文章编号：2095-2163(2020)02-0348-05 中图分类号：TP391.1 文献标志码：A

网页搜索排序模型研究

李明琦
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要： 互联网发展至今，搜索引擎在人们生活中扮演着不可或缺的角色，网页搜索排序对于搜索引擎至关重要。优化网页排序，可以使用户节约大量甄别信息的时间，得到满意的结果。如今，排序学习即基于统计的排序方法被广泛应用于各个搜索引擎的网页排序中，可以结合多种文档特征，对文档进行深度语义理解，能够十分有效地对检索结果进行排序。本文主要研究基于排序学习的各个排序模型，同时探索文档的有效表示方法，旨在通过对比实验，得到更优的排序结果。

关键词： 网页排序；排序学习；文档表示

Research of Web search ranking model

LI Mingqi
（School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China）

【Abstract】 Since the development of the Internet, search engines have played an indispensable role in people´s lives. Web search ranking is very important for search engines. Optimizing Web ranking can save users a lot of time to identify information and get satisfactory results. Nowadays, learning to rank, which is based on statistical ranking methods, is widely used in Web ranking of various search engines. It could combine various document features and deeply understand the documents, thus rank the documents effectively. This paper mainly studies various ranking models based on learning to rank, and explores the effective representation method of documents. It aims to obtain better ranking results through comparative experiments.

【Key words】 Web search ranking; learning to rank; document representation

0 引 言

随着互联网的发展以及相关技术的不断提升与完善，人们获取信息的主要途径从查阅书籍、报纸等纸质材料转向了搜索引擎。然而，用户往往需要耗费大量时间从纷繁复杂的网页中去甄别有用信息。因此，高效地进行信息检索的主要挑战是优化网页搜索的排序以检索出与用户查询相关的文档。时下，许多搜索引擎仍在进一步研究新的排序算法，来提升用户的满意度。

目前，第二代搜索引擎还有一些不足之处。其一是相关性问题，即用户使用的检索词与文档的相关程度。通常使用的语言是复杂的，难以通过表面的文档特征来判断该文档是否与检索词相关。一方面，这种判断会使搜索引擎返回大量网页，且容易发生排序作弊现象，另一方面，这种判断无法返回不包含检索词、但相关性高的文档。其二，大多搜索引擎是根据关键词匹配，经常给出许多混合结果，不能有效解决问题。虽然搜索引擎展示的结果与用户检索的关键词相关，但是这并不能满足用户对信息的需求与期待。

改善这些问题的一种方法是更好地理解用户行为，在不断地检索过程中，搜索引擎收集到了大量的用户行为数据，通过分析和利用这些数据，可以有效提升排序效果。同时，如果能在文档特征中加深语义理解，就能使检索词与文档的相关程度分析得更为精准，从而能够提高用户的满意度。

近期的研究者已经转而关注起新的研究任务，并不是因为网页搜索排序问题已经完全解决了，而是因为这个任务到达了一个平台期。网页搜索排序问题仍然有着实际重要性，因此还需展开深入系统研究，推动该领域的发展与进步。

1 相关工作

网页搜索排序，即给定一个查询 Q 和一个网页文档集合 D，基于文档和查询的相关性得分，给出最相关 k 个文档的顺序。迄今为止，许多学者尝试了各种方法来解决这个问题，而且取得了较为可观的成果。

搜索引擎在早期时，主要用到的网页排序思想是根据关键词在文档中出现的位置和频率进行排序。基本原理是，关键词在文档中的词频越高，出现的位置越重要，则被认为和检索词的相关性越好。OkapiBM25[1]是一个流行的基于 tf/idf 的排序函数。

作者简介：李明琦(1994-)，女，硕士研究生，主要研究方向：自然语言处理。

收稿日期：2019-06-06

然后，出现了链接分析排序技术，其思想源于文献引文索引机制，若网页被其他网页引用的次数越多，或者被越有价值的网页所引用，该网页的价值就越大。

斯坦福大学 Page 等人[2]提出了 PageRank 算法，基本思想是，以 PageRank 值来判断网页的重要程度，PageRank 值取决于 2 个特征，其一是引用该网页的网页个数，其二是引用该网页的网页重要程度。但 PageRank 算法会严重排斥新加入的网页，并且没有将网页的主题相关性考虑到排序中。

