文章编号：1003-0077(2012)05-0053-06

排序学习中数据噪音敏感度分析

牛树梓，程学旗，郭嘉丰

（中国科学院 计算技术研究所，北京 100190）

摘　要：排序学习是当前信息检索领域研究热点之一。为了避免训练集中噪音的影响，当前排序学习算法较多关注鲁棒性。已有的工作发现相同的排序学习方法的性能在不同的数据集上会有截然不同的噪音敏感度。模型改变是导致性能下降的直接原因，而模型又是从训练集学习到的，因此根源在于训练数据的某些特性。该文根据具体排序学习场景分析得出影响噪音敏感度的根本原因在于训练集中文档对分布的结论，并在 LETOR3.0 上的实验验证了这一结论。

关键词：排序学习；数据质量；噪音敏感

中图分类号：TP391　　**文献标识码**：A

Noise Sensitivity in Learning to Rank

NIU Shuzi, CHENG Xueqi, GUO Jiafeng

(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Learning to rank is one of the most attractive areas in information retrieval. Much attention has been paid on the robustness of ranking algorithms to deal with noise which is inevitable in the training set. Previous work observes that ranking performance of the same algorithm showed totally different noise sensitivities. The performance degradation of ranking models boils down to the training set. Thus the underlying reason for different sensitivities lies in some attribute of training data. Experimental results on LETOR3.0 suggest that if the document pairs of the same training set scatter more dispersedly, the model from this training set is less influenced by the error document pairs and the training set is thus less sensitive to noise.

Key words: learning to rank; data quality; noise sensitivity

1　引言

排序学习是当前信息检索领域研究热点之一。随着其逐渐成熟与完善，越来越多的工作专注于排序学习算法的鲁棒性。现实世界中的数据噪音是不可避免的。尤其是在有监督的学习问题中，标注数据的获得的过程中更容易引入噪音。在信息检索领域，对于文档与查询的相关程度的标注也是如此。在单人的标注中由于偏见与主观带来的错误标注是在所难免的；在多个人的标注中，尽管通过投票的方式可以减少犯错误的概率，但仍不能避免[1]。因此对排序学习算法性能的鲁棒性研究是必要的。

越来越多的研究工作[2]表明算法的鲁棒性是一个复杂的现象，它不但与排序学习算法有关，与数据集也有密切联系[3]。文献[2]中描述了 RankSVM、SVM^{MAP} 等算法在一些数据集上具有较好抗噪性，随着噪音水平的增加性能下降不大，而在另外一些数据集上即使噪音水平很小性能也有很大程度的下降。从数据角度出发，这种现象可以理解为数据具有不同的噪音敏感度。本文从 Pairwise 方法和 Listwise 方法的两个具体的学习场景出发，重点研究

收稿日期：2011-09-13　**定稿日期**：2012-04-13

基金项目：国家自然科学基金资助项目(60903139, 60873243, 60933005)；国家 863 计划重点资助项目(2010AA012502, 2010AA012503)

作者简介：牛树梓(1985—)，女，博士研究生，主要研究方向为信息检索；程学旗(1971—)，男，博士，研究员，主要研究方向为信息检索、复杂网络分析与社会计算、网络安全；郭嘉丰(1980—)，男，博士，助研，主要研究方向为信息检索。

导致数据具有不同的噪音敏感度的潜在原因，并希望能够依据实验得到的结论，指导训练集的构建。

2 相关工作

现有关于排序学习的工作大部分关注新算法的提出。由于训练集中噪音是不可避免的，因此对于算法抗噪性的研究不但是一个具有挑战性的任务[4]，而且受到越来越多的工作[5-6]关注。核函数主成分排序算法(KPCRank)[5]是非线性主成分回归方法的扩展，适合于解决属性数据中存在噪音的问题。文献[6]采用两阶段优化策略对任何线性排序模型进行非凸优化，从而使得学习到的模型对噪音不敏感。

