一种应用奇异值分解的 RankBoost 排序学习方法

林原　林鸿飞　苏绥

大连理工大学计算机科学与工程系，大连，116024

E-mail: yuanlin@student.dlut.edu.cn

摘要：Learning to rank(排序学习)已经成为当今信息检索领域研究和讨论热点。它运用信息检索和机器学习领域的方法，结合相关性判断条件提供与查询更加相关的信息。当前的排序算法主要集中于相关性标注数据的使用，本文通过对相关性标注数据集以及非标注数据集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中，引入非标注数据信息。通过实验对比了新特征集加入前后的 RankBoost 算法排序性能差异，以及新特征集合的大小对于排序结果的影响。实验表明，加入奇异值分解后选取的特征集合，有助于排序效果的提高。

关键词：信息检索；learning to rank；机器学习；奇异值分解；RankBoost

Learning to Rank Using SVD and RankBoost

Yuan Lin,Hongfei Lin,Sui Su

Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024

E-mail: yuanlin@student.dlut.edu.cn

Abstract: In today's field of information retrieval, learning to rank has become a hot issue for experts and scholars to study and explore. It's using the methods of information retrieval and machine learning with the relevance judgements, to provide more relevant information for queries. The previous ranking algorithm focused on the relevance labeled data , this paper uses SVD on the collection made of labeled data set and unlabeled data set to extract new feature vectors which are added to training sets for RankBoost, so as to import unlabeled data information. We experimentally compared the performance of RankBoost using the training set with new feature vectors and not, as well as the impact of the size of the new features to the result of sort. The experimental results show that it is useful to improve the result of sort by adding the SVD feature vectors into training set.

Key words: Information Retrieval; Learning to rank; Machine learning; SVD; RankBoost

1　引言

Learning to rank 是一个信息检索与机器学习相结合的研究领域，目的是结合相关性判断条件从训练数据中学习 Rank 函数，通过 Rank 函数对文档的相关性进行排序。排序学习研究的核心问题是如何构造一个函数或模型反映文档对于查询的相关度。排序学习在信息检索中所定义的任务内容如下：给定文档的训练集合 D，其中每个文档表示为<q,d,r>的三元组的形式，q 为查询；d 为文档特征集合$\{f_1,f_2,\ldots,f_n\}$,这里的文档特征特是查询和文档的复合特征；r 为文档与查询的相关性判断条件取值一般为{0，1}，0 代表不相关，1 代表相关。测试集合用 T 表示，也以三元组<q,d,?>形式表示，只有查询和文档特征集合两个元素，而文档的相关性未知。排序模型就是由

基金项目：国家自然科学基金资助项目（编号：60373095，60673039）和国家 863 高科技计划资助项目（编号：2006AA01Z151）。

文档训练集合三元组训练得到，用于预测测试集文档相关性分数，进而计算文档相关性排名。

尽管有许多基于标注数据集的算法和模型不断的被引入到排序学习问题当中，但是非标注数据集运用方面的探讨和研究还是略有不足。鉴于这种情况，本文通过对标注数据和非标注数据进行 SVD 联合分解，引入非标注数据信息，借此提高 RankBoost 排序算法的性能，进而提高排序结果的 Map 值。

本文余下部分组织结构如下：第二部分着重介绍下当今排序学习方法相关工作。第三部分提出了的通过奇异值分解将非标注数据信息引入排序学习的算法和思想。第四部分对于实验结果进行介绍和分析。第五部分为本文的结论部分，并对今后的工作进行展望。

2 相关工作

为了获取排序函数，很多模型和理论都先后被引入到排序学习领域当中,Freund 率先将机器学习中的 AdaBoost 方法引入到电影参考相关性排序当中，提出了 RankBoost 方法[1]用于排序学习。Herbrich 则把经典的 SVM 理论引用于排序学习当中，提出了 RankSVM 方法[2]。清华大学 Zhe cao 等人基于神经网络和梯度下降优化方法的提出了 Listnet 方法[3]，对排序结果的提高有着很大作用。Guiver 等人则成功地将高斯过程[4]应用于排序任务,也获得较好的效果。Adriano Veloso 以关联规则的相关理论和算法为基础，从特征集合中提取特征的区间关联条件，并以此建立规则对文档的相关性进行预测[5]。