斯坦福大学的 HaveliWala[3]提出了主题敏感的 PageRank 算法，解决了主题漂流问题，然而这个算法没有用主题的相关性来提高链接得分的准确性。

Google 的工程师 Bharat 等人[4]获得了 HillTop 算法的专利，解决了 PageRank 算法的查询无关性的问题，文档链接如果与查询主题相同会认定为更具可靠性，并只考虑专家页面，由专家页面对用户查询进行链接。这就有效处理了一些想通过增加循环链接数量提升 PageRank 值来作弊排序的网页。然而，专家页面在查询过程中权重非常大，这忽略了许多非专家页面的影响。

Kleinberg[5]提出的 HITS 算法是另一个基于超链接分析的著名排序算法，但仍然无法解决主题漂流的问题，并且可能产生主题泛化等问题。

在网络搜索领域，机器学习算法自动训练排序模型越来越流行，因为网页搜索中，有很多信息可以用来确定 query-doc 对相关性，并且可以利用大量的搜索日志。

Cao 等人[6]将 Ranking SVM 应用于文档检索，并对高排名的文档加强训练提出了用新的损失函数解决排序问题，应用了梯度下降和二次规划来优化损失函数。

Burges 等人[7]提出一种基于 PairWise 的 RankNet 方法，使用神经网络来训练模型，训练的损失函数为交叉熵，使用梯度下降来优化损失函数，时间复杂度优于 Ranking SVM。RankNet 优化的是 pairwise 错误的数量，但这与检索特征并不匹配。而其后提出的 LambdaRank 模型[8]，其设计思想是直接求梯度，而不是利用代价。LambdaMART 模型[9]则是结合了 MART 和 LambdaRank，也是基于 RankNet 的算法。

对于网页搜索的点击模型最初是考虑点击偏置来估计相关性。Richardson 等人[10]提出了位置模型，用户点击取决于文档位置和 query-doc 对相关性。Craswell 等人[11]提出了瀑布模型，研究中假设一个用户从头开始逐个地浏览文档，并且在遇到不相关文档后继续浏览，但在遇到相关文档后停止。许多复杂的点击模型都是基于这两个模型，如 UBM[12]、DBN[13]、CCM[14]。Chapelle 等人提出了 DBN，该模型假设了一次点击当且只当用户检测并且认为网页是可能相关的。

本文采用基于统计的排序学习方法，用不同的方法对文档进行表示，输入到排序模型中，进行对比实验，期望在小数据集上得到较好的排序效果。

2　文档表示方法和排序模型

本节拟探讨实验中采取的文档表示方法和排序模型。这里，排序任务是机器学习问题，需要抽取不同的特征来代表 query-doc 对，以输入到排序模型中进行训练。对此可做分析论述如下。

2.1　query-doc 对表示方法

首先，采用手工抽取特征的方式对 query-doc 对进行表示，每个 query-doc 对都由多维向量表示，每个维度都是一个特征。本任务抽取了文本特征、相似度特征、匹配特征、点击特征等 14 个特征，见表 1。

表 1　特征类型

Tab. 1　Type of features

类型	特征
文本特征	查询长度
	文档标题长度
	文档内容长度
相似度特征	查询与文档标题的 Word2Vec 相似度
	查询与文档内容的 tf-idf 相似度
	查询与文档内容的 Doc2Vec 相似度
匹配特征	查询与标题完美匹配
	查询与标题非完美匹配
	查询与内容完美匹配
	查询与内容非完美匹配
点击特征	DBN
	TCM
	PSCM
	UBM

文本特征考虑到了查询和文档，是非常传统的用于排序学习任务的特征，长度是基于中文分词后的结果进行统计。相似度特征是基于查询和文档关系的特征，由 3 种模型得到不同的文档向量表示，然后计算得到查询和文档的余弦相似度。匹配特征是

指关键词在文档标题或内容的出现情况,完美匹配即关键词以连续顺序出现于文档标题或内容,非完美匹配即关键词以不连续顺序出现于文档标题或内容。点击特征是由 4 种流行的点击模型(DBN, TCM[15], PSCM[16], UBM)训练得到的点击概率值,搜索引擎收集的各种用户行为信息揭示了查询和点击文档的相关信息,因此这些点击模型可以根据用户行为建立起来并且可以预测下次用户点击的位置,这些点击概率给研究者提供了极具价值的查询和文档相关程度信息。

这种用特征表示文档的方法是人们凭经验提取组合的,可能并没有足够好的表达 query-doc 对,并且人工抽取特征也较为耗费时间和人力。所以,研究尝试使用不同的深度学习方法对 query-doc 对进行表示,以期能够表示 query-doc 对的深度语义信息,再将这些特征向量作为排序模型的输入,可能提高整体模型的表现效果。