仅有少量的工作关注于排序学习问题中训练集的质量。采用恰当的文档选择策略[7-8]可以提高训练集的质量。文献[7]利用四类已有的文档选择策略构建具有不同特性的数据集，重点研究了数据集的相关与不相关文档的比例以及两类文档间的相似性两个特性对排序学习算法性能的影响。文献[8]将文档选择问题转化为最优化问题，提出了一个称为PPC(Pair-wise Preference Consistency)的量作为优化目标，将PPC最大的子集作为训练集。文献[3]等试图通过加入一些辅助信息，如用户的点击日志来检测并修正数据的标注值，用以改善训练集中标注数据的质量。随着外包平台的兴起，以较少的代价获得大量的标注数据已经变得非常容易，但由于标注者的知识背景未知，标注数据的质量不再有保证，一些工作提出通过重复标注[1]或综合多个标注结果[9]来改善标注质量。

文献[10]和[11]试图研究数据集的特性对排序学习算法性能的影响。前者说明排序学习算法的性能随着标注数据准确度的提高而提高，这一点在不同数据集上是一致的。与其不同的是，本文工作关注于性能随数据质量变化在不同数据集上的差异，并重点分析了造成这种差异的根本原因。此外，文献[10]通过实验证明了模拟噪音的方法对这种一致性是没有影响的，因此本文为了简化实验，只采用一种均匀分布来模拟噪音。文献[11]研究了相关性级别在训练集中的分布情况对学习到的排序模型的性能的影响。综合上述工作，本文提出了文档对属性向量的方向分布这一特征，并研究该特征对数据集噪音敏感性的影响。

3 实验环境

3.1 实验数据集

本文采用微软亚洲研究院公布的LETOR3.0[12]的OHSUMED和TD2003数据集。OHSUMED来源于医学检索任务。TD2003来自于TREC任务的语料数据集。两个数据集中的文档都被表示为属性数据向量的形式。数据集的大小以及更详细的信息见表1。

表1 数据集的统计信息

数据集	查询个数	每个查询的平均文档数	属性个数	相关性级别
OHSUMED	106	152	45	{0, 1, 2}
TD2003	50	981	64	{0, 1}

3.2 排序学习算法

Pairwise方法与Listwise方法是目前排序学习问题中比较常用的两类算法。Pairwise学习方法从原始训练集中根据标注的相关性的某种大小关系构建偏序文档对的集合，将排序模型的学习问题转换为这些偏序文档对集合上的二分类问题。当这个二分类问题采用支持向量机的方法解决时就是RankSVM算法。RankSVM[13]旨在寻找一个超平面w使得两类之间的间距最大化。在一个线性可分的空间中(在原空间中线性不可分的点经过核函数的变换到高维空间)，超平面的法方向w可以视为这些文档对的主方向。ListNet[14]是一个典型的Listwise算法。该方法将排序问题建模为概率模型，n个元素构成的排列有O(n!)种可能，每个排列的出现概率不同。当该文档根据标注数据得到的排列概率分布与根据排序函数得到概率分布的交叉熵最小时的w即为ListNet的最优解。

3.3 实验流程及基本实验结果

为了研究噪音敏感度不同的原因，本文分别比较了RankSVM算法与ListNet算法在OHSUMED和TD2003上性能变化曲线，重现了同一排序学习算法在不同数据集上具有不同的噪音敏感度这一现象。这里排序性能的评价指标采用的是MAP和NDCG@10。

假设原始数据集OHSUMED和TD2003中的

标注是干净的，采用模拟噪音的方式研究噪音对不同数据集上算法性能的影响。本文采用文献[3,15-16]中提到的加噪音的方法。设噪音水平为 x，对每个查询中的文档集，按相关性级别分层采样，每级中样本占总体的比例为 x，改变样本数据的相关性级别。改变方法是将其标注均匀的翻转为其他级别。文献[10]通过实验证明了噪音分布对研究算法性能与噪音水平关系影响不大，因此本文采用一种简单均匀分布也是可以说明问题的。

