融合位置相关和概率排序的Lucene排序算法改进

胡 博 蒋宗礼

(北京工业大学计算机学院 北京 100124)

摘 要 文档检索结果的排序和文本分类技术是解决垂直搜索、个性化信息检索、信息过滤等相关问题的核心技术。为了提高检索系统的性能,针对Lucene的基础排序算法,提出了一种融合位置相关和概率排序的改进方法。考虑到查询词在文档中出现的位置信息和概率排序对文档相关性的影响,利用位置相关的查询词权值和基于朴素贝叶斯分类算法的文档相关性概率值,对Lucene基础排序算法的评分公式进行改进。实验表明,该改进方法能够有效提高垂直搜索的准确率,使用户拥有更好的垂直搜索体验。

关键词 位置相关,概率排序,Lucene,排序算法,垂直搜索

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.9.049

Improvement of Lucene Sorting Algorithm Fusing Location-related and Probabilistic Sorting

HU Bo JIANG Zong-li

(College of Computer Science, Beijing University of Technology, Beijing 100124, China)

Abstract Sorting document retrieval results and text classification technology is the core technology to solve vertical search, personalized information retrieval, information filtering and other related issues. In order to improve the performance of retrieval systems, an improved method for integrating location-related and probabilistic sorting was proposed for Lucene default sorting algorithm. Taking into account the document relevance impact of query's location information and probabilistic sorting, the scoring formula of Lucene default sorting algorithm is improved using the probability value of document relevance based on naive Bayesian classification algorithm and the weights of location-related query. Experimental results show that this improvement can effectively improve the accuracy of vertical search, allowing users to have better vertical search experience.

Keywords Location-related, Probabilistic sorting, Lucene, Sorting algorithm, Vertical search

1 引言

垂直搜索是针对某一特定领域、某一特定人群或某一特定需求而提供的有一定价值的信息和相关服务的搜索,它不仅仅是普通的网页索引,而且包括信息的加工和结构化的信息,尤其是针对特定的行业内容做了专业和深入的分析挖掘、精细分类、过滤筛选等,信息定位更精准[1]。Lucene是一个高性能、可伸缩的信息搜索(IR)库[2],它可以为应用程序添加索引和搜索功能。Lucene排序算法基于向量空间模型(Vector Space Model, VSM),使用TF-IDF(Term Frequency-Inverse Document Frequency)的统计方法来计算查询语句与文档的相关度并为文档打分,根据得分高低对文档进行排序。其基础排序算法保证了检索结果与查询语句有一定的相关性。

文献[3,4]对基于Lucene的小型搜索引擎进行了初步的探索与研究;使用网络爬虫工具Heritrix抓取Web页面,分析并提取出结构化的文档数据,利用Lucene对文档集合建立索引和检索,设计实现了一个小型的垂直搜索引擎系统。但其只是简单地使用开源检索工具Lucene实现了对文档的索引和检索功能,在中文分词、检索结果排序等方面没有做更加深入的研究,还存在着很大的发展空间。

文献[5-8]主要是对基于Lucene的搜索引擎的改进进行了研究。其中,文献[5,6]在Lucene的基础排序算法上加入了PageRank技术,提高了网页检索的准确率。Page-Rank算法是通过网页中的链接来计算网页的重要性,在Lucene基础排序的评分公式中加入PageRank值,从而影响网页文档检索结果的排序。一般来说,PageRank算法只适用于网页文档,而且PageRank算法仅关注网页的链接信息,没有考虑网页内容的重要性,所以在应用到垂直搜索引擎中时可能会出现主题漂移现象。

文本分类作为处理和组织大量文本数据的关键技术,可以在较大程度上解决信息杂乱的问题,方便用户准确地定位所需的信息和分流信息[9]。另一方面,对文本分类的研究能为文本提取、文本过滤及文档聚类等其他文本领域的研究提供重要的参考依据[10]。通过将文本分类技术与搜索引擎、信息推送、信息过滤等技术相结合,能够有效地提高检索结果的