Word2Vec 方法可以将单词映射到向量空间,不仅考虑到了词与词之间的语义信息,而且还能将词语映射到低维度的向量,解决了 one-hot 向量稀疏的问题,常见的用词向量表示文档的方式有:对词向量的每一个维度取平均值,最大、最小值等。通常,直接拼接组合词向量是简单有效的方法,通过实验证明该方法能够在不同的 NLP 任务中取得较好的效果。

基于 Word2Vec 的文档表示方法,考虑到了词与词之间的语义信息,并且能够降低向量的维度,然而,研究时将文档中的所有词取平均值或最大、最小值会忽略词与词之间的顺序,同时对文本表示信息有一定影响。基于 Doc2Vec 的文档表示方法是对 Word2Vec 的扩展和改进,其段落向量保留了段落的主题信息,对段落进行记忆。Doc2Vec 模型可以将文档映射到固定维度的向量,既可以学习到词与词之间的语义信息,又可以保存词与词之间的顺序信息,用 Doc2Vec 对文档进行表示,可以很容易得进行文档相似度等计算,对于许多含有长文本的任务都有所帮助。

2.2 排序模型

排序学习是一个有监督的机器学习过程,通过对每一对给定的 query-doc 对,抽取查询文档的特征表示,然后通过训练排序模型,使得输出与实际数据相似。常用的排序学习分为 3 种类型:PointWise, PairWise 和 ListWise。其中,PointWise 方法只处理单独的文档,将文档转换为特征向量,根据训练数据得到的模型对其进行打分,再将所有文档按照得分结果进行排序。PairWise 方法将相关性得分转换为文档对关系,例如 A 的相关性得分为 3,B 为 2,C 为 1,则可得到 $A > B$, $B > C$, $A > C$ 等关系。这样就把排序问题转化成了二分类问题,利用训练模型,对所有文档进行分类得到偏序关系,从而构造全集的排序关系。ListWise 方法的输入为一个文档序列,通过构造合适的度量函数来优化排序,得到排序模型。

本课题通过调研选取 3 种稳定的排序模型进行实验,分别为:Ranking SVM、RankNet 和 LambdaMART。其中,Ranking SVM 和 RankNet 是基于 PairWise 方法的,LambdaMART 是基于 ListWise 方法的。

过程中,分别用前文所述方法对 query-doc 对进行表示,再与不同的排序模型进行组合,这里以使用 Word2Vec 取平均表示 query-doc 对,排序模型采取 RankNet 为例,设计模型如图 1 所示。

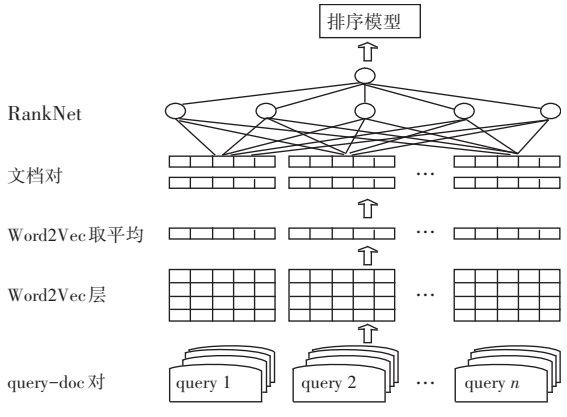

图 1 基于 Word2Vec 取平均的文档表示输入 RankNet 模型

Fig. 1 RankNet model based on the average of Word2Vec

3 实验

3.1 实验数据集

本课题用到的数据集来自 NTCIR-14 的 WWW2 任务,包含 2 万个 query-doc 对,含有相关性标签,其中 15 000 对作为训练集,5 000 对作为测试集。对于每对 query-doc 对,提供 4 种弱相关标签,由 4 种流行的点击模型得到,分别为:UBM, DBN, TCM, PSCM。这些点击标签利用了大量用户行为,如点击、跳过、停留时间。

3.2 数据预处理

原始数据为 xml 格式,包含查询内容、查询频率、查询 id、文档 url、文档 id、文档标题、文档内容、html、文档频率、文档点击标签等内容。

首先，需要对这些数据进行预处理。本课题只提取了查询内容、文档标题、文档内容、点击标签等信息。然后过滤掉内容、标题等数据中的非中文、空格、停用词、空数据等信息。最后，将数据进行繁简转换、分词等操作。

3.3　评价指标

针对回归、分类、排序等不同类型的问题，研究时用到的评价指标也不相同。网页搜索排序返回的结果通常是有序的，所以需要考虑其位置信息，本课题采用信息检索的常用评价指标如 NDCG、nERR、Q-measure，来度量排序结果的优劣。