本文考察了同一排序学习算法在不同数据集上噪音水平从 0 到 0.5，步长为 0.05 的变化区间上性能变化情况。对每一个噪音比例，为了避免随机选择带来的偏差，重复加噪音的方法 10 次，这样得到 10 个同样噪音比例的数据集。每个加噪音后的数据集按照原始划分，得到五个大小近似相等子集，用于五折交叉验证。这样该算法在每个数据集上得到的模型的性能(MAP，NDCG@10)由五个不含噪音的 test 集上的平均得到。算法在该噪音比例下的性能由 10 个数据集上的性能评价值进行平均得到。采用上述实验流程，分别得到 RankSVM 和 ListNet 的性能变化曲线，如图 1 所示。

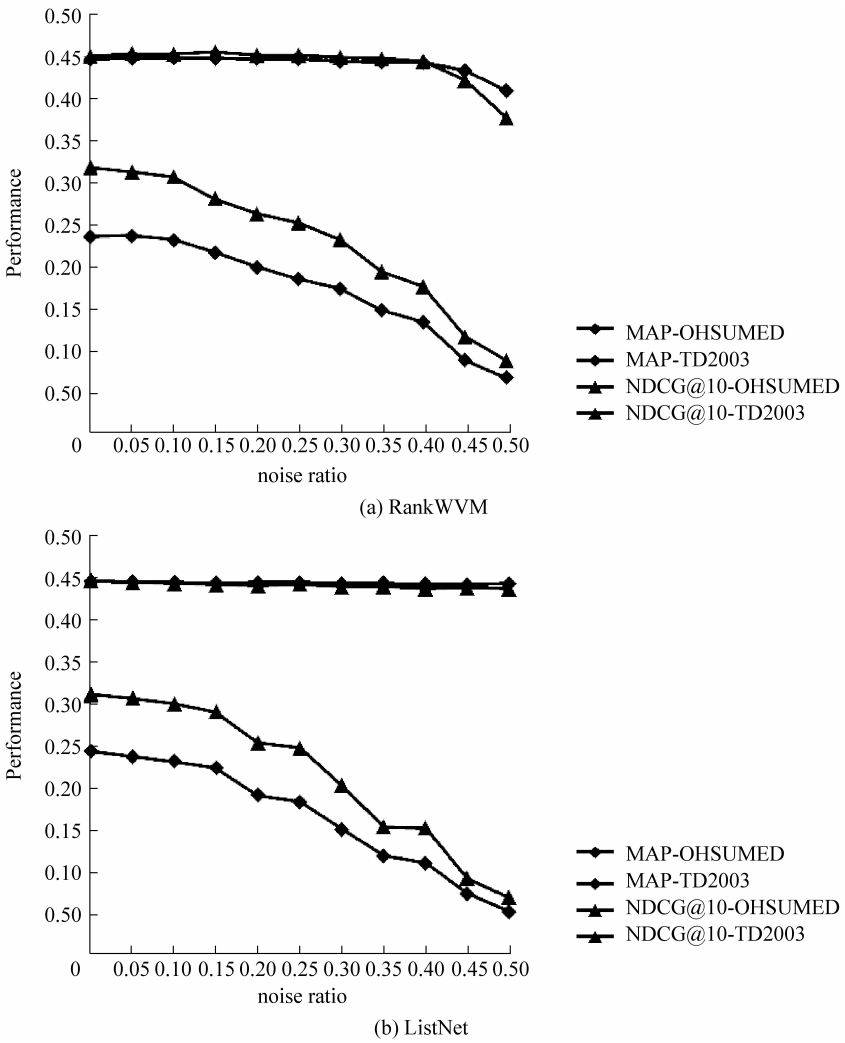

图 1　RankSVM 与 ListNet 在不同数据集性能随噪音加入的变化情况

从实验结果中不难看出无论采用哪种评价指标，也无论采用哪种学习算法，在噪音水平达到 0.4 之前在 OHSUMED 上性能随噪音增加降幅不明显，而在 TD2003 上的性能在噪音水平刚到 0.1 时便开始急剧下降。如图 1(a)所示，在 OHSUMED 上，噪音水平为 0.3 时的 NDCG@10 较噪音水平为 0.25 时的降低了 0.53%，在 TD2003 上较噪音水平为 0.25 时降低了 7.9%。这种现象表明同一排序学习算法在不同的数据集上噪音敏感度不同。这种明显的差异促使对其产生原因的研究。