到稿日期:2015-08-21 返修日期:2015-12-01 本文受计算机软件与理论北京市重点学科基金(007000541215042)资助。

胡 博(1992-),女,硕士,主要研究方向为网络信息检索、搜索引擎,E-mail:453982841@qq.com;蒋宗礼(1956-),男,教授,博士生导师,CCF会员,主要研究方向为网络信息搜索与处理、计算学科教育、工程教育。

相关度和准确率。

本文在上述文献的基础上，深入研究了基于 Lucene 的信息检索系统，并针对 Lucene 排序算法的不足之处进行改进，得到更加适用于垂直搜索引擎的排序算法。其中涉及到 3 个重要的相关性计算环节：1)基于 Lucene 的查询语句与文档的相关性计算；2)位置相关的查询词的权值计算；3)基于朴素贝叶斯分类算法的文档概率排序值计算。

2 Lucene 排序算法的不足

Lucene 基础排序算法的核心是其评分公式，如式(1)所示：

$$score(q,d)=coord(q,d)*queryNorm(q)*\sum_{t in q}(tf(tind)*idf(t)^2*t.getBoost*norm(t,d)) \quad (1)$$

其中，$coord(q,d)$为查询语句 q 在文档 d 中出现的频率，查询语句 q 可以由多个查询词 t 组成；$queryNorm(q)$为归一化的参数，与文档的相关度无关；$tf(t\ in\ d)$为词 t 在文档 d 中出现的频率，$idf(t)$ 为词 t 在整个文档库中的逆文档频率；$t.getBoost$ 为词 t 的加权值(默认值为 1.0)；$norm(t,d)$为可以在索引期设置域的加权值的索引期因子。

在式(1)中，使用了经典的 TF-IDF 算法对查询语句与文档的相关度进行评分，文档的评分越高，该文档在检索结果的位置越靠前。TF-IDF 算法的基本思想是，通过查询词在文档中出现的频率(TF)和查询词在整个文档集合中出现的逆文档频率(IDF)这两个值对词的权重进行计算。因此在基于 Lucene 的搜索引擎中，查询词在文档中出现的频率和逆文档频率是文档检索结果排序的最重要的指标。

从式(1)可以看出，Lucene 基础排序算法没有考虑查询词在文档中的位置信息对文档评分的影响。一般情况下，词出现在正文位置的概率较出现在标题、摘要位置的概率高，但是，词出现在标题位置的重要程度要比出现在正文位置的高。因此，查询词在文档中出现的位置也是影响文档相关性的重要因素之一。

例如，有两个文档 A、B，查询词 q 出现在文档 A 的标题位置 1 次、正文位置 5 次；查询词 q 只出现在文档 B 的正文位置 8 次。采用式(1)计算得出，q 与文档 B 的相关度评分值大于 q 与文档 A 的相关度评分值，则在检索结果中文档 B 排在文档 A 前面。如果考虑查询词在文档 A、B 中出现的位置信息，则 q 与文档 B 的相关度评分值不一定大于 q 与文档 A 的相关度评分值。

另外，Lucene 基础评分公式通过改进向量空间模型的余弦相似度公式来提高检索结果的相关性和实际可用性。向量空间模型对相关性采用了隐含的假设，即相关性是与查询语句和文档的相似度有关联的，因此对检索系统的验证是基于经验的，缺乏恰当的理论模型支撑。向量空间模型假设词之间相互无关，忽视了词与词之间的内在联系，如“跳水”和“郭晶晶”的出现是相关的，所以这一假设对计算结果的可靠性造成了一定的影响[11]。

概率检索模型的基本思想是：用户输入查询语句后，检索系统能够根据文档集合与用户需求的相关性的高低返回检索结果。如果把文档集合划分为两个类别：相关文档集合(即垂直检索针对的特定行业类别)和非相关文档集合(即与特定行业无关的类别)，就可以将这种相关性衡量转换为一个文本分类问题。让检索系统遵循概率排序原则，可以弥补向量空间模型的理论缺陷(即隐含假设)，提高检索系统的理论性和可靠性。

