◎数据库、信号与信息处理◎

面向排序学习的特征分析的研究

花贵春,张 敏,邝 达,刘奕群,马少平,茹立云
HUA Guichun,ZHANG Min,KUANG Da,LIU Yiqun,MA Shaoping,RU Liyun
清华大学 计算机系,智能技术与系统国家重点实验室,清华信息科学与技术国家实验室(筹),北京 100084
Tsinghua National Lab for Information Science and Technology,State Key Lab of Intelligent Technology and Systems,Department of Computer Science and Technology,Tsinghua University,Beijing 100084,China

HUA Guichun,ZHANG Min,KUANG Da,et al.Feature analysis methods for learning to rank.Computer Engineering and Applications,2011,47(17):122-127.

Abstract: Ranking is an essential part of information retrieval.Nowadays there are hundreds of features for constructing ranking functions and it is a hot research topic that how to use these features to construct more efficient ranking functions.So learning to rank,an interdisciplinary field of information retrieval and machine learning,has attracted increasing attention.The construction methods of ranking features show that the features are not independent from each other.However, the state-of-the-art learning to rank approaches merely analyze the features from the aspects of feature recombination and selection for constructing more efficient ranking functions.In this paper,the model structure is proposed.Firstly the features are analysed for constructing the ranking functions.Secondly the features are recombined and selected,and finally ranking functions are learnt through learning to rank methods.And four methods are proposed based on this structure:feature recombination based on principal component analysis,feature selection based on MAP,forward selection and feature selection implied by learning to rank methods.The experimental results show that ranking functions learned through learning to rank methods based on the feature analysis methods outperform the original ones.

Key words: learning to rank;ranking function;feature recombination;feature selection

摘 要: 排序是信息检索中一个重要的环节,当今已经提出百余种用于构建排序函数的特征,如何利用这些特征构建更有效的排序函数成为当今的一个热点问题,因此排序学习(Learning to Rank),一个信息检索与机器学习的交叉学科,越来越受到人们的重视。从排序特征的构建方式易知,特征之间并不是完全独立的,然而现有的排序学习方法的研究,很少在特征分析的基础上,从特征重组与选择的角度,来构建更有效的排序函数。针对这一问题,提出如下的模型框架:对构建排序函数的特征集合进行分析,然后重组与选择,利用排序学习方法学习排序函数。基于这一框架,提出四种特征处理的算法:基于主成分分析的特征重组方法、基于MAP、前向选择和排序学习算法隐含的特征选择。实验结果显示,经过特征处理后,利用排序学习算法构建的排序函数,一般优于原始的排序函数。

关键词: 排序学习;排序函数;特征重组;特征选择
DOI:10.3778/j.issn.1002-8331.2011.17.033 **文章编号:** 1002-8331(2011)17-0122-06 **文献标识码:** A **中图分类号:** TP311

1 引言

当用户需要从大规模数据集或者互联网上快速查找某一方面信息的时候,往往需要求助于信息检索技术,而排序是其中一个十分重要的环节。截至目前,已经提出了百余种用于构建排序函数的特征,例如:基于内容的特征,如TFIDF、BM25;基于链接的特征,如PageRank、HITS;基于点击数据的用户行为的特征等等。综合利用这些特征对文档进行排序,是信息检索的一个核心问题,因此如何利用这些特征构建更为有效的排序函数,成为当今研究的一个热点问题,于是人们提出了排序学习(Learning to Rank)方法,它是利用机器学习的方法,基于特征集合,自动学习最终的排序函数,作为信息检索与机器学习的交叉学科,排序学习方法越来越受到人们的重视。排序学习方法与传统的机器学习方法存在一定的区别:第一,传统的机器学习的学习结果只与作为训练样例的文档以及文档的标注有关,而排序学习方法从查询的角度,文档的标注结果是由文档与查询之间的相关度决定的。第二,在用户使用搜索引擎时,绝大多数用户只访问第一页的结果,并且文档排名越靠前,越可能受到用户的关注,因此对文档绝对

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60736044,No.60903107);高等学校博士学科点专项科研基金(No.20090002120005)。
作者简介:花贵春(1983—),男,博士研究生,主要研究方向为信息检索,机器学习;张敏(1977—),女,博士,副教授;刘奕群(1981—),男,博士,讲师;马少平(1961—),男,教授,博士生导师;茹立云(1979—),男,博士研究生。E-mail:huaguichun@gmail.com
收稿日期:2010-04-15;修回日期:2011-01-25