4 不同数据集上噪音敏感度差异的根本原因

直观来讲,影响排序学习算法在同一测试集上性能差异的最直接原因是该算法在训练集上学习到的模型的好坏。本文首先通过验证模型和性能的变化曲线是否吻合来证明这一点。究其根本,排序学习算法的模型来源于训练集。受 Pairwise 方法的启发,本文给出了偏序文档对方向分布散度这一指标。该指标与训练集相关,因此从本质上解释了导致噪音敏感度不同的原因。最后通过进一步的比较实验验证其正确性。

4.1 直接原因:模型因素

为了验证模型与性能变化的一致性,需要对模型 w 量化。简单起见,可以采用与同一向量的余弦相似度[17]作为该量化指标,并将该不变的向量简称为基向量。理论上讲,基向量是可以随便选择的,但为了使其物理意义更明显,这里采用从不含噪音的训练集上得到的模型 w_0 作为基向量。因此每个模型 w 的量化指标值定义为其与 w_0 的余弦相似度。

LETOR 中每个数据集上的性能评价来源于五折交叉验证,因此模型性能是五个训练集上平均作用的结果,此处对模型的量化值也应采用五个对应训练集上的平均。RankSVM 较 ListNet 特殊之处在于,模型中存在一个正则化参数 C 需要选择以保证最优。本试验中采用的 C 的选择范围为$[10^{-5}, 1]$,从 40 个候选值中选择最优的。在含有噪音的情况下,根据实验流程中指出的加噪音的方法,对于每个噪音比例,需要得到 10 个数据集,每个数据集上采用上述方法得到模型量化值,对这 10 个数据集上的模型量化值平均才是具有该噪音比例的数据集上的结果。得到的模型变化曲线如图 2 所示。

由图 2 可知当不含噪音时该量化值为 1,随着噪音的加入,该值减小。对比图 1 与图 2 可知同一算法在同一数据集上的性能变化曲线与模型变化曲线基本一致。以 OHSUMED 为例,在噪音比例小于 0.45 时,图 1(a)的灰线所示的 RankSVM 的 MAP 值和图 2(a)的灰线所示模型的量化值都基本保持不变,但是超过 0.45 两者则有明显下降的趋势。在 TD2003 上图 2(a)中黑线所示的模型下降趋势比图 1(a)所示的性能变化趋势更加明显。这种一致性在 ListNet 的情况下(图 1(b)与图 2(b)所示)也是显而易见的。由此性能的好坏直接取决于模型得到验证。