本文在 Lucene 的基础排序算法的基础上，对其评分公式进行了改进：1)在评分公式中融入查询语句在文档中的位置加权值；2)融入基于朴素贝叶斯分类算法的文档概率排序值。

3 Lucene 排序算法的改进

3.1 位置相关的查询词的权值计算

计算权值时，若全面体现查询词在文档中的位置，则会过多地增加算法的复杂性，可能导致算法的整体有效性的降低。所以，本文选择了查询词在文档中的标题(name)、摘要(abstract)、正文(text)3 个位置，认为它们体现了查询词最关键的位置特征，据此给予查询词不同的权值。本文采用的加权公式如式(2)所示：

$$weight=m_0+m_1*name+m_2*abstract+m_3*text \quad (2)$$

其中，m_0 为常数；m_1，m_2，m_3 为权重系数，三者的比值代表了查询词分别出现在 3 个位置对文档相关度的影响。查询词出现在某一位置时，该位置的值记为 1，不出现记为 0，如查询词出现在标题中，则 $name=1$。

3.2 文档相关性的概率值计算

每个查询语句 Q 都会有两组文档：相关文档集合 R 和非相关文档集合 NR，即每个 Q 都有两个类别的文档。当 $P(R|D)>P(NR|D)$时，文档 D 是相关的，其中 $P(R|D)$是文档 D 属于相关集合(即文档 D 相关)的条件概率，$P(NR|D)$是文档 D 属于非相关集合(即文档 D 不相关)的条件概率。利用贝叶斯公式：$P(R|D)=\frac{P(D|R)P(R)}{P(D)}$，可得出，当 $\frac{P(D|R)}{P(D|NR)}>\frac{P(NR)}{P(R)}$时，文档 D 是相关的。因此，如果文档 D 属于相关文档集合且 $\frac{P(D|R)}{P(D|NR)}$值较高，那么文档 D 的排序靠前。

文档表示为一组二元向量特征，$D=(d_1,d_2,\cdots,d_n)$[12]。其中 d_i 只有 0 和 1 两个值，1 表示词在文档中出现，0 表示不出现。由于朴素贝叶斯分类是基于属性之间条件独立性假设的，因此有：$P(D|R)=\prod_{i=1}^{n}p(d_i|R)=\prod_{i=1}^{n}p_i$。

本文采用的计算文档相关性的概率公式如式(3)所示：

$$P=\frac{P(D|R)}{P(D|NR)}=\prod_{i:d_i}\frac{p_i}{s_i}\cdot\prod_{i:d_i}\frac{1-p_i}{1-s_i}\propto\sum_{i:d_i}\log\frac{p_i(1-s_i)}{s_i(1-p_i)} \quad (3)$$

其中，p_i 为词项 i 在相关集合的某片文档中出现的概率，s_i 为词项 i 在非相关集合的某片文档中出现的概率。

3.3 改进 Lucene 排序算法的评分公式

本文将查询语句与被检索文档特征向量的相似度评分值、查询语句在文档中的位置加权值和经过朴素贝叶斯算法得到的文档相关性的概率值这 3 个值进行组合计算，得出新的评分公式，如式(4)所示：

$$newscore=k_0+k_1*score_1+k_2*score_2+k_3*score_3 \quad (4)$$

式中，$newscore$ 为排序算法的新评分值；$score_1$ 表示查询语句与被检索文档特征向量的相似度评分值，由式(1)计算得到；$score_2$ 表示查询语句在文档中的位置加权值，由式(2)计算得到；$score_3$ 表示经过朴素贝叶斯算法得到的文档相关性的概率值，由式(3)计算得到；k_0 为常数，k_1，k_2，k_3 为权重系数，可以通过更改三者的比值确定 $score_1$，$score_2$，$score_3$ 对检索结果