的评分数值本身是不重要的，仅给出文档的分类也是不够的，而文档之间的顺序更为重要。

现在排序学习方法主要的研究方向可以归结为三类[1]：(1)提出一种排序学习算法，用于构建更有效率或者更有效果的排序函数。(2)构建一种优化函数，能够指导排序学习算法学得一个更优的排序函数。(3)设计新的特征，能够更好地描述某一数据集，从而用于构建更为有效的排序函数。然而，除此之外，从传统机器学习的研究方法中可以看出，特征以及特征的组合方式对分类等方法的性能影响很大，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特征[2]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本文的研究重点。

提出先利用特征重组与选择方法对构成排序函数的特征进行分析，然后利用排序学习方法学习最终的排序函数的模型框架，基于这一框架，提出四种不同的特征处理算法，并利用两种不同类型的排序学习算法进行对比实验。实验结果表明，从整体上而言，排序学习算法基于特征处理后的特征集合，可以学习得到更为有效的排序函数。这一工作是对排序学习方法在特征分析这一研究领域的一种有效的尝试。

2　相关工作

有两方面的研究和本文工作很相关：排序学习方法和特征重组与选择算法。

排序学习方法是在信息检索问题中，利用机器学习的方法，基于特征集合，自动学习得到用于检索的最终的排序函数。排序学习算法需要的数据是由三部分构成的：查询、与该查询相对应的文档的特征序列，以及由人工进行标注的查询与文档之间的相关度。已有的排序学习算法可以根据训练样例的不同分为三类：Pointwise、Pairwise和Listwise方法。Pointwise方法，如Pranking with ranking算法[3]，采用一个查询对应的单一文档作为训练样例，而不考虑此文档与该查询对应的其他文档之间的关系；Pairwise方法，如Ranking SVM算法[4-5]，采用查询对应的文档对(Document Pairs)作为训练样例；而Listwise方法，如ListNet算法[6]，采用查询对应的文档序列(Document Lists)作为训练样例。已有的研究[7-8]表明，一般来说Pairwise和Listwise方法优于Pointwise方法，所以本文中采用两种经典的算法：Pairwise方法的Ranking SVM算法(以后简写为RankSVM)和Listwise方法的ListNet算法进行比较研究。但是这些方法主要关注于算法本身，如学习方法、优化函数等，并不对排序特征进行特征的处理，本文正是对面向排序学习的特征分析的研究。

特征重组方法，是将特征由原特征空间映射到新的特征空间，如PCA(Principal Component Analysis)[9]，将特征重组为线性不相交的主元，在本文中，基于PCA的特征重组方法属于这一类别。特征选择算法按照它与后续学习方法(如分类、回归等问题应用的方法)的结合方式，可以分为过滤式(Filter)、封装式(Wrapper)和嵌入式(Embedded)[10]三种。过滤式特征选择算法是与具体的学习方法无关，可以基于数据集，直接获得特征子集，例如文献[11-12]，在分析特征权重与相互之间相关程度的基础上，对特征进行选择，在本文中，基于MAP(详见4.1节)的特征选择方法属于这一类别；封装式特征选择算法[13]是利用后续学习方法，但不涉及学习方法的内部逻辑，例如常见的前向选择算法(Forward Selection)和后向消去算法(Backward Elimination)，在本文中，采用前向选择算法；嵌入式特征选择算法的算法本身是学习方法的一个组成部分，例如文献[14]的实验利用决策树对特征进行选择，在本文中，算法隐含的特征选择属于这一类别。

3　模型框架，特征重组、选择算法，和排序学习算法

3.1　模型框架

根据传统信息检索框架的线下和线上的工作模式，模型框架的具体设计如图1所示。线下工作：首先基于排序学习方法对应的训练集、验证集和测试集，利用特征重组或者选择算法对特征进行重组或者选择，得到新的特征集合，排序学习方法基于特征集合学习得到最终的排序函数。线上工作：对于一个新来的查询，利用排序学习方法得到的排序模型，对文档索引库中相应的文档进行排序，并将结果文档列表返回给用户。

