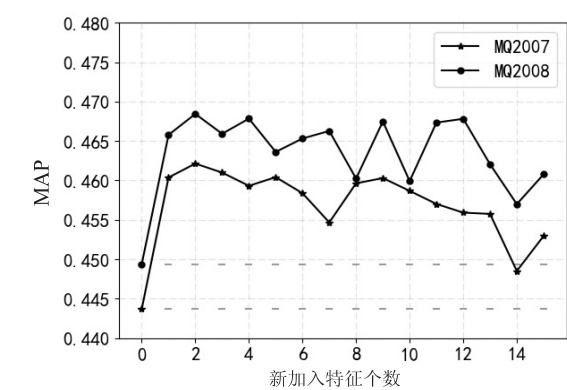


图 1 新加入的特征个数对排序结果的影响

加入较多的 PCA 主特征向量会增加排序模型的训练时间,引入较少的特征向量则会因数据特征信息丢失过多导致最终的排序结果改善不明显。综合考虑排序效果和模型训练时间,选择添加前 4 位主特征向量。在扩展后的数据集上进行排序算法隐含的特征选择,随着新加入特征个数的逐渐增加,排序函数在两个测试集上的 MAP 值的变化趋势如图 2 所示。

从图中可以看出,随着选择的特征个数的逐渐增加,查询集的排序精度并不是一直增高,甚至有所下降,这也验证了进行特征选择的必要性。在 MQ2007 数据集,特征大小为 25 时,MAP 值就达到了特征选择之前的效果,并趋向于稳定,选择特征集合的大小为 31。MQ2008 数据集上的特征大小为 7 时就达到了特征选择之前的效果,但波动较大,大小为 29 之后趋于稳定,选择特征集合的大小为 33。

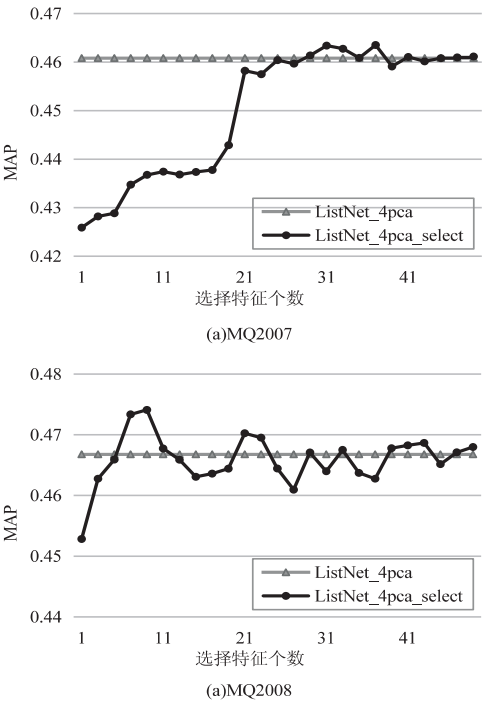


图 2 特征子集的大小对 MAP 值的影响

确定了要选择的特征个数,在基于扩展的数据集上进行特征选择的实验结果如图 3 所示,其中 ListNet 表示不进行特征处理,ListNet_select 表示只进行特征选择,ListNet_4pca 表示只添加前 4 位主特征向量,ListNet_4pca_select 表示在数据集扩展的基础上进行特征选择。