排序的不同影响，从而应用到不同的实际需求中。

本文采用多元线性回归分析方法求解 $newscore$，$score_1$，$score_2$，$score_3$ 这 4 个变量之间存在的统计线性关系。为式(4)建立多元线性回归模型，其中 k_0 为常数项；k_i 为参数，是 $score_i$ 的回归系数，表示在其他所有自变量不变的情况下，$score_i$ 每变化一个单位都能引起 $newscore$ 的变化。使用 SPSS 对数据进行统计分析，建立因变量为 $newscore$ 且自变量为 $score_1$，$score_2$，$score_3$ 的线性模型，解出各系数的值，$k_0=2.015$，$k_1=0.852$，$k_2=0.388$，$k_3=0.693$。同样采用多元线性回归分析方法，为式(2)建立多元线性回归模型，解出各系数的值，$m_0=0.366$，$m_1=0.631$，$m_2=0.475$，$m_3=0.419$。

3.4 改进的 Lucene 排序算法

输入：用户输入的查询语句

输出：与查询语句相关的文档集合

步骤 1　对查询语句 Q 进行分词，对文档集合进行索引和分词处理；

步骤 2　利用 Lucene 的检索功能，根据式(1)，得到相关文档的基础评分值；

步骤 3　根据式(2)计算查询语句在相关文档中的位置加权值；

步骤 4　根据式(3)计算文档相关性的概率值；

步骤 5　将步骤 2－步骤 4 得到的值代入到式(4)中，得到新的排序算法的评分公式；

步骤 6　根据步骤 5 中的新评分值，对文档重新排序，并将排好序的文档集合返回给用户。

4 实验与结果分析

为了测试改进后的排序算法的检索效果，本文设计了 Lucene 排序算法改进前后的对比实验。首先以太平洋汽车网站主页作为种子 URL 抓取了 10000 个网页，经过网页信息处理与抽取，得到结构化的文档集，将其作为实验所用的测试数据集。然后选择了 10 名测试者，每个测试者随机输入查询语句进行 20 次查询，并在系统返回的前 100 个检索结果中统计符合查询的文档数量。使用 Lucene 排序算法改进前后的垂直检索系统进行实验时，测试者、查询次数、每次的查询语句保持一致，以便于实验结果的统计与分析。

4.1 信息检索系统评价指标

召回率(Recall)、准确率(Precision)和 F 值度量(F measure)是信息检索系统中最常用的 3 个效果评价指标。召回率是指相关文档被检索出的比率。准确率是指检索出的文档中相关文档的比率。F 值度量是召回率和准确率的调和平均数，是综合召回率和准确率的效果评价指标。

用 R 表示召回率、P 表示准确率、F 表示 F 值度量，则 3 个评价指标的计算公式为：

$$R=\frac{\text{被检索到的相关文档的数量}}{\text{文档库中所有相关的文档数量}}$$

$$P=\frac{\text{被检索到的相关文档的数量}}{\text{被检索到的所有文档数量}}$$

$$F=\frac{1}{\frac{1}{2}(\frac{1}{R}+\frac{1}{P})}=\frac{2RP}{R+P}$$

4.2 实验结果分析

选取查询词个数分别为 1，2，3，5，10 时，使用基于 Lucene 基础排序算法的检索系统和基于改进 Lucene 排序算法的检索系统进行实验，并对检索结果的准确率、召回率和 F 值进行比较。

实验结果如表 1－表 3 所列。

表 1　Lucene 排序算法改进前后检索准确率的比较(%)

查询词个数	1	2	3	5	10
算法改进前	82.1	85.6	90.1	88.5	85.2
算法改进后	85.5	89.7	91.0	94.2	93.8

表 2　Lucene 排序算法改进前后检索召回率的比较(%)

查询词个数	1	2	3	5	10
算法改进前	77.6	73.1	65.9	66.4	66.2
算法改进后	83.5	80.3	80.7	76.6	79.6

表 3　Lucene 排序算法改进前后检索 F 值的比较(%)