基于标注数据的信息检索和机器学习的理论模型在排序学习任务中取得了巨大的成功，而非标注数据的方面的探讨则显得力度不够。尽管如此，仍有一些学者试图从非标注相关性的数据集入手，提高排序的效果。Massih-Reza Amini,考虑用寻找与标注文档相似度最大的非标注文档，赋之以相同的相关性条件加入训练集文档中，实验表明这种方法可以提高排序效果[6]。Kevin Duh 则应用基于核的主成分分析法对测试集合进行主成分模式提取，并采用这种模式对训练集合提取新的特征集合，从而隐式地将测试集合中的特征加入到训练集合当中，获得了很好的结果[7]。但是该方法只是单独利用测试集合的信息获得特征提取模式，应用该模式在训练集合中进行特征提取，而没有考虑到同时将训练集合中的信息引入到测试集合当中，也有助于模型对文档的相关性预测。如何将测试集和训练集的信息同时引入到彼此当中，这是本文研究的一个问题之一。

本文借鉴前人的研究方法，试图更有效的运用非标注语料进一步提高模型对于文档相关性的预测效果。奇异值分解对于挖掘文档特征之间信息和关系效果更好。本文将训练集和测试集合并，对合并后的集合进行奇异值分解，这样做不但能够引入非标注信息而且可以将二者的特征集合投影到统一的维度空间，并从中提取新的特征向量集合，并结合使用 RankBoost 方法利用该向量集合进行模型训练和相关性预测，用以提高排序效果。

3 SVD 和 RankBoost 算法的应用

SVD 与 PCA 提取的主特征向量有着相似的性质[8]，本文借鉴文献[7]的思想，将非标注数据信息引入排序模型的训练过程当中。与其不同的是本文通过 SVD 对测试集合以及训练集合进行联合分解，同时构造标注数据（训练集）与非标注数据（测试集）的关联关系，将二者的信息同时引入到对方当中。在 RankBoost 的模型训练过程中就可以利用到非标注语料集合的相关信息，对于下一步测试集合的相关性分数预测也是很有帮助的。

同时，SVD 本身是一种主特征提取方法，可以通过该方法隐式地获取更为有效的主特征向

量，加入这样具有复合信息的向量，对于训练迭代中的特征弱学习器的选取以及测试集的相关性分数预测都是十分有意义的。本文所采用的方法整体运行过程，如图 1 所示。

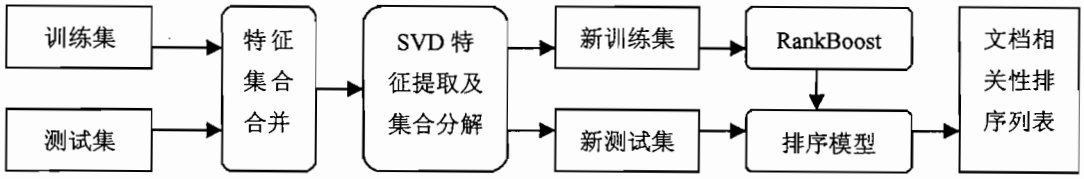

图 1　SVD-RankBoost 排序学习过程

3.1 SVD 特征提取

奇异值分解是一种基于矩阵变换的特征抽取方法。通过对文档—特征项矩阵进行投影分析，得到三个矩阵，第一个矩阵就是文档—主特征矩阵 U，中间矩阵是主特征值对角阵 S，而最后一个矩阵则是主特征值—原特征矩阵 P。同时 SVD 是一种隐式的特征降维方法，通过矩阵变换选取较大的特征值所对应的向量把原始空间降到低维空间，使这个空间则拥有原始特征空间的大部分信息量。这里所选取的新的特征向量就包含在文档—主特征项 U 矩阵中，选取较大的特征值所对应的特征向量加入到训练集合当中。算法思想如下：

