第 35 卷 第 2 期 计 算 机 工 程 2009 年 1 月
Vol.35 No.2 Computer Engineering January 2009
·软件技术与数据库· 文章编号：1000—3428(2009)02—0047—03 文献标识码：A 中图分类号：TP391

一种基于多特征融合的博客文章排序算法

卢 刚
(浙江财经学院信息学院，杭州 310018)

摘 要：随着博客数据的迅速增长，在网络媒体中进行信息检索时的效率问题日益受到人们的关注。该文在针对博客搜索中特有的用户需求以及博客系统自身特点进行分析的基础上，提出一种基于博客文章相关性、时效性、查询类型和博客作者兴趣特征一致性等多特征融合的博客文章排序算法。实验结果证明了该算法性能优于传统算法。

关键词：博客；信息检索；排序；多特征融合；兴趣特征

Sorting Algorithm for Blog Articles Based on Fusion of Multi Features

LU Gang
(College of Information, Zhejiang University of Finance and Economics, Hangzhou 310018)

【Abstract】With the rapid development of Blog data, the efficiency of information retrieval in them is of great concern. Based on the analysis of the peculiar users' needs and the special feature of Blog system, this paper proposes a new sorting algorithm on the basis of the integration of multi features like relativity, timeliness, query type, Blog writer's interest characteristics and so on. Experimental result proves that the performance of the new algorithm is more efficient than the traditional ones.

【Key words】 Blog; information retrieval; sorting; fusion of multi features; interest characteristics

1 概述

随着互联网技术的发展，大规模的博客作者迅速聚集起了海量的数据，如何保证博客中信息检索的效率变得至关重要。在博客搜索中，用户的信息需求和传统的网页搜索有所不同，主要集中在以下 3 个方面：主题搜索，博客作者搜索以及评论搜索[1]。由于此类信息需求在传统网页搜索中是不存在的，因此对传统的结果排序算法提出了挑战。目前，百度以及Google等主流的中英文搜索引擎已经在博客搜索服务方面开展了一些工作，然而他们所采用的排序策略与传统的网页搜索十分类似，无法很好地满足博客搜索用户的特定需求。

本文综合考虑博客文章的相关性、时效性、查询类型与博客作者兴趣特征的一致性等因素，提出一种新的排序算法以满足博客搜索用户的信息需求。采用经典文本分类方法完成查询分类工作，同时提出一种结合兴趣遗忘因子的用户建模算法描述博客作者的兴趣特征，相关性和时效性则依据向量空间模型和博客文章的时间信息进行度量。

2 相关工作

目前，针对博客搜索的工作还较少。在K. Fujimura等人的工作中，作者在考虑博客搜索需求的基础上提出了多面(multi-faceted)博客搜索技术，通过提供不同类型的服务满足不同的用户需求[1]。在之前的工作中，他们提出了一个名为“EigenRumor”的博客文章排序算法。该算法基于特征向量计算，通过对博客作者hub值和authority值的计算得出博客文章的权值，这种计算方式使得那些由较好的博客作者提交，但未被基于该用户前期工作认同的其他博客链接到的博客文章得到较高的分数[2]。

然而，上述算法均没有考虑博客搜索中的查询类型以及博客作者的兴趣特征。

与本文相关的查询类型识别工作是一种特定类型的文本分类任务，与传统文本分类任务所不同的是其分类对象是由少量词汇构成的词序列。目前在该识别任务上已有一些工作，大致可以分为 2 类：非监督学习分类和监督学习分类。在D. Beeferman等人提出的非监督学习分类工作中，通过对点击文章的分析来挖掘查询之间的潜在联系，从而将类似的查询进行聚类[3]。但这种无监督的聚类方法不仅需要付出昂贵的计算代价，而且对最终将获得的类别信息也一无所知。在监督学习方面，L. Gravano等人基于地理位置信息，采用了PIPPER对数线性回归和支持向量机等 3 种机器学习方法对查询分类进行了实验。D. Shen等人实现了基于支持向量机的分类方法，将查询归类到预先定义的 67 个类别[4]。本文也采用了支持向量机的方法实现查询分类。

另外的相关工作是个性化，尤其是基于使用的网络个性化(Usage-based Web Personalization)。网络个性化的目标是在用户显式提出请求之前提供给用户所需的相关信息[5]。典型的个性化过程由 5 个模块组成：用户建模，日志分析和网络使用挖掘，内容管理，网址发布和信息获得与搜索。本文主要关注用户建模，该模块显式或隐式地采集每位到访者的信息，包括用户的属性信息、兴趣特征甚至是其在浏览网站时

基金项目：浙江省科技厅科技专项和优先主题基金资助重大项目(2007C13050)

作者简介：卢 刚(1974－)，男，副教授、硕士，主研方向：网络安全，计算机视觉，嵌入式系统，计算机图形图像处理

收稿日期：2008-12-20 **E-mail：**hz_lugang@163.com

的行为特征。此前关于用户建模的工作大致可分为 2 类：基于知识的用户建模和基于行为的用户建模。基于知识的用户建模设计了若干静态的用户模型，然后将用户动态地匹配到最接近的模型中。该方法通常采用调查问卷和采访等形式来获得与该用户相关的信息。基于行为的方法采用用户的行为作为模型，并通常采用机器学习方法挖掘行为中的有用模式。本文的建模方法属于后者。

