文本搜索排序中构造训练集的一种方法①

王 黎 帅建梅 (中国科学技术大学 自动化系 安徽 合肥 230027)

摘 要: 在文本搜索领域，用自学习排序的方法构建排序模型越来越普遍。排序模型的性能很大程度上依赖训练集。每个训练样本需要人工标注文档与给定查询的相关程度。对于文本搜索而言，查询几乎是无穷的，而人工标注耗时费力，所以选择部分有信息量的查询来标注很有意义。提出一种同时考虑查询的难度、密度和多样性的贪心算法从海量的查询中选择有信息量的查询进行标注。在 LETOR 和从 Web 搜索引擎数据库上的实验结果，证明利用本文提出的方法能构造一个规模较小且有效的训练集。

关键词: 信息检索；自学习排序；构造训练集

Construct Training Set for Learning to Rank in Web Search

WANG Li, SHUAI Jian-Mei

(Department of Automation, University of Science and Technology of China, Hefei 230027, China)

Abstract: Learning to rank has become a popular method to build a ranking model for Web search. For the same ranking algorithm, the performance of ranking model depends on a training set. A training sample is constructed by labeling the relevance of a document and a given query by a human. However, the number of queries in Web search is nearly infinite, and the human labeling cost is expensive. Therefore, it is necessary to select a subset of queries to construct an efficient training set. In this paper, a algorithm is developed to select queries by simultaneously taking the query difficulty, density, and diversity into consideration. The experimental results on LETOR and a collected Web search dataset show that the proposed method can lead to a more efficient training set.

Keywords: information retrieval; learning to rank; construct training set

1 引言

互联网的发展带来了海量的信息，如何快速准确地找到需要的信息变得尤为重要。文本搜索引擎是一种信息检索的工具，它的作用是根据用户的查询对它索引的网页进行排序。由于网页有固定的结构，可以提取很多有用的特征，帮助搜索引擎进行文档排序。但以往使用的文本排序算法(如 BM25[1]和 LMIR[2,3])的参数需要人工设定，当增加新的特征时，需要调整更多的参数，增加计算量。为了充分利用更多的特征信息，一些研究人员提出用机器学习的方法来解决文本排序问题。近年，自学习排序的方法引起学术和商业界的普遍关注。自学习排序是一种有监督的机器学习方法。首先根据训练集学习排序模型，然后对任意给定的查询，使用学到的排序模型对文档按相关性进行排序。具体来说，自学习排序的过程可以看作以下三步：

(a)构造训练集。自学习排序算法将查询和文档组成的对看作一个样本，记作 $(\Phi(q,d),r)$，其中 q 代表查询，d 代表文档，$\Phi(q,d)$ 表示根据文档和查询的关系提取的特征，r 表示人工标注文档与查询的相关程度，比如文档与查询非常相关标为 2，相关标为 1，完全不相关标为 0。

(b)根据训练集学习排序模型。目前自学习排序模型主要分为三类：基于样本点[4]，基于样本对[5-9]，基于样本序列[10]。McRank[4]把 $(\Phi(q_i,d_j),r_{ij})$ 这样的样

① 基金项目:国家高技术研究发展计划(863)(2006AA01Z449)
收稿时间:2010-01-18;收到修改稿时间:2010-02-26

本点作为输入，排序问题被简化为一个普通的分类或回归问题。Ranking SVM[5,9], Frank[6], RankBoost[7], RankNet[8],把 $(\Phi(q_i,d_j),r_{ij})$ 和 $(\Phi(q_i,d_k),r_{ik})$ 这样的样本对作为输入，排序问题转化为分类问题。ListNet[10] 则把关于查询的样本序列 $(\Phi(q_i,d_1),r_{i1})$，$(\Phi(q_i,d_2),r_{i2})$，$\cdots$ $(\Phi(q_i,d_n),r_{in})$，作为输入，通过最小化损失函数得到排序模型。

(c)排序模型应用。任意给定查询 q_i，对数据库中所有文档 $d_j \in D$，提取特征 $\Phi(q_i,d_j)$，根据排序模型预测文档与查询的相关性，然后按相关性对文档进行排序。