输入：训练文档集合 D,测试文档集合 T;

输出：新的训练集 D′,测试集 T′；

1:合并集合 D 与集合 T =>得到总文档集合 DT;

2:SVD(DT) =>得到文档-主特征项矩阵 U，主特征值对角阵 S，主特征-值原特征矩阵 P;

3:选取 N 个特征值对应的 U 中的 N 个特征向量组成主特征向量矩阵 F;

4:合并 DT 中的特征向量集合与 F =>得到新的总特征集合 (DT)′；

5:将 (DT)′ 按原训练集和测试集合的维数将测试集合分解为新的训练集合 D′ 和 T′；

通过上述算法得到了带有非标注测试集信息并且含有原特征集合主要信息的新的训练集合。对训练集合和测试集合进行联合分解的意义在于可以将二者的特征向量同时投影在同维度的特征空间中，会使投影后获取的新向量与原组合矩阵的关系更密切，以此构造他们的关联关系，为下一步排序模型中弱学习器的选取，以及对测试集合文档相关性的预测提供统一的平台。最后本文综合训练时间和排序效果选择适合的特征向量数目加入训练集当中。

3.2　RankBoost 算法

RankBoost[1]是一种 Pairwise 方法，Pairwise 方法是排序学习中一种把排序问题转化为分类问题的常用方法。其首要任务就是把训练集文档以查询为单位根据相关性判断条件做正负样本对。这里设定这个文档对的模式为$<x_0,x_1>$其中 x_0 为不相关文档，而 x_1 为相关文档。RankBoost 是一种为完成排序任务而改进的 AdaBoost 算法。每一个文档对赋予一个权重值，每一次迭代中，会产生一个使顺序正确的文档对数目最多的弱学习器，如果该弱学习器使一个文档对中的两个文档排序错误，那么就增大该文档对的权重，相反则降低该文档对的权重。保留每一步所得的学习器，

并且加以权重累加起来，迭代完成后得到了一个预测文档相关性的线性累加函数。

4 实验结果及分析

4.1 实验数据集

实验的结果是建立在微软亚洲研究院所发布的 LETOR2.0[9]数据集的基础之上的。LETOR2.0 数据集合包括 3 个文档检索数据集：TD2003,TD2004,OHSUMED。前两个数据集来自于 TREC 任务的语料数据集，而 OHSUMED 则来自于医学检索任务的语料。三个数据集都提供了文档的相关性判断条件以便用于结果排序模型的训练。

LETOR2.0 数据集包含大量的文档—查询的复合特征，其中包括 BM25[10]、 HITS[11]、PageRank[12]以及语言模型[13]中提取的特征等。其中 TREC 文档包含有 44 个特征，而 OHSUMED 则包含 25 个特征。其中每一个特征都可以作为对文档相关性进行排序的依据，而排序学习的目的则正是综合这些特征的排序效果得到更好的文档相关性排名列表。

4.2 实验结果

对于实验的结果评价，本文采用的是排序学习的主流评估方法 MAP[14],该方法的主要思想是首先计算每个测试集合中的每个查询的 AP（average precision）值，AP 评价方法的基本思想就是考虑每个相关文档在排名列表的具体位置，并且以此为依据给出每个文档排名的分数。再对这些 AP 值加和求平均值得到了该测试集合的 MAP 值。

LETOR2.0 共有 3 个数据集，每个数据集包含 5 组训练集和测试集合，本文对 TREC 语料中文档集合的 44 个特征构成的矩阵进行 SVD 分解，以分解后的特征向量的信息量的大小为依据[15]，经过反复验证，从中选取 10 个特征向量加入到训练集合当中。而对于包含 25 个特征的 OHSUMED 数据集则选取 5 个特征向量加入训练集合当中。实验给出每组测试集合的用 RankBoost 方法训练的 Baseline(不加入 SVD 主特征向量)，以及加入主特征向量的排序结果，表 1 为两种方法对应于 OHSUMED，TD2003,TD2004 测试集预测结果的 Map 值的列表。其中 OHSUMED 加入了前 5 个 SVD 特征向量,TD2003 和 TD2004 则分别加入了前 10 个 SVD 特征向量。