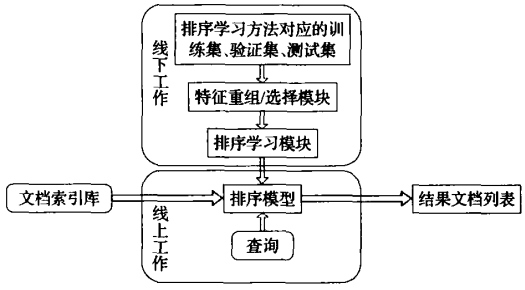


图1　模型框架图

其中：排序学习方法对应的训练集、验证集和测试集的组成详见4.1节，特征重组/选择模块对应的特征重组和选择算法详见3.1节，排序学习模块对应的排序学习算法详见3.2节。

3.2　特征重组和选择算法

基于3.1节介绍的模型框架，面向排序学习方法，提出了一种特征重组算法和三种特征选择算法：基于主成分分析的特征重组算法、基于MAP的特征选择算法、前向选择算法和排序学习算法隐含的特征选择算法。算法的流程如下所示。

算法1　基于主成分分析的特征重组算法

(1)利用训练集对应的特征矩阵，进行主成分分析。

(2)利用得到的前K个主元对原训练集、验证集和测试集进行处理，得到由新的特征构成的训练集、验证集和测试集。

算法2　基于MAP的特征选择算法

(1)将原训练集F_1中的N个特征，每一个特征值取相反数，并归一化后得到N个新的特征，利用这$2N$个特征构建新的训练集F_2。

(2)将F_2中的特征作为排序函数计算对训练集排序的MAP值，将它作为这个特征的权重。

(3)取前$K+m$个权重最大的特征，保证它们在F_1中对应K个不重复的特征。

(4)利用这K个特征对应的特征子集构建新的训练集、验证集和测试集。

算法3 AlgBase前向选择算法

(1)初始的特征子集为 $F_0=\Phi$。

(2)For $i=0,1,\cdots,K-1$

①将每个不在 F_i 中的特征分别加入特征子集，并执行AlgBase算法的流程得出测试集5个集合上的平均MAP值。

②在多个平均MAP值中选出最大值对应的特征子集，记为 F_{i+1}。

(3)利用 F_K 对应的特征子集构建新的训练集、验证集和测试集。

算法4 排序学习算法隐含的特征选择算法

(1)基于训练集和验证集，利用排序学习算法AlgBase获得线性排序函数RF。

(2)将RF中特征系数的绝对值作为特征的权重，取前 K 个权重最大的特征构建新的训练集、验证集和测试集。

根据算法与后续的学习算法之间的关系对算法进行分类，三种特征选择算法的对比如表1所示。基于MAP的特征选择算法不涉及后续的学习算法，因此属于过滤式特征选择算法；前向选择算法涉及后续学习算法，但是不涉及学习算法的内部逻辑，因此属于封装式特征选择算法；排序学习算法隐含的特征选择算法，涉及后续学习算法的内部逻辑，因此属于嵌入式特征选择算法。

表1 三种特征选择算法对比表

	基于MAP	前向选择	算法隐含
排序学习算法相关	否	是	是
排序学习算法内部逻辑相关	否	否	是
属于分类	过滤式	封装式	嵌入式

3.3 排序学习算法

Ranking SVM算法：Ranking SVM[4-5]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：

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

$$\forall(d_k^i,d_k^j)\in d_k\times d_k:\omega\Phi(q_k,d_k^i)>\omega\Phi(q_k,d_k^i)+1-\varepsilon_{i,j,k}$$

在这里，ω 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，$\Phi(q_k,d_k^i)$ 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。

ListNet算法：ListNet[6]定义了一个置换概率：

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

其中 π 是作用于 $n(q_i)$ 个文档的置换，$\Phi(.)$ 是一个递增、恒正的函数，S 为一个评分函数，计算文档与查询相关度的评分，$S_{\pi(j)}$ 表示在置换 π 中第 j 个位置的文档的评分。基于这一置换概率，以交叉熵的形式定义损失函数，也就是优化目标，公式如下：

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

4 实验

4.1 实验设置

实验数据：实验数据集为MSRA提供的LETOR数据集[15]，它作为排序学习领域的一个标准测试集合，广泛地被研究学者采用，实验中采用的是最新的一个版本：LETOR4.0，发布于2009年7月份。LETOR4.0采用Gov2的文档集合，有25 205 179个网页，采用TREC 2007和TREC 2008中Million Query Track的查询，查询数目分别为1 692个和784个，无论从文档数目，还是查询数目，LETOR4.0都是排序学习领域，乃至信息检索领域中规模最大的数据集合之一。每个查询、文档对的相关度标注，分为三级，分别是非常相关、相关和不相关(分别为2,1,0)；抽取基于内容、链接等46个特征。实验中采用的是LETOR4.0中的MQ2007和MQ2008两个数据集。