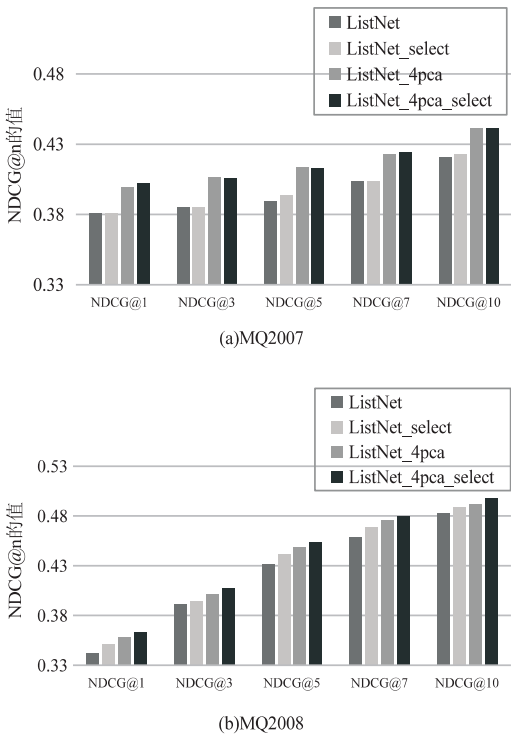


图 3 应用特征选择的排序结果对比

从图 3 可以看出,三种特征处理的方法基本都使排序结果有所提高。在基于 PCA 特征重组方法扩展

后的数据集上进行排序算法隐含的特征选择之后得到的排序效果总体上是最优的,虽然在 MQ2007 数据集上 NDCG@3 的值要略低于特征选择之前的值,但是显著优于原始的排序函数得到的效果。

4 结束语

使用特征处理方法对构建排序函数的特征进行处理,并用排序学习算法 ListNet 进行验证,实验结果表明,经过特征处理后,对排序学习的排序结果的提高有积极意义。

在特征选择算法中一个很难解决的问题是什么样的特征子集是最优的。希望用尽量少的特征,得到最好的输出结果。文中没有对选择的特征子集的大小做分析,只是简单地增加选择的特征个数。在后面的工作中,将继续研究如何在特征数和输出精度中得到一个很好的平衡。

参考文献:

[1] 张俊林. 这就是搜索引擎:核心技术详解[M]. 北京:电子工业出版社,2012:26-27.

[2] 李 敏,卡米力·木依丁. 特征选择方法与算法的研究[J]. 计算机技术与发展,2013,23(12):16-21.

[3] 黄震华,张佳雯,田春岐,等. 基于排序学习的推荐算法研究综述[J]. 软件学报,2016,27(3):691-713.

[4] 印 鉴,王智圣,李 琪,等. 基于大规模隐式反馈的个性化推荐[J]. 软件学报,2014,25(9):1953-1966.

[5] 程 凡. 基于排序学习的信息检索模型研究[D]. 合肥:中国科学技术大学,2012.

[6] CRAMMER K, SINGER Y. Pranking with ranking[C]//Processing of the conference on neural information processing systems. Vancouver, British Columbia: [s. n.], 2002:641-647.

[7] SONG Yang, WANG Hongning, HE Xiaodong. Adapting deep RankNet for personalized search[C]//Proceedings of the 7th ACM international conference on web search and data mining. New York:ACM,2014:83-92.

[8] CAO Yunbo, XU Jun, LIU Tieyan, et al. Adapting ranking SVM to document retrieval[C]//Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval. Seattle, Washington, USA:ACM,2006:186-193.

[9] CAO Houwei, VERMA R, NENKOVA A. Speaker-sensitive emotion recognition via ranking: studies on acted and spontaneous speech[J]. Computer Speech & Language, 2015, 29(1):186-202.

[10] CAO Zhe, QIN Tao, LIU Tieyan, et al. Learning to rank: from pairwise approach to listwise approach[C]//International conference on machine learning. Corvallis, OR:ACM,

会提升,因此挑选阈值 T 为 0.9 时,精确度较高的结果进行下一轮迭代,结果如表 2 所示。

表 2 迭代结果 %

迭代次数	precision	recall	F -score
1	65.29	58.19	61.54
2	63.12	64.50	63.81
3	63.16	64.77	63.96

从表 2 可以看出,随着迭代次数的增加,种子集合扩充,有交互蛋白质对的召回率提升,精确度略有下降,整体 F 值上升。实验结果表明,该方法以少量的初始种子取得了较高的精确度和召回率,3 次迭代后的 F 值可达到 63.49%。

3 结束语

文中提出了一种基于分布式假设的弱监督蛋白质交互识别方法。该方法仅需少量有交互关系的蛋白质对作为种子集,根据分布式假设构建向量空间模型,利用相似性识别出有交互的蛋白质对。实验结果表明,该方法以较少的种子取得了较高的精确度与召回率。

目前该方法只考虑了利用签名档中两个蛋白质中间部分的文本来构造词汇模式,之后的研究将考虑提取句子中其他部分的关键信息作为特征来表示蛋白质对的交互关系。

参考文献:

[1] PRASAD T S K,GOEL R,KANDASAMY K,et al. Human protein reference database-2009 update[J]. Nucleic Acids Research,2009,37:767-772.