表 1 LETOR 2.0 数据集的 Map 值

	OHSUMED		TD2003		TD204	
	Baseline	SVD	Baseline	SVD	Baseline	SVD
Fold1	0.3523	0.3560	0.1406	0.1569	0.4539	0.4731
Fold2	0.4747	0.4755	0.3001	0.3100	0.3486	0.3748
Fold3	0.4504	0.4459	0.1693	0.1743	0.3410	0.3614
Fold4	0.5228	0.5176	0.1964	0.2343	0.3516	0.3335
Fold5	0.4710	0.4701	0.1628	0.1575	0.3140	0.3212
Average	0.4542	0.4530	0.1938	0.2066	0.3618	0.3729

4.3 结果分析

表 1 的实验结果表明，加入 SVD 处理得到的主特征集合使绝大部分测试集合的排序结果有

了显著地提高，但是仍有部分测试集合的排序效果不理想，这可能是由于弱学习器的特征阈值选择以及迭代次数的选择过于泛化造成的，同时也可能是由于 RankBoost 的备选特征集合没有进一步优化选择引起的，这是需要今后要继续研究改进的。

通过 SVD 分解把有相关性标注的文档集合和非标注相关性的文档集合联系起来，将其投影到低维空间中，获取复合信息的主特征向量集合，加入原训练集以及测试集特征向量的特征集合中，分别进行排序模型训练，和相关性分数预测。从整体上看，排序效果有了比较明显地提高的。SVD 特征项向量所含原特征集合的信息量的大小往往由该特征所对应的特征值的大小所决定，所以选择特征值较大的特征向量加入训练集合对改善 RankBoost 排序结果的性能帮助会更大。同时，所加入的新特征的数目也影响着结果排序的最终效果，加入较多的 SVD 特征向量无疑会大幅度增加模型训练的时间，相反引入较少的 SVD 特征则会使信息丢失过多，改善效果不明显。所以本文着重对 SVD 特征的数目进行了选择实验，通过对 3 种数据训练集合的加入不同数目的特征向量，得到的测试集合的平均 MAP 值的比较表明，对于 OHSUMED 测试集合，从 5 个集合的平均 Map 角度来看，加入新向量改进效果不明显，基于排序效果和模型训练时间的综合考虑，选择特征值大小在前 5 位的 SVD 特征向量比较合适。而对于 TD2003 和 TD2004 测试集合则选择特征值集合在前 10 位左右的特征值较好，加入特征 1-5 个，MAP 值变化不明显，而 5-10 的区间属于上升阶段，10 到 34 左右略有震荡，总体平缓，34-44 左右缓慢下降，但最终的排序效果仍好于 Baseline。从训练时间上来看，新特征数目的增加，大大的延长了模型的训练时间。最终 OHSUMED 测试集合特征数目选定为 5，TREC 数据集合选择加入的特征数目为 10，在尽量减少训练时间的基础上提高了排序效果。

本文融合 Baseline 模型和加入 SVD 特征向量的模型得出的相关性判断分数，给出新的文档相关性排名列表。该方法的主要思想是借助两个模型的相关性评分，以查询为单位进行标准化，加和后作为新的相关性分数。观察该方法是否有助于排序结果的进一步提高，并与加入 PCA 向量[7]方法的排序结果进行比较，表 2 为上述几种方法在 3 个测试集合中的平均 MAP 值，从中可以看出，虽然 SVD 方法与 PCA 方法都有助于排序效果的提高，排序效果对于不同的训练集还是略有不同的。同时融合排序模型结果的方法对于改善排序效果起到了一定的作用，对两个模型的结果起到了一定的中和作用，如何更好的融合不同方法的排序模型是排序学习研究一种新的思路。