实验采用五份交叉验证的方式，数据被平分为五份，三份用于训练，一份用于验证，一份用于测试。实验中给出的实验结果，是五个测试集上的平均值，如：MAP的值，是五个测试集上MAP的平均值。

实验评测函数：在信息检索领域中，MAP和NDCG@n广泛被研究学者与商用搜索引擎公司用于对排序函数的评价，因此实验中采用MAP和NDCG@n两种评测函数，用于评测排序函数的性能。

MAP是信息检索领域最常用的评测函数之一，是综合考虑准确率与召回率的一个评测函数，它的定义为：$MAP=\frac{1}{m}\sum_{j=1}^{m}AP_j=\frac{1}{m}\sum_{j=1}^{m}\frac{1}{R_j}(\sum_{i=1}^{k}rel_j(d_i)\cdot(P@i)_j)$，其中：$m$ 是查询的总数，R_j 是指第 j 个查询的相关文档的数目，k 是指系统为这个查询返回的结果中文档的总数，$rel_j(d_i)$ 和 $(P@i)_j$ 分别是相对于第 j 个查询，$rel(d_i)$ 和 $P@i$ 的值。$rel(d_i)$ 定义为：

$$rel(d_i)=\begin{cases}1,\text{如果}d_i\text{和查询是相关的}\\0,\text{其他情况}\end{cases}$$

$P@i$ 是前 i 个结果中的准确率，它定义为：$P@i=\frac{1}{i}\sum_{k=1}^{i}rel(d_k)$。

在返回的结果序列中，综合考虑查询与文档之间的相关度和文档在排序中所处的位置对排序效果的影响，人们进一步提出NDCG@n[16-17]，它广泛用于网络搜索与其他相关的任务中，它的定义为：$NDCG@n=\frac{1}{Z_n}DCG@n=\frac{1}{Z_n}\sum_{i=1}^{n}\frac{2^{rel(d_i)}-1}{\log(1+i)}$，这里 Z_n 是在结果序列是最优的情况下的DCG@n的值，在论文中 $n=1,2,\cdots,10$。

4.2 实验结果及结论

排序函数的命名方式定义为SetName_AlgBase_FS(其中_FS是可选的)，表示在数据集SetName上，基于特征重组或者选择算法FS得到新的特征集合(没有_FS，则基于所有的特征对应的特征集合)，利用排序学习算法AlgBase学习得到的排序函数。其中：

(1)SetName可以取：①MQ2007，表示MQ2007数据集；②MQ2008，表示MQ2008数据集。

(2)AlgBase可以取：①RankSVM，表示RankSVM算法；②ListNet，表示ListNet算法。

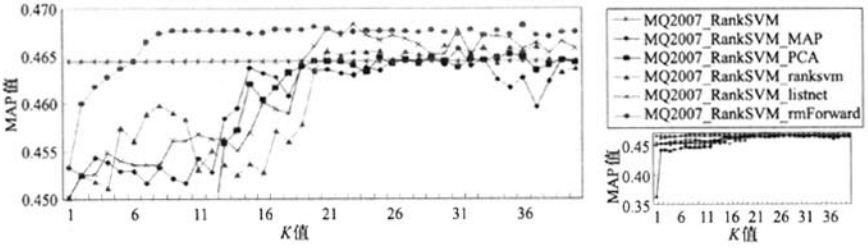
(3)FS可以取：①PCA，表示基于PCA的特征重组算法；②MAP，表示基于MAP的特征选择算法；③rmForward和lnForward，分别表示基于RankSVM前向选择算法和基于ListNet前

向选择算法；④ranksvm和listnet，分别表示RankSVM算法隐含的特征算法和ListNet算法隐含的特征选择算法。

为了确定合适的特征子集的大小，基于四种特征重组或者选择算法得到的特征集合，分别利用RankSVM和ListNet算法学习得到排序函数，随着特征集合的特征数K的不同，各个排序函数在测试集上的MAP值的变化如图2和图3所示。这里图2中的MQ2007_RankSVM和图3中的MQ2007_ListNet分别是基于所有的特征对应的特征集合得到的排序函数，它们作为其他的基于特征集合得到的排序函数的对比实验。