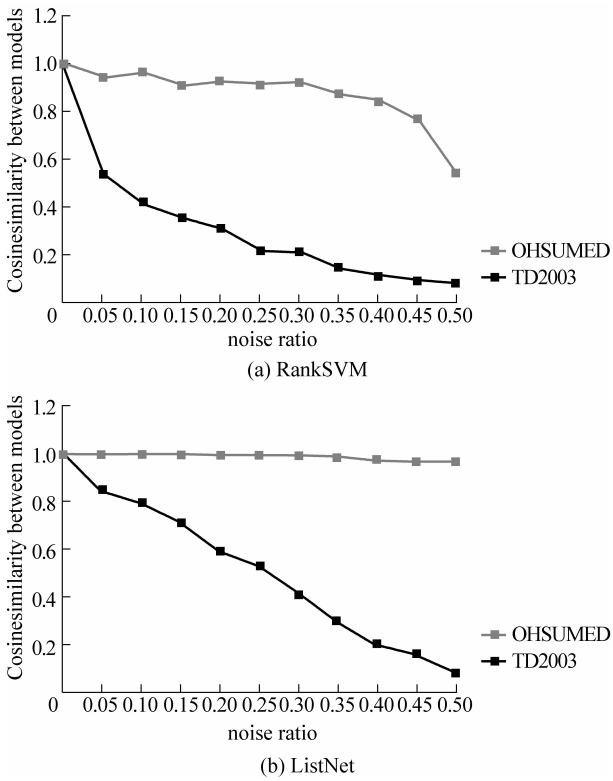

图 2 RankSVM 与 ListNet 模型在不同数据集上随噪音变化情况

4.2 根本原因

模型是从训练集中学习到的,训练集的质量决定了模型的好坏。在 Pairwise 方法这类具体的学习场景中,训练集是以偏序文档对的形式存在的。因此单个文档的标注错误对全序关系的影响并非是线性的。受此启发,本文中借助偏序文档对的分布,来探究单个文档上的标注错误对整个训练集带来的影响。研究发现正是由于单个文档上的标注错误对不同数据集的影响不同才导致的数据集的噪音敏感度不同。

为了描述训练集中文档对的分布情况,采用文档对与基向量的余弦相似度来刻画文档对的方向。选择一个基向量,用文档对的方向向量与该基向量的余弦相似度作为其量化值。采用从不含噪音的训练集上得到的模型 w_0 作为基向量,物理意义是明显的。当 w_0 与文档对的方向向量的余弦相似度为正值时,表示该模型可以对该文档对是可以正确分类的;否则是错误的。可以预见在不含噪音的训练集上,文档对的量化指标为正值的比例应该是最大

的。因此，模型与文档对的余弦相似度定义如式(1)所示，其中 d_i 与 d_j 为两个文档的属性向量，p 为文档对的属性向量，即两个文档属性向量之差。

$$\cos(w_0, p) = \cos(w_0, d_i - d_j) = \frac{w_0 \cdot (d_i - d_j)}{\| w_0 \| \| d_i - d_j \|} \quad (1)$$

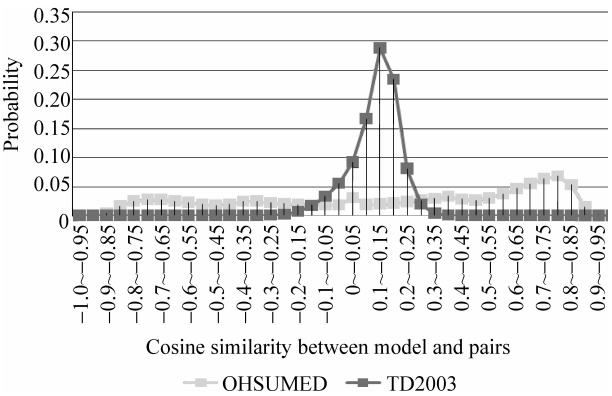

图 3　不同数据集上文档对方向的分布情况

根据上述定义，当 w_0 为 RankSVM 模型时，分别统计 OHSUMED 和 TD2003 的在不含噪音的训练集中的文档对与对应模型的夹角的余弦平均分布，按间隔 0.05，得到的概率分布如图 3 所示。从图 3 中可以直观看到，TD2003 上的文档对的方向分布较 OHSUMED 上的更集中。为了节省空间，便于从数量上比较，本文采用信息熵来度量文档对的散乱程度，如式(2)所示。同时探究这种散乱程度的大小关系与 w_0 的选取无关，表 2 中比较了 OHSUMED 和 TD2003 在 RankSVM 和 LsitNet 两个模型下的文档对的分布熵，记为 $\text{Entropy}_{\text{RankSVM}}$，$\text{Entropy}_{\text{ListNet}}$。

$$Entropy = -\sum_x p(x)\log p(x) \quad (2)$$

表 2　熵度量文档对方向分布的散度

数据集 X	OHSUMED	TD2003
EntropyRankSVM	5.028 5	2.775 5
EntropyListNet	5.024 5	4.663 3