从自学习排序的过程可以看出排序模型的性能很大程度上依赖于训练集。训练集的好坏取决于所含样本的信息量。如果训练集的样本能够很好地反映整体样本的分布情况，那么学习得到的排序模型能更准确地预测文档与查询的相关性。选择样本的一个重要因素是查询。一般来说，一个查询对应一系列文档。人工标注文档与查询的相关性费力耗时，与分类问题的标注不同，它不仅需要判断文档是否与查询相关，而且需要判断文档与查询在多大程度上相关。显然，增加了标注的难度。所以，我们认为构建训练集最重要的是如何从海量的查询中选择有信息量的查询进行标注。在这篇文章中我们提出一种贪心算法来选择需要标注的查询。该算法基于以下三个准则：查询的难度，密度和多样性。

2　选择查询算法

通过对查询空间的研究，我们定义了选择查询的三个准则：查询的难度，密度和多样性，并且把这些准则运用到贪心算法，构造训练集。

2.1 训练集构造

基于选择查询的准则，我们设计一个贪心算法逐个选择查询来构建训练集，具体算法如图 1。

2.2 查询的难度

所谓有难度的查询就是用排序模型不容易预测其文档相关性的查询，也就是机器学习中常见的不确定性。在分类问题中，不确定准则得到了广泛的应用，比如 SVM[11]的支持向量就是难以预测其类别的样本。基于同样的道理，我们认为如果把有难度的查询加入训练集会使排序模型更加健壮。

对查询难度的预测，有很多方法。比如 Clarity score[12]，不过这种方法需要对每个文档训练一个模

条件：候选查询集合 Q，训练集 Q_0 中包含查询的数目

初始化：$Q_0 \leftarrow f$

根据本文提出的方法分别计算查询的难度 $Diff(q)$，密度 $Dens(q)$ 和多样性 $Divs(q,Q_0)$

for i=1 to k do

for all query in $Q-Q_0$ do

$I(q)=a*Diff(q)+b*Dens(q)+(1-a-b)*Divs(q,Q_0)$

end for

add $q=\arg\max_{q\in Q-Q_0} I(q)$ into Q_0

end for

return Q_0

图 1　训练集构造算法

型，这样的要求对网络搜索是不可能实现的。而委员会主动学习方法“query by committee”[13]采用最大相反策略，构建多个模型，每个模型对样本进行评估，得到最大相反评估的样本被看作为难度大的样本。受此启发，本文中我们采用 BM25 排序方法，计算文档的不同部分与查询的相关程度，给出相应的排序结果。比如根据标题与查询的相关性对文档进行排序结果为 r^1，根据正文排序结果为 r^2，$\cdots$，查询的难度计算如下：

$$Diff(q)=\sum_{i=1}^{N}\sum_{j=i+1}^{N} D_R(r^i,r^j) \qquad (1)$$

其中 N 表示文档被分成多少个不同的部分，D_R 用来计算两个排序的距离，排序距离大的查询被看作难度大的。本文中我们使用肯德尔等级相关系数[14](kend- alltau rank coefficient)来估计两个排序的距离，公式如下

$$D_R(r^i,r^j)=\frac{n_c-n_d}{n(n-1)/2} \qquad (2)$$

其中 n_c 表示在 r^i,r^j 中排序一致的文档对数，n_d 表示

在 r^i, r^j 中排序相反的文档对数，n 表示文档数目。

2.3 查询的密度

在查询样本空间中位于高密度区域的查询更具有代表性。而选择有代表性的样本作为训练集的一部分是机器学习算法中常用准则。我们利用核密度估计的方法计算在整个查询空间中每个查询的密度，公式如下

$$Dens(q)=\frac{1}{|Q|h}\sum_{i=1}^{|Q|}\frac{1}{\sqrt{2p}}e^{\frac{D_Q(q,q_i)^2}{2h^2}} \quad (3)$$

其中是 h 窗宽参数，|Q|表示候选集中查询的数目，核采用标准高斯函数，均值为 0，方差为 1，$D_Q(q_i,q_j)$ 代表查询 q_i 和 q_j 的距离。