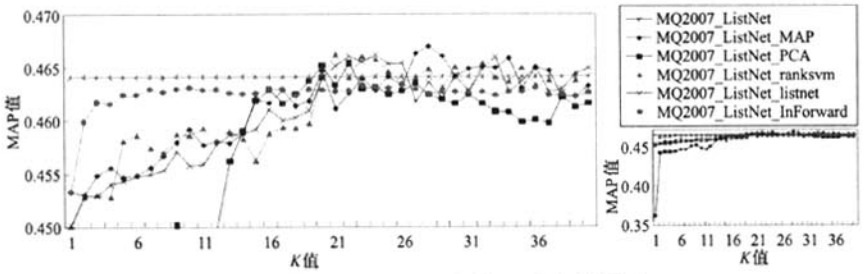
可以看出，大多数的排序函数在特征集合大小K为21到30之间，都趋向于稳定，除了前向选择算法，它在K为9到31之间，趋向于稳定，在这一稳定的区间，MAP值都达到了最高值或邻近最高值，将所有特征重组或者选择算法的特征集合的大小固定为23。则基于RankSVM算法和ListNet算法的NDCG值如图4所示，因为绝大多数的用户只是浏览第一页的返回结果，所以只是给出NDCG@1~10，针对NDCG@n的值，对经过特征重组或者选择的排序学习算法和原始算法做成对T检验，结果如表2所示。

从图4和表2的结果中可以看出：(1)基于PCA的特征重组算法，利用RankSVM算法和ListNet算法，基于特征重组后



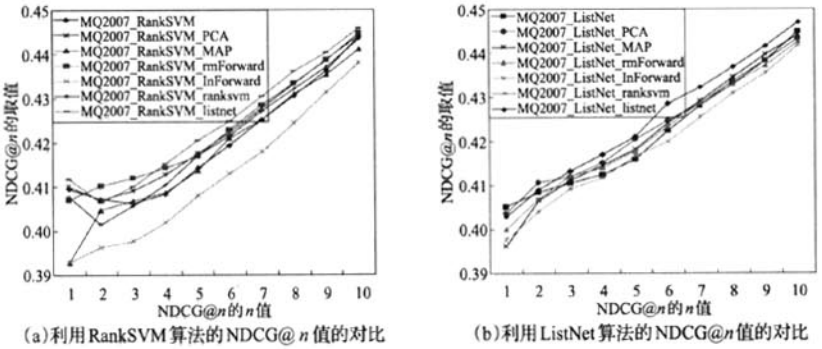
注：右下方图和左侧图只是区别于横纵坐标的取值范围。

图2　RankSVM算法在所有特征对应的集合以及利用各种特征重组或者选择算法得到的特征集合上的对比实验



注：右下方图和左侧图只是区别于横纵坐标的取值范围。

图3　ListNet算法在所有特征对应的集合以及利用各种特征重组或者选择算法得到的特征集合上的对比实验



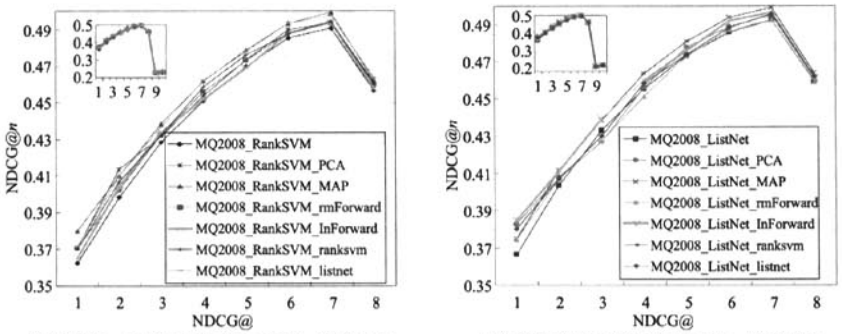
(a)利用RankSVM算法的NDCG@n值的对比　(b)利用ListNet算法的NDCG@n值的对比