查询词个数	1	2	3	5	10
算法改进前	79.8	78.9	76.1	75.9	74.5
算法改进后	84.5	84.7	85.5	84.5	86.1

为了更加直观地观察排序算法改进前后准确率、召回率和 F 值的数据变化，将表 1－表 3 中的数据采用折线图的方式显示，如图 1－图 3 所示。

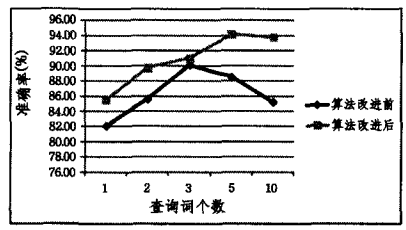

图 1　Lucene 排序算法改进前后检索准确率的比较

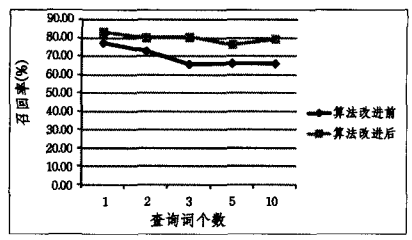

图 2　Lucene 排序算法改进前后检索召回率的比较

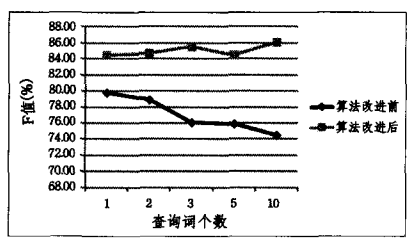

图 3　Lucene 排序算法改进前后检索 F 值的比较

实验结果分析：

(1)如图 1 所示，Lucene 排序算法改进后检索结果的准确率比改进前有了较大程度的提高。

(2)如图 2 所示，由于检索系统的召回率受到了朴素贝叶斯分类的影响，Lucene 排序算法改进后检索结果的召回率比改进前有了提高，但提高程度不明显。

(3)如图 3 所示，由于 Lucene 排序算法改进后检索系统的准确率和召回率均有不同程度的提高，F 值作为二者的调和平均数，也有了较大程度的提高且稳定在 85%左右。

结束语　本文将位置相关的查询词权值和基于朴素贝叶斯分类算法的文档相关性概率值，与 Lucene 基础排序算法的评分公式进行融合，得出新的排序算法的评分公式，检索系统

(下转第 273 页)

robic sludge blanket reactor[J]. Bioresource Technology, 2014, 165(8): 233-240

[4] Hosmer D W, Lemeshow S. Introduction to the logistic regression model[M]// Applied Logistic Regression(Second Edition). John Wiely & Sons, Inc. 2005: 1-30

[5] Ng A Y, Jordan M I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes [J]. Advances in Neural Information Processing Systems, 2002, 2 (3): 169-187

[6] Lemeshow S, Hosmer D W. A review of goodness of fit statistics for use in the development of logistic regression models[J]. American Journal of Epidemiology, 1982, 115(1): 92-106

[7] Cortes C, Vapnik V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297

[8] Rendle S. Factorization machines[C]// 2010 IEEE 10th International Conference on Data Mining (ICDM). 2010: 995-1000

[9] Rendle S. Factorization machines with libfm[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2012, 3 (3): 57

[10] Rendle S. Scaling factorization machines to relational data[C]// Proceedings of the VLDB Endowment. 2013: 337-348

[11] Rendle S, Gantner Z, Freudenthaler C, et al. Fast context-aware recommendations with factorization machines[C]// Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2011: 635-644

[12] Storn R, Price K. Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces [J]. Journal of Global Optimization, 1997, 11(4): 341-359

[13] Ilonen J, Kamarainen J K, Lampinen J. Differential evolution training algorithm for feed-forward neural networks[J]. Neural Processing Letters, 2003, 17(1): 93-105

[14] Kotha S R, Vij S, Sahoo S K. A study on strategies and Mutant factor in differential evolution algorithm for FIR filter design[C]// 2014 International Conference on Signal Processing and Integrated Networks (SPIN). 2014: 50-55

[15] Wang G-G, Gandomi A H, Alavi A H, et al. Hybrid krill herd algorithm with differential evolution for global numerical optimization[J]. Neural Computing and Applications, 2014, 25(2): 297-308

[16] Zhou Y L, Zhu Y H, Zhang J. Discrete Differential Evolution with Learning Mechanism[J]. Computer Science, 2011, 38(7): 225-227(in Chinese)
周雅兰，朱耀辉，张军. 具有学习机制的离散差分演化算法[J]. 计算机科学，2011，38(7)：225-227