计算查询距离时，我们把查询对应的一系列文档根据不同的特征进行排序，用来表示该查询。其中 r^i 表示根据第 i 个特征对文档排序的结果，M 表示使用了多少种不同的特征。通过计算不同特征对应的文本排序距离的均值，得到查询的距离，公式如下：

$$D_Q(q_i,q_j)=\frac{1}{M}\sum_{k=1}^{M}D_R(r_i^k,r_j^k) \quad (4)$$

其中 D_R 同上，用来计算排序 r_i^k 和 r_j^k 的距离，r_i^k 表示查询 q_i 对应的文档关于第 k 个特征的排列顺序，r_j^k 表示查询 q_j 对应的文档关于第 k 个特征的排列顺序。

2.4 查询的多样性

样本的多样性也是机器学习中选择样本的一个常用准则。多样性能够保证所选择的样本不局限于高密度区域，从而使所选择的样本能更全面地反映样本空间的信息。我们采用的多样性准则就是在给定的候选集中，选择与已经被选中的集合距离最远的样本。计算查询 q 与已经被选中的查询集合 Q_0 的距离 $Divs(q,Q_0)$ 定义，公式如下：

$$Divs(q,Q_0)=\min_{q'\in Q_0}D_Q(q,q') \quad (5)$$

3 自学习排序算法

Ranking SVM，用分类方法解决排序问题，是一种常用的排序模型。对给定的查询 q，若文档 d_i 比文档 d_j 更相关，排序模型的目标就是找到向量 w，满足 $w\Phi(q,d_i)>w\Phi(q,d_j)$，其中 $\Phi(q,d_i)$ 表示根据查询 q 和文档 d_i 的关系提取的特征。借鉴传统的 SVM 方法，Ranking SVM 的具体公式如下：

$$\min_{w,x}\frac{1}{2}\|w\|^2+C\sum_{i,j,k}x_{ijk} \quad (6)$$

Subject to: $w(\Phi(q,d_i),\Phi(q,d_j))\geq 1-x_{ijk}$

其中 C 是惩罚因子，决定了排序模型惩罚误判样本的程度。x_{ijk} 是松弛因子，表示样本错分程度。

4 实验分析

4.1 实验数据

为了评价本文算法的有效性，我们分别在 LETOR 和 Web 搜索引擎的数据库上进行了实验。

LETOR2.0 是微软亚洲研究院发布的一个用于研究自学习排序算法的标准数据集。其中 TD2004 包含 75 个查询(45 个查询是训练集，15 个查询是验证集，剩下的是测试集)，每个查询大约对应 1000 个文档，人工标注文档与查询关系，不相关标为 0，相关标为 1。根据文档和查询的关系提取 44 维特征，包含在文档的不同部分(如链接锚文本，the URL，文档标题，文档正文等)单词出现频率(term frequency,TF)，逆向文件频率(inverse document frequency,IDF)，BM25 等特征。

从Web搜索引擎收集数据，总共有2024个查询，随机选 5 个查询作为验证集，500 个查询作为测试集，剩下的是训练集。文档与查询的相关程度从 0 到 3，0 表示完全不相关，3 表示完全相关。根据文档和查询关系提取 354 维特征。

4.2 实验结果

为了证明较小规模训练集的可适用性，我们在 LETOR2.0 TD2004 上进行了试验。该数据集原本包含 45 个查询作为训练集，我们用本文提出的方法从中分别选择 5，10，15，20，25 查询构成训练集，基于不同规模的训练集，使用 Ranking SVM 学到不同的排序模型，作用于相同的测试集。我们用测试集中所有查询的平均 NDCG[15]评价排序模型，结果如图 2，可以看出随着查询数目的增多，排序模型的性能提高，当训练集中包含 25 个查询时，排序模型的性能可以与原始训练集（45 个查询）相当。也就是说，我们可以从原始的训练集中，选择部分查询来构造较小规模的训练集可以学得很好的排序模型。