图4　RankSVM算法与ListNet算法的对比实验图

表2　对NDCG结果的成对T检验

与MQ2007_RankSVM进行成对T检验的排序函数	T检验的α值	与MQ2007_ListNet进行成对T检验的排序函数	T检验的α值
MQ2007_RankSVM_PCA	0.000 70	MQ2007_ListNet_PCA	0.047 2
MQ2007_RankSVM_MAP	0.126 60	MQ2007_ListNet_MAP	0.463 9
MQ2007_RankSVM_rmForward	0.002 50	MQ2007_ListNet_rmForward	0.297 1
MQ2007_RankSVM_lnForward	0.000 03	MQ2007_ListNet_lnForward	0.002 1
MQ2007_RankSVM_ranksvm	0.363 70	MQ2007_ListNet_ranksvm	0.001 7
MQ2007_RankSVM_listnet	0.000 20	MQ2007_ListNet_listnet	0.001 7

注：$\alpha<0.01$为非常显著；$0.01<\alpha<0.05$为显著；$\alpha>0.05$为不显著。



(a)利用RankSVM算法的NDCG@n值的对比 (b)利用ListNet算法的NDCG@n值的对比

注:左上角的小图与大图的区别只是横纵坐标的取值范围。

图5 RankSVM算法与ListNet算法的对比实验

表3 对NDCG结果的成对T检验

与MQ2008_RankSVM进行成对T检验的排序函数	T检验的α值	与MQ2008_ListNet进行成对T检验的排序函数	T检验的α值
MQ2008_RankSVM_MAP	0.000 060	MQ2008_ListNet_MAP	0.000 008
MQ2008_RankSVM_PCA	0.001 400	MQ2008_ListNet_PCA	0.059 700
MQ2008_RankSVM_ranksvm	0.000 200	MQ2008_ListNet_ranksvm	0.016 000
MQ2008_RankSVM_listnet	0.000 003	MQ2008_ListNet_listnet	0.015 900
MQ2008_RankSVM_rmForward	0.000 100	MQ2008_ListNet_rmForward	0.308 700
MQ2008_RankSVM_lnForward	0.012 500	MQ2008_ListNet_lnForward	0.003 300

的特征集合得到的排序函数都优于原始的排序函数,并且利用RankSVM算法的提升非常显著(0.000 7<0.01),利用ListNet算法的提升显著(0.01<0.047 2<0.05)。(2)基于MAP的特征选择算法,利用RankSVM算法和ListNet算法,基于特征选择后的特征子集得到的排序函数与原始的排序函数是可比的(0.126 6>0.05,0.463 9>0.05)。(3)前向选择算法,利用RankSVM和ListNet算法,原始算法都显著优于基于ListNet前向选择算法后的特征子集得到的排序函数(0.000 3<0.01,0.002 1<0.01);而基于RankSVM前向选择算法后的特征子集,利用RankSVM算法得到的排序函数显著优于原始的排序函数(0.002 5<0.01),利用ListNet算法得到的排序函数与原始的排序函数是可比的(0.297 1>0.05)。(4)ListNet算法隐含的特征选择,利用RankSVM算法和ListNet算法,基于特征选择后的特征子集得到的排序函数都显著优于原始的排序函数(0.000 2<0.01,0.001 7<0.01);RankSVM算法隐含的特征选择,基于特征选择后的特征子集,利用RankSVM算法得到的排序函数与原始的排序函数差别不显著(0.363 7>0.05),而利用ListNet算法得到的排序函数却比原始排序函数有显著的提升(0.001 7<0.01)。

由此可以得出结论:(1)在所有特征选择方法中,基于ListNet算法隐含的特征选择后的特征子集,无论利用RankSVM算法还是利用ListNet算法得到的排序函数相比于同系列其他的排序函数,性能都是最优的,也就是说此特征选择方法对排序函数的性能提升最大。(2)在所有的特征选择算法下,基于ListNet前向选择算法后的特征子集,无论利用RankSVM算法还是利用ListNet算法,原始的排序函数都显著优于特征选择后得到的排序函数,这种特征选择算法也是唯一一个显著降低排序函数性能的特征选择方法,其他的特征选择方法得到的排序函数至少与原始的排序函数是可比的。(3)基于PCA的特征重组算法可以提高排序函数的性能,而且较稳定,虽然对排序函数性能的提高不是最优的,但是它的重组过程不涉及排序函数以及评测函数,应用最为方便。