从表 2 中可知，无论 w_0 采用 RankSVM 模型还是 ListNet 模型，OHSUMED 上文档对分布的熵总比 TD2003 上的要大，从而也说明了 OHSUMED 上的偏序文档对分布更分散。至此可以猜测，文档对分布越集中越容易受噪音文档对的影响。若上述猜测成立，那么用训练集上文档对分布的分散程度来解释噪音敏感度在不同数据集上的差异就是正确的。因此需要进一步验证当文档对的分布分散时，加入同等噪音后分布的改变情况是否会比文档对分布集中时的要小。

4.3　正确性验证

为了验证文档对的分布的集中与分散是造成噪音敏感程度不同的根本原因，本文采用 TD2003(文档对分布集中)和 OHSUMED(文档对分布分散)两个典型的数据集，以同样的方法向各自的训练集中加入同等噪音的比例后，统计其概率分布，并计算与不含噪音的训练集上的概率分布的差异，比较其差异的大小。

当 w_0 采用 RankSVM 模型时，在 OHSUMED 和 TD2003 上取了 0.1 和 0.2 两个噪音比例来达到清晰比较的目的，得到的结果如图 4 所示。从图 4(b)中可以很容易看出 TD2003 不含噪音的数据集上的文档对的分布与噪音水平为 0.1 时的文档对分布的差异要小于与噪音水平为 0.2 时的文档对分布的差异，这一点在图 4(a)中 OHSUMED 数据集上是不明显的，但是可以通过下面提到的一个量化值得到。图 4(a)与图 4(b)中从均值和方差的角度对比分布间的差异，可以很直观看出 OHSUMED 噪音水平为 0.1 的数据集与不含噪音的数据集的文档对分布的差异要小于 TD2003 上两者的差异。受篇幅所限，本文中仅对 RankSVM 的情形给出了直观分布图，对于 ListNet 的情形只能从数量上进行比较。

本文采用 Kullback-Leibler 散度[18]即相对熵来度量任意含噪音数据集上文档对的分布(记为 $p(x)$)与原始的不含噪音数据集 X 上文档对分布(记为 $p_0(x)$)间的差异，如式(3)所示。利用 Kullback-Leibler 的定义，可以对这种差异从数量上进行比较，比较结果如表 3 所示。

$$I(p, p_0, X) = \sum_{i=1}^{n} p(x_i)\log\frac{p(x_i)}{p_0(x_i)} \quad (3)$$

表 3　Kullback-Leibler 度量含噪音数据集与不含噪音数据集上的文档对分布的差异

数据集 X	RankSVM		ListNet	
	$I(p_{0.1}, p_0, X)$	$I(p_{0.2}, p_0, X)$	$I(p_{0.1}, p_0, X)$	$I(p_{0.2}, p_0, X)$
OHSUMED	0.175 3	0.265 9	0.005 4	0.014 0
TD2003	0.821 5	3.551 5	0.579 1	0.632 9