[17] Wang C J, Wang X H, Xiao J M. Discrete Differential Evolution with Parameter Adaptive Mechanism [J]. Computer Science, 2014, 41(1): 279-282(in Chinese)
王丛佼，王锡淮，肖建梅. 具有参数自适应机制的改进离散差分进化算法[J]. 计算机科学，2014，41(1)：279-282

[18] Chun L Y, Song H, Yang J. Research on Music Classification Based on MFCC and BP Neural Network[C]// 2nd International Conference on Information, Electronics and Computer. 2014: 57-68

(上接第 249 页)

根据新的评分公式对文档集合重新排序。实验结果表明，与 Lucene 基础排序算法相比，使用改进后的排序算法时，检索系统的准确率、召回率和 F 值均有不同程度的提高，用户有更加良好的垂直搜索体验。但其还存在一些缺点，例如，在朴素贝叶斯分类模型中，分类的效率、准确率和召回率可能对检索系统的效率、准确率和召回率有所影响。下一步的研究工作是考虑用户兴趣和行为特征等因素，优化分类和检索过程，从而提高检索系统的性能。

参 考 文 献

[1] Liu J X, Sheng Y. The differences and case analysis of vertical and general search engines[J]. Modern Information, 2009, 29 (3): 143-149(in Chinese)
刘俊熙，盛宇. 垂直和通用搜索引擎的差异和案例分析[J]. 现代情报，2009，29(3)：143-149

[2] 牛长流，尚宇. Lucene 实战(第 2 版)[M]. 北京：人民邮电出版社，2011

[3] Bai K, Geng G H. Research and Application of vertical search engines based on Lucene/Heritrix[J]. Computer Applications and Software, 2009, 26(1): 212-215(in Chinese)
白坤，耿国华. 基于 Lucene/Heritrix 的垂直搜索引擎的研究与应用[J]. 计算机应用与软件，2009，26(1)：212-215

[4] Zhang X, Liu X F. Design and implementation of full-text search engine based on Lucene and Heritrix[J]. Modern Computer, 2013(22): 74-77(in Chinese)
张宣，刘晓飞. 基于 Lucene 和 Heritrix 的全文搜索引擎的设计与实现[J]. 现代计算机，2013(22)：74-77

[5] Cai F. Research and improvement of Lucene sorting algorithm [J]. New Technology and New Products of China, 2011(4): 15-16(in Chinese)
蔡峰. Lucene 排序算法的研究和改进[J]. 中国新技术新产品，2011(4)：15-16

[6] Chen J X, Huang R, Ma Z B. Optimization and implementation of Lucene sorting algorithm based on PageRank[J]. Computer Engineering and Science, 2012, 34(10): 123-127(in Chinese)
陈建峡，黄日，马忠宝. 基于 PageRank 的 Lucene 排序算法优化与实现[J]. 计算机工程与科学，2012，34(10)：123-127

[7] Mohd M. Development of Search Engines using Lucene: An Experience[J]. Procedia-Social and Behavioral Sciences, 2011, 18: 282-286

[8] Milosavljevic, Branko, Boberic, et al. Retrieval of bibliographic records using Apache Lucene[J]. The Electronic Library, 2010, 28(4): 525-539

[9] Rong G, Zhang H X. Application of text classification in the search engine[J]. Guide of Scitech Magazine, 2008, 12(2): 14-15 (in Chinese)
荣光，张化祥. 文本分类在搜索引擎性能中的应用[J]. 科技致富向导，2008，12(2)：14-15

[10] Lewis D D. Representation and learning in information retrieval [D]. Graduate School of the University of Maassachusetts, 1992

[11] Zhang X F. Analysis and evaluation of several common information retrieval model[J]. Journal of Intelligence, 2008, 27(3): 121-123(in Chinese)
张小芳. 几种常见信息检索模型的分析与评价[J]. 情报杂志，2008，27(3)：121-123

[12] Croft W B, Metzler D, Strohman T. Search Engine: Information Retrieval in Practice[M]. Pearson, 2010