3.4　实验结果

本节将 BM25 算法作为本课题研究的基线，该算法是文档检索的常用算法，思路非常简单。这里对比了加入点击特征对 query-doc 对进行表示的情况下，Ranking SVM、RankNet 和 LambdaMART 模型的表现，具体实验结果见表 2。

表 2　排序模型在不同特征组合下的实验结果

Tab. 2　The results of ranking model under different combination of features

排序模型	特征	NDCG@10	Q@10	ERR@10
LambdaMART	加入点击特征	**0.546 1**	**0.540 7**	**0.686 7**
LambdaMART	未加点击特征	0.467 9	0.473 5	0.602 6
RankNet	加入点击特征	0.502 8	0.514 2	0.654 3
RankNet	未加点击特征	0.433 7	0.434 4	0.571 7
Ranking SVM	加入点击特征	0.445 8	0.453 5	0.569 5
Ranking SVM	未加点击特征	0.400 6	0.409 4	0.535 3
BM25		0.326 7	0.332 2	0.464 1

由此可以看出，LambdaMart 模型表现效果最好，并且点击特征对排序结果非常有帮助。

采用 Word2Vec，Doc2Vec 等模型对文档进行表示，在排序模型上选择稳定性较好的 Ranking SVM、RankNet 和 LambdaMART 进行实验，继而比较 3 种评价指标的好坏，实验结果见表 3。

由此可以看出，基于深度学习的表示方法整体优于不加入点击特征时的手工提取特征的方法，Doc2Vec 模型表现最优。

4　结束语

如今，人们在日常生活中广泛使用互联网，对信息的获取主要求助于搜索引擎，因此对网页搜索排序结果进行优化是有着重要研究价值的，好的排序结果可以节省用户浏览大量低相关度网页的时间，并且返回用户满意的结果，从而解决人们生活中的实际问题。

表 3　基于不同文档表示方法的实验结果

Tab. 3　The results based on different document representations

表示方法	排序模型	NDCG@10	Q@10	ERR@10
Word2Vec 取平均	Ranking SVM	0.434 7	0.442 5	0.552 9
Word2Vec 取平均	RankNet	0.473 7	0.487 6	0.532 5
Word2Vec 取平均	LambdaMART	0.534 8	0.523 9	0.661 8
Word2Vec 取最大	Ranking SVM	0.452 5	0.451 8	0.583 7
Word2Vec 取最大	RankNet	0.491 8	0.517 5	0.645 5
Word2Vec 取最大	LambdaMART	0.547 1	0.549 8	0.678 2
Word2Vec 取最小	Ranking SVM	0.446 5	0.451 4	0.582 6
Word2Vec 取最小	RankNet	0.490 3	0.512 6	0.651 3
Word2Vec 取最小	LambdaMART	0.537 4	0.547 2	0.664 2
Doc2Vec	Ranking SVM	0.487 7	0.494 6	0.618 2
Doc2Vec	RankNet	0.532 5	0.546 9	0.671 7
Doc2Vec	LambdaMART	**0.561 4**	**0.553 7**	**0.692 3**

本文在少量标注样本数据集上，采用不同的 query-doc 对表示方法，对不同的排序模型如 Ranking SVM、RankNet、LambdaMART 进行对比实验。实验结果表明，点击特征对于提升排序效果非常重要，并且 LambdaMART 模型在本实验中的排序效果最好，稳定性较高。本文探索了多种基于深度学习的文档表示方法，如 Word2Vec 分别取平均值、最大值、最小值，Doc2Vec 模型，将以上模型生成的文档表示向量输入到排序模型中进行了对比试验。实验结果表明，用 Doc2Vec 模型来表示 query-doc 对，最终得到的排序结果是最好的，可以很好地捕捉到文档的语义信息。本文在网页搜索排序问题上取得了一定的研究成果，但是仍然存在一些不足。一方面，排序模型的实验以及基于深度学习方法表示文档的实验对比并不充足，未能尝试基于 pointwise 方法的排序模型，而且也没有用更多的深度学习方法（如 GRU 模型）对文档进行表示，这样会使实验结果不全面，不足以进行有效的论证。另一方面，数据样本较小，且数据存在不平衡性，这对提升排序效果的表现产生一定影响。

后续的研究工作可以从半监督学习方面开展，排序模型效果的表现与训练数据的多少相关，由前文可看出，即使研究尝试了多种文档表示方法、排序方法，排序结果的评价指标仍然没达到最理想的状态。因为研究中很容易地获取到大量的无标注网页，其中蕴含的信息对于训练排序模型也是很有价值的，因此可以利用半监督学习方法，自动标注一部分数据，这样就可以扩充训练集，同时也能尽量保证标签的准确性。

参考文献

[1] ROBERTSON S E, ZARAGOZA H. The probabilistic relevance framework: BM25 and beyond[J]. Foundations and Trends® in Information Retrieval, 2009, 3(4): 333.