表 2 方法比较

	BaseLine	SVD	Merge	PCA
OHSUMED	0.4543	0.4530	0.4535	0.4455
TD2003	0.1939	0.2066	0.2041	0.3226
TD2004	0.3619	0.3729	0.3686	0.3703

5 结论和展望

实验结果表明，加入 SVD 主特征向量及非标注数据信息，对于排序学习的排序结果的提高是有积极意义，同时该方法仍不能对所有的测试集合结果起到提高的作用，这可能是由于在实验细节上不能准确把握尺度和技巧造成的，该方面的研究将是下一阶段工作的重点。在提取新的特征以及利用非标注语料的同时，仍将从其它角度来提高排序效果，本文所利用的 RankBoost 方法主要是基于正负相关文档对的，然而训练文档中存在着大量相关性相等的文档对，如何利用如此

丰富的信息，进一步从相关性判断条件中挖掘其它重要信息，将成为排序学习的一个重要的方向. 再者，应用其它机器学习的算法和模型于排序学习任务，也将是未来工作的重要方向。总之，排序学习领域涵盖的问题是广泛的，留给了研究人员足够探索和学习的空间。

参 考 文 献

[1] Y.Freund,R.Iyer, R.Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences.Journal of Machine Learning Research, 2003, 4: 933-969.

[2] R. Herbrich, T. Graepel, and K. Obermayer. Support vector learning for ordinal regression. In international Conference on Artificial Neural Networks (ICANN1999),Edinburgh, UK, 1999, 1: 97-102.

[3] Zhe Cao , Tao Qin , Tie-Yan Liu , Ming-Feng Tsai and Hang Li . Learning to Rank: From Pairwise Approach to Listwise Approach. In International conference on Machine learning (ICML2007), Corvalis,Oregon，USA，2007: 129－136.

[4] John Guiver and Edward Snelson . Learning to Rank with SoftRank and Gaussian Processes.In ACM Special Interest Group on Information Retrieval (SIGIR2008), Singapore, 2008: 259-266.

[5] Adriano Veloso, Humberto M. Almeida, Marcos Gonçalves and Wagner Meira Jr . Learning to Rank at Query-Time using Association Rules. In ACM Special Interest Group on Information Retrieval (SIGIR2008), Singapore, 2008: 267-274.

[6] Massih-Reza Amini, Tuong-Vinh Truong and Cyril Goutte . A Boosting Algorithm for Learning Bipartite Ranking Functions with Partially Labeled Data. In ACM Special Interest Group on Information Retrieval (SIGIR2008), Singapore, 2008: 99-106.

[7] Kevin Duh and Katrin Kirchhoff. Learning to Rank with Partially-Labeled Data. In ACM Special Interest Group on Information Retrieval (SIGIR2008), Singapore, 2008: 251-258.

[8] 吴春国，梁艳春，孙延风，周春光，吕英华.关于 SVD 与 PCA 等价性的研究. 计算机学报, 2004, 27(2): 286-288.

[9] T.-Y. Liu, T. Qin, J. Xu, W. Xiong, and H. Li. LETOR: Benchmark dataset for research on learning to rank for information retrieval.SIGIR2007 Workshop on Learning to Rank for IR(LR4IR), ACM SIGIR Forum ,2007,41(2): 58-62.

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

[11] Kleinberg, J. Authoritative sources in a hyperlinked environment.Journal of the ACM.1999, 46(5): 604-622.

[12] Page, L., Brin, S., Motwani, R., and Winograd, T. The PageRank citation ranking: bringing order to the Web, Technical report, Stanford University, 1998.

[13] Zhai, C. and Lafferty, J. A study of smoothing methods for language models applied to Ad Hoc information retrieval. Proceedings of SIGIR, 2001: 334-342.

[14] Kalervo Järvelin and Jaana Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA,2000: 41–48.

[15] 林鸿飞,姚天顺. 基于潜在语义索引的文本浏览机制. 中文信息学报,2000, 14(5): 49-56.