从表 3 的每一列来看，它对比了在相同的噪音

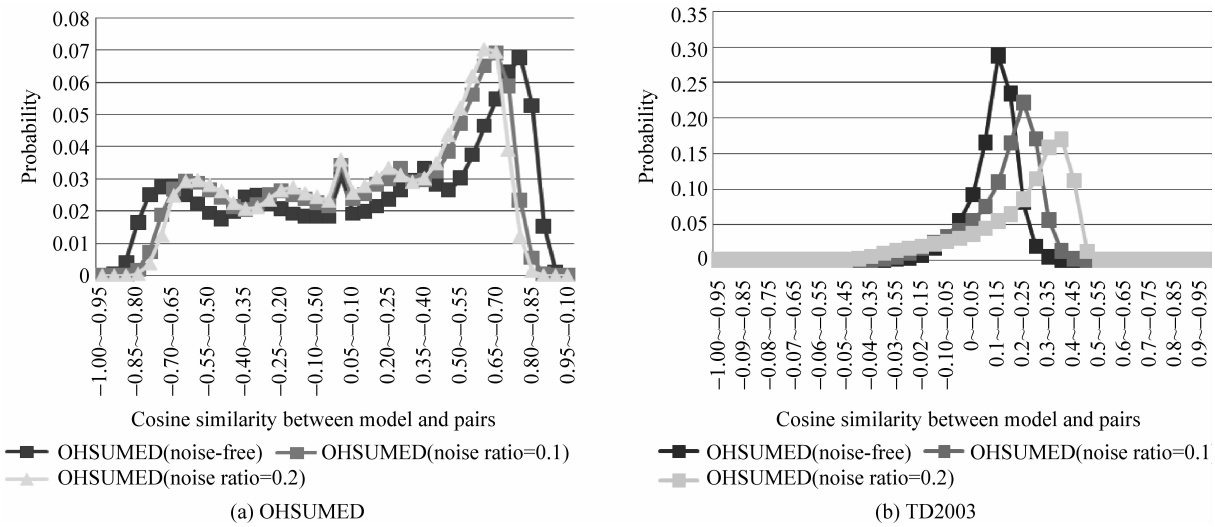

图4 不含噪音的与含噪音的数据集上文档对分布的差异

比例下，OHSUMED和TD2003上的文档对分布分别相对于各自的不含噪音时的分布的改变程度，很明显TD2003上的要远远大于OHSUMED的。当噪音比例为0.2时，OHSUMED上文档对分布相较于原始不含噪音时的分布改变量为0.265 9，而TD2003上为3.551 5。由于TD2003文档对的方向分布较集中，即使有少量的噪音文档对存在，也会使得分布的改变很大，很容易对学习到的模型的方向，即这些文档对的主方向，造成很大影响；OHSUMED中的文档对分布较分散，加入少量的噪音文档对，不会对分布改变太大。这一结论是明显的。由此也验证了文档对分布的不同是不同数据集上噪音敏感度差异的根本原因。

5 结论与展望

在将排序学习算法应用到实际的搜索问题中的第一步便是要构建一个好的训练集。由于在标注过程中受各种主观与客观因素的影响，噪音是不可避免的。针对这个问题，以往的工作多关注如何构建一个鲁棒的算法能够对噪音更不敏感，考虑到训练集的构建这一具体应用背景，本文从数据集的角度分析了在不同数据集上噪音敏感度差异的根本原因。首先本文验证了影响模型性能的噪音敏感型的最直接因素在于模型本身，而不是评价指标或者测试数据集；其次追究模型的来源，受Pairwise学习场景的启发，提出了训练集上的文档对的方向分布才是影响噪音敏感度的根本原因，因此从直观上得到了这样的结论：文档对的方向分布越分散，数据集对噪音越不敏感。最后对该结论的正确性进行验证。

下一步的工作主要包含两个方面：其一，还需要利用更多的实验数据来证明这一结论；其二，上述结论用来指导训练集的构建还是不够具体的，需要结合数据集本身的一些属性来得到进一步的结论，同时也需要进一步的实验证明。

参考文献

[1] Sheng, et al. Get another label? improving data quality and data mining using multiple, noisy labelers[C]// Proceeding of the 14th ACM SIGKDD. New York: ACM, 2008: 614-622.

[2] Xu Jingfang, Chen Chuanliang, Xu Gu, et al. Improving quality of training data for learning to rank using click-through data[C]//Proceedings of the third WSDM. New York: ACM, 2010: 171-180.

[3] Nettleton D. F., Orriols-Puig A., Fornells A., et al. A study of the effect of different types of noise on the precision of supervised learning techniques [J]. Artificial Intelligence Review, 2010, 33: 275-306.

[4] Chapelle O., Chang Yi, Liu Tie-Yan. Future directions in learning to rank [J]. Journal of Machine Learning Research, 2011, 14: 91-100.

[5] Tsivtsivadze E., Cseke B., Heskes T. Kernel Principal Component Ranking: Robust Ranking on Noisy Data[C]//Proceedings of the ECML/PKDD-Workshop on Preference Learning. Pascal Lecture Series, 2009: 101-113.

[6] Carvalho V. R., Elsas J. L., Cohen W. W., et al.

(下转第128页)