[2] KERRIEN S,ALAM-FARUQUE Y,ARANDA B,et al. IntAct-open source resource for molecular interaction data [J]. Nucleic Acids Research,2007,35:561-565.

(上接第 33 页)

2007:129-136.

[11] BURGESS C J C. From ranknet to lambdarank to lambdamart: an overview[M]. [s. l.]:[s. n.],2010.

[12] DING Yuxin,ZHOU Di,XIAO Min,et al. Learning to rank relational objects based on the listwise approach[C]//International joint conference on neural networks. San Jose,CA, USA:IEEE,2010:1818-1824.

[13] 奚凌然,王小平. 一种结合 LPA 半监督学习的排序学习算法[J]. 计算机应用与软件,2016,33(1):286-290.

[14] AMINI M R,TRUONG T V,GOUTTE C. A boosting algorithm for learning bipartite ranking functions with partially labeled data[C]//International ACM SIGIR conference on research and development in information retrieval. [s. l.]:ACM,2008,99:106.

[3] LICATA L,BRIGANTI L,PELUSO D,et al. MINT,the molecular interaction database:2012 update[J]. Nucleic Acids Research,2007,40:572-574.

[4] KOIKE A,KOBAYASHI Y,TAKAGI T. Kinase pathway database:an integrated protein-kinase and NLP-based protein-interaction resource[J]. Genome Research,2003,13:1231-1243.

[5] 杨志豪,洪莉,林鸿飞,等. 基于支持向量机的生物医学文献蛋白质关系抽取[J]. 智能系统学报,2008,3(4):361-369.

[6] 唐楠,杨志豪,林鸿飞,等. 基于多核学习的医学文献蛋白质关系抽取[J]. 计算机工程,2011,37(10):184-186.

[7] GRIMES G R,WEN T Q,MEWISSEN M,et al. PDQ Wizard:automated prioritization and characterization of gene and protein lists using biomedical literature[J]. Bioinformatics,2006,22(16):2055-2057.

[8] ANANIADOU S,KELL D B,TSUJII J. Text mining and its potential applications in systems biology[J]. Trends in biotechnology,2006,24(12):571-579.

[9] NIU Y,OTASEK D,JURISICA I. Evaluation of linguistic features useful in extraction of interactions from PubMed; application to annotating known,high-throughput and predicted interactions in I2D[J]. Bioinformatics,2010,26(1):111-119.

[10] 崔宝今,林鸿飞,张霄. 基于半监督学习的蛋白质关系抽取研究[J]. 山东大学学报:工学版,2009,39(3):16-21.

[11] 董美豪. 基于文本挖掘的蛋白质相互作用对抽取方法的研究[D]. 哈尔滨:哈尔滨工业大学,2015.

[12] 高飞. 基于 MapReduce 的蛋白质相互作用信息抽取系统的设计与实现[D]. 杨凌:西北农林科技大学,2016.

[13] 刘敏捷. 基于组合学习和主动学习的蛋白质关系抽取[D]. 大连:大连理工大学,2015.

[14] HARRIS Z S. Distributional structure[J]. Word,1954,10(2-3):146-162.

[15] DUH K,KIRCHHOFF K. Learning to rank with partially-labeled data[C]//ACM special interest group on information retrieval. Singapore:ACM,2008:251-258.

[16] LIN Yuan,LIN Hongfei,YANG Zhihao,et al. A boosting approach for learning to rank using SVD with partially labeled data[C]//5th Asia information retrieval symposium on information retrieval technology. Sapporo, Japan:[s. n.],2009.

[17] 边肇祺,张学工. 模式识别[M]. 第 2 版. 北京:清华大学出版社,2000.

[18] 程凡,李龙澍. 基于 Listwise 的新型排序算法[J]. 计算机工程,2011,37(23):165-167.

[19] QIN Tao,LIU Tieyan,XU Jun,et al. LETOR:a benchmark collection for research on learning to rank for information retrieval[J]. Information Retrieval,2010,13(4):346-374.