基于MQ2007下经过特征选择算法得到的特征集合以及经过主成分分析方法得到的特征向量,直接用于MQ2008,特征子集的大小也设为K=23,实验结果以及成对T检验结果如图5和表3所示。可以看出,利用RankSVM算法和ListNet算法,原始的排序函数性能几乎都是最差的。基于特征重组或者选择后得到的特征集合,利用排序学习模型学得的排序函数,显著,或非常显著优于原始排序函数,除MQ2008_ListNet_PCA和MQ2008_ListNet_rmForward两个排序函数与原始的排序函数MQ2008_ListNet是可比的。和MQ2007数据集中的结论相似,ListNet算法隐含的特征选择算法对排序函数性能的提升几乎都是最显著的,除了利用ListNet算法,基于MAP的特征选择后的特征子集得到的排序函数MQ2008_ListNet_MAP的性能优于MQ2008_ListNet_listnet。

4.3 特征子集的分析与结论

除了基于主成分分析的特征重组算法,其他三种特征选择算法对应的五个特征子集,当子集规模K=23时,它们之间的重叠率如表4所示。可以看出,这三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。

表4 五个特征子集之间的重叠率 (%)

	MAP	ranksvm	listnet	rmForward	lnForward
MAP	100.00	65.22	56.52	47.83	43.48
ranksvm	65.22	100.00	56.52	34.78	34.78
listnet	56.52	56.52	100.00	73.91	47.83
rmForward	47.83	34.78	73.91	100.00	52.17
lnForward	43.48	34.78	47.83	52.17	100.00

四种特征重组或者选择算法都能在一定程度上提高检索

性能,与此同时,他们有着不同的特点。基于主成分分析的特征重组算法,可以针对数据集的特点对特征进行重组,构建新的特征,不需要借助于后续的排序学习算法;基于MAP的特征选择算法最为简单,也不需要借助后续的学习算法,但是从实验结果可以看出它不是很稳定;前向选择算法与排序学习算法有很大的关系,作用不明显,甚至会降低排序函数的性能;排序学习算法隐含的特征选择,也与排序学习算法有很大的关系,在两个数据集上都表现很稳定。在所有的方法中,算法隐含的特征选择对性能的提升最为明显,尤其是基于ListNet算法隐含的特征选择,对排序函数性能的提升最为显著。

5 总结与未来工作

在本文中,使用四种特征重组或者选择算法对构建排序函数的特征进行重组或选择,并用排序学习算法Ranking SVM和ListNet算法进行验证,实验结果显示,排序学习算法在经过特征重组或者选择后的特征集合上学习排序函数,可以显著提高检索性能。

本文主要的贡献是:利用特征重组和选择的方法,对构建排序函数的排序特征进行分析,对四种特征重组或者选择算法进行多方面的比较,并利用Pairwise和Listwise的排序学习算法进行验证,实验结果显示经过特征重组或者选择后,排序函数的性能确实得到了显著的提高。

未来的工作可以从以下两个方面继续进行:(1)考察排序学习方法与特征重组、选择方法的关系,为不同的排序学习算法设计更为契合的特征重组、选择算法,进而学习性能更优的排序函数。(2)区分查询的类别,并进一步区分不同的特征在每个类别中的适用性,为不同的类别进行特征的重组、选择。

参考文献:

[1] Duh K,Kirchhoff K.Learning to rank with partially-labeled data[C]//SIGIR 2008,2008:251-258.

[2] Robertson S E.Overview of the okapi projects[J].Journal of Documentation,1997,53(1):3-7.

[3] Crammer K,Singer Y.Pranking with ranking[C]//NIPS 2002,2002.

[4] Herbrich R,Graepel T,Obermayer K.Large margin rank boundaries for ordinal regression[C]//Advances in Large Margin Classifiers,2000:115-132.

[5] Joachims T.Optimizing search engines using clickthrough data[C]//KDD 2002,2002:133-142.

[6] Cao Z,Qin T,Liu T,et al.Learning to rank:from pairwise approach to listwise approach[C]//ICML 2007,2007,227:129-136.

[7] Zhang M,Kuang D,Hua G C,et al.Is learning to rank effective for Web search[C]//SIGIR 2009 Workshop:Learning to Rank for Information Retrieval,2009.

[8] Liu T.Learning to rank for information retrieval[J].Foundation and Trends on Information Retrieval.[S.l.]:Now Publishers,2009,3(3):225-331.

[9] Jolliffe I T.Principal component analysis[M]//2nd ed.Series:Springer Series in Statistics.NY:Springer,2002.

[10] Blum A,Langley P.Selection of relevant features and examples in machine learning[J].Artificial Intelligence,1997,97:245-271.

[11] Cover T M.The best two independent measurements are not the two best[J].IEEE Transactions on Systems,Man,and Cybernetics,1974,4:116-117.