[2] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: Bringing order to the web[R]. USA: Stanford InfoLab, 1999.

[3] HAVELIWALA T H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for Web search[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 784.

[4] BHARAT K, MIHAILA G A. Hilltop: A search engine based on expert documents[R]. Toronto: University of Toronto, 2000.

[5] KLEINBERG J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM (JACM), 1999, 46(5): 604.

[6] CAO Y, XU J, LIU T Y, 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. Washington: ACM, 2006: 186.

[7] BURGES C, SHAKED T, RENSHAW E, et al. Learning to rank using gradient descent[C]//Proceedings of the 22nd International Conference on Machine learning. Bonn, Germany: ACM, 2005: 89.

[8] SCHÖLKOPT B, PLATT J, HOFMANN T. Learning to rank with nonsmooth cost functions[C]//Advances in Neural Information Processing Systems. USA: MIT Press, 2007: 193.

[9] BURGES C J C. From RankNet to LambdarMANK: An overview [R]. Redmond: Microsoft Research, 2010.

[10] RICHARDSON M, DOMINOWSKA E, RAGNO R. Predicting clicks: Estimating the click-through rate for new ads[C]// Proceedings of the 16th International Conference on World Wide Web. New York: ACM, 2007: 521.

[11] CRASWELL N, ZOETER O, TAYLOR M, et al. An experimental comparison of click position-bias models[C]// Proceedings of the 2008 International Conference on Web Search and Data Mining. Palo Alto, CA: ACM, 2008: 87.

[12] DUPRET G, PIWOWARSKI B. A user browsing model to predict search engine click data from past observations[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore: ACM, 2008: 331.

[13] CHAPELLE O, ZHANG Y. A dynamic bayesian network click model for Web search ranking[C]//Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain: ACM, 2009: 1.

[14] GUO F, LIU C, KANNAN A, et al. Click chain model in Web search[C]//Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain: ACM, 2009: 11.

[15] ZHANG Yuchen, CHEN Weizhu, WANG Dong, et al. User-click modeling for understanding and predicting search-behavior [C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, CA: ACM, 2011: 1388.

[16] WANG C, LIU Y, WANG M, et al. Incorporating non-sequential behavior into click models[C]//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago, Chile: ACM, 2013: 283.

(上接第 347 页)

工作,无人作陪的孤独感会愈发强烈。虽有街坊邻居也不可能随时随地都陪在身边解闷,此时借助互联网技术,让社区参与互助养老的老人足不出户,就可通过网络加入到更多的娱乐活动中,参与各类有益身心的在线游戏;而且,也可以运用社交网络平台,随时随地与子女、家人联系,包括进行视频通话等。网络的便捷度也可以让同社区的老人距离更加亲近,当需要帮助时可以更快地传递消息。同时,老人体验了这种交互性强的社交娱乐方式,感觉新颖,从中得到乐趣,获得更好的休闲娱乐体验,同时也提高了生活质量。与子女、朋友的沟通方式多样化,也更方便快捷,孤单空虚感降低,生活也更加精彩。

5 结束语

现如今,互联网的影响已然无处不在。在此背景下,本文即对互联网时代下的城市社区互助养老模式中可提供的服务、及可行化建议进行了系统研究与讨论,不言可喻的是,这一问题的有效解决已与社会上的大多数家庭都密切相关。本文的研究初衷,在于探索寻求在新形势下最能满足老年养老生活需求的养老模式,力求用最高效的方式,用最小的投入,最终让社区居民家庭及老人用户的满意程度最大化,进而为解决社会养老问题及助力城市社区互助养老模式的良性、可持续发展发挥有益的推动作用。

参考文献

[1] 方静文. 超越家庭的可能:历史人类学视野下的互助养老—以太监、自梳女为例[J]. 思想战线,2015,41(4):78.

[2] 曹莹,苗志刚,李明杰,等. 基于互联网+的智慧互助养老服务模式研究[J]. 管理观察,2018(14):74.

[3] 常红林. “互联网+”背景下的社区居家养老模式构建[J]. 新闻研究导刊,2016,7(23):18.

[4] 曾小辉. 优势视角理论下的“‘互联网+’以老养老”新型养老模式研究—以沈阳市为例[D]. 沈阳:辽宁大学,2017.