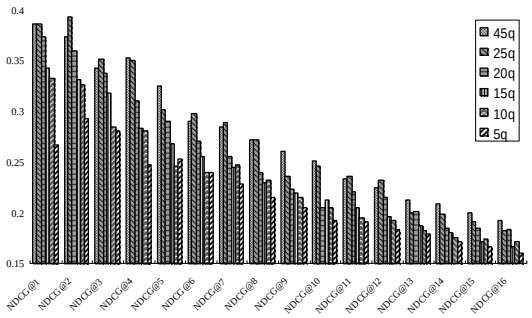

图 2　TD2004 不同规模训练集的性能

为了证明本文提出的选择查询方法的有效性，我们在 Web 搜索引擎收集的数据上进行了实验。首先，设定训练集中查询的个数分别为 5，10，20，30，50，100，200，500。接着采用本文提出的贪心策略选择查询构造了相应的训练集，标记为“Select”。然后用随机方法构造训练集，标记为“Random”。本文采用 Ranking SVM 学习排序函数，将其作用在相同的测试集上得到对应模型的性能指标。我们使用 NDCG 和 MAP[15]评价模型性能，结果如图 3，其中随机选择方法构造训练集时，相同规模的训练集随机构造十次，分别对这十个训练集求性能指标，取均值作为随机训练集的性能。可以看出，本文提出的选择查询方法明显优于随机选择的方法。

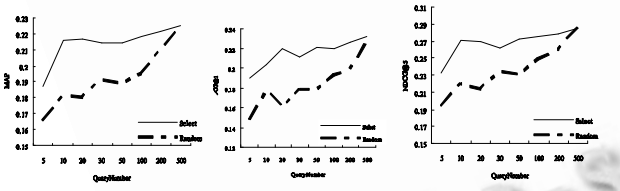

图 3　Live Web 数据上训练集性能

5　结论

训练集很大程度上影响学习模型的性能。本文基于查询的难易程度，密度，多样性准则，提出了一种贪心选择算法构建训练集，在 LETOR 和 Web 文本搜索引擎数据库上的实验结果，表明利用该算法可以构造一个有效且规模小的训练集。

参考文献

1 Robertson SE, Walker S, Jones S, Hancock-Beaulieu M, Gatford M. Okapi at TREC-3. Proc. of the Third Text Retrieval, Gaithersburg, USA, 1994.

2 Lafferty J, Zhai C. Document Language Models. Query Models and Risk Minimization for Information Retrieval. SIGIR, 2001.111－119.

3 Ponte JM, Croft WB. A Language Modeling Approach to Information Retrieval. SIGIR, 1998,275－281.

4 Li P, Christopher JC. Burges, Wu Q. McRank: Learning to Rank Using Multiple Classification and Gradient Boosting. NIPS, 2007.

5 Herbrich R, Graepel T, Obermayer K. Large Margin Rank Boundaries for Ordinal Regression. MIT Press, Cambridge, 2000.

6 Tsai MF, Liu TY. FRank: A Ranking Method with Fidelity Loss. SIGIR, 2007.383－390.

7 Freund Y, Iyer R, Schapire RE, Singer Y, Dietterich G. An Efficient Boosting Algorithm for Combining Preferences. Journal of Machine Learning Research, 1998:170－178.

8 Burges C, Shake T, Renshaw E, Lazier A, Deeds M, Hamilton N, Hullender G. Learning to Rank Using Gradient Descent. ICML, 2005.89－96.

9 Joachims T. Optimizing Search Engines using Clickthrough Data. SIGKDD 02 Edmonton, Alberta, Canada.

10 Cao Z, Qin T, Liu TY, Tsai MF, Li H, Learning to Rank: From Pairwise Approach to Listwise Approach. Proceedings of the 24th International Conference on Machine Learning, Corvallis, 2007.

11 Tong S, Koller D. Support vector machine active learning with applications to text classification. JMLR, 2002,45－66.

12 Cronen-Townsend S, Zhou Y, Croft BW. Predicting query performance. In SIGIR, 2002:299－306.

13 Freund Y, Seung SH, Shamir E, Tishby N. Selective sampling using the query by committee algorithm. Machine Learning, 1997,28(2-3):133－168.

14 Kendall GM. A new measure of rank correlation. Biometrika, 30(1/2):81－93, June 1938.

15 Jarvelin K, Kekalainen J. Ir evaluation methods for retrieving highly relevant documents. SIGIR, 2000: 41－48.