[12] Geng X,Liu T,Qin T,et al.Feature selection for ranking[C]//SIGIR 2007,2007:407-414.

[13] John G H,Kohavi R,Pfleger K.Irrelevant features and the subset selection problem[C]//ICML 1994,1994:121-129.

[14] Langley P,Sage S.Scaling to domains with many irrelevant features[M]//Greiner R.Computational Learning Theory and Natural Learning Systems.Cambridge,MA:MIT Press,1997,4.

[15] Liu T Y,Xu J,Qin T,et al.LETOR:benchmark dataset for research on learning to rank for information retrieval[C]//SIGIR 2007 Workshop on Learning to Rank for Information Retrieval,2007.

[16] Jarvelin K,Kekalainen J.IR evaluation methods for retrieving highly relevant documents[C]//SIGIR 2000,2000:41-48.

[17] Jarvelin K,Kekalainen J.Cumulated gain-based evaluation of IR techniques[J].ACM Transactions on Information Systems,2002,20(4):422-446.

(上接94页)

[5] Khurana S.Performance evaluation of distributed co-ordination function for IEEE 802.11 Wireless LAN protocol in presence of mobile and hidden terminals[C]//Proceedings of the 7th International Symposium on Modeling,Analysis and Simulation of Computer and Telecommunication Systems,1999:40-47.

[6] 饶国威,杜明辉,万泉.无线局域网中移动隐藏终端的自适应算法[J].计算机工程,2007,33(4):105-107.

[7] Kim J,Kim S,Choi S,et al.CARA:collision-aware rate adaptation for IEEE 802.11 WLANs[C]//25th IEEE International Conference on Computer Communications Proceedings,2006:1-11.

[8] Chen Xi,Qiao Daji.Probabilistic-based rate adaptation for IEEE 802.11 WLANs[C]//IEEE Global Telecommunications Conference,2007:4904-4908.

[9] Kim Y.A novel hidden station detection mechanism in IEEE 802.11 WLAN[J].IEEE Communications Letters,2006,10(8):608-610.

[10] 严少虎,卓永宁,吴诗其,等.802.11 DCF中优化吞吐率的RTS门限调整算法[J].系统工程与电子技术,2004,26(9):1172-1175.

[11] NS2使用说明手册[EB/OL].http://140.116.72.80/~smallko/ns2/ns2.htm.

[12] NS-网络仿真软件[EB/OL].http://www.baisi.net.

面向排序学习的特征分析的研究

作者： 花贵春， 张敏， 邝达， 刘奕群， 马少平， 茹立云， HUA Guichun， ZHANG Min， KUANG Da， LIU Yiqun， MA Shaoping， RU Liyun

作者单位： 清华大学计算机系,智能技术与系统国家重点实验室,清华信息科学与技术国家实验室(筹),北京100084

刊名： 计算机工程与应用 ISTIC PKU

英文刊名： COMPUTER ENGINEERING AND APPLICATIONS

年，卷(期)： 2011,47(17)

参考文献(17条)

1. Jarvelin K;Kekalainen J Cumulated gain-based evaluation of IR techniques 2002(04)

2. Jarvelin K;Kekalainen J IR evaluation methods for retrieving highly relevant documents 2000

3. Liu T Y;Xu J;Qin T LETOR:benchmark dataset for research on learning to rank for information retrieval 2007

4. Langley P;Sage S Scaling to domains with many irrelevant features 1997

5. John G H;Kohavi R;Pfleger K Irrelevant features and the subset selection problem 1994

6. Geng X;Liu T;Qin T Feature selection for ranking 2007

7. Cover T M The best two independent measurements are not the two best 1974

8. Blum A;Langley P Selection of relevant features and examples in machine learning[外文期刊] 1997

9. Jolliffe I T Principal component analysis 2002

10. Liu T Learning to rank for information retrieval 2009(03)

11. Zhang M;Kuang D;Hua G C Is learning to rank effective for Web search 2009

12. Cao Z;Qin T;Liu T Learning to rank:from pairwise approach to listwise approach 2007

13. Joachims T Optimizing search engines using clickthrough data 2002

14. Herbrich R;Graepel T;Obermayer K Large margin rank boundaries for ordinal regression 2000

15. Crammer K;Singer Y Pranking with ranking 2002

16. Robertson S E Overview of the okapi projects 1997(01)

17. Dub K;Kirchhoff K Learning to rank with partially-labeled data 2008