3 基于多特征的博客文章排序

本文提出的排序算法综合考虑 3 种特征：博客文章的相关性，时效性以及查询类型与博客作者兴趣特征的一致性。采用线性融合方法，将上述 3 种特征进行融合，博客文章的排序值按如下公式计算：

$$Score=\alpha\times\frac{documentRelevance}{documentFreshness}+\beta\times typeCoherence \quad (1)$$

其中，排序值与文章的相关性，查询类型与博客作者兴趣特征的一致性呈正比，与博客的时效性呈反比；α 和 β 两个参数通过实验估计得到，用于权衡不同因素的重要程度。下面对公式中的各个因素进行解释。

博客文章与查询之间的相关性*documentRelevance*由余弦相似度计算方法进行衡量。给定查询Q以及博客文章D，分别以N维向量$<q_1, q_2,\cdots, q_n>$和$<d_1, d_2,\cdots, d_n>$表示，则它们之间的相关性计算如下：

$$documentRelevance(Q,D)=\frac{\sum_{i=1}^{n}q_i d_i}{\sqrt{\sum_{i=1}^{n}q_i^2}\sqrt{\sum_{i=1}^{n}d_i^2}} \quad (2)$$

其中，q_i和d_i是查询和文章中对应词汇的权重，通过TF-IDF方法量化确定。

文章的时效性程度*documentFreshness*则通过文章发表时间与查询执行时间的时间差进行度量。G. Mishne等人对博客搜索引擎的日志进行研究，结果显示有大量的博客查询属于“近期查询”，即用户在搜索结果中选择相关文章时，更倾向于点击近期的文章[6]。因此，在排序策略中，将相对于查询时间较近的“近期”相关文章赋以较高的排序值会更合适。本文给定一篇发表于日期T_{post}的文章D，其“近期”程度，即时效性用下式表示：

$$documentFreshness(D)=\mathrm{lb}(2+T_{issue}-T_{post}) \quad (3)$$

其中，T_{issue}指查询执行日期，为防止对数真数为 0，加入常数 2 作为平滑。

查询类型与博客作者兴趣特征之间的一致性*typeCoherence* 在之前的博客搜索工作中均未涉及。本文考虑博客搜索的主要意图之一，即寻找拥有某些兴趣特征的博客作者，提出将此因素融入博客文章的排序中。然而，在实际搜索过程中，搜索用户只能通过查询表述他们的兴趣需求，因此，本文通过度量查询兴趣特征(即查询类型)与博客作者兴趣特征之间的一致性来衡量搜索用户与博客作者之间的兴趣一致性。

根据本文定义的类别体系(表 1)，查询类型和博客作者的兴趣特征均使用类别向量表示，其中，每个维度的值即为查询类型和博客作者的兴趣特征属于对应类别的概率。查询类型和博客作者兴趣的一致性可直接利用式(2)对上述 2 个向量进行余弦相似度计算得到。在衡量该值的过程中，如何获得查询类型和获得博客作者的兴趣特征是关键，下文将分别就这 2 个问题展开讨论。

表 1　本文定义的类别

类别	示例
IT	X60 价格
经济	微软 股价
健康	禽流感 传播途径
教育	高考
军事	歼 10
旅游	北京 自助游
运动	NBA 姚明
文化	功夫之王 评价
招聘	硬件工程师 杭州

3.1 查询分类

在上文提到，查询类型的识别是一种特定类型的文本分类任务。因此，在此项工作中可采用经典的文本分类技术，如朴素贝叶斯模型、支持向量机等。在综合考虑各分类算法的查准率、查全率之后，本文采用支持向量机完成分类工作(libsvm 工具包实现)。

分类器所需要的训练集采用搜狗实验室提供的网页分类数据集 SogouC。该数据集包括 17 910 个网页，其类别定义与表 1 中的分类定义相同，每一个类包含了 1 990 个网页。采用该数据集的原因是查询往往仅由少量词汇构成，不能完全反映出类别特征，因此，使用由查询组成的数据集训练得到的分类器存在“欠学习”的问题。而网页数据集则包含了足够多的词汇和分类特征，经由网页数据集训练所得的分类器将能更准确地完成分类工作。在分类器的训练阶段，每个类别选取了 1 330 个网页，整个训练集由 11 970 个网页构成。

3.2 博客作者建模

博客系统中的博客文章是能反映博客作者兴趣特征的直接来源之一，且易获得，不存在侵犯用户隐私等问题。本文基于对博客文章内容的分析，提出一种博客作者建模方法，具体如下：

博客作者发表的文章被分类到表 1 所定义的类别中，分类所需要的分类器直接采用查询分类时构造的分类器。给定一篇文章 $\boldsymbol{P}$ 和类别集合$\{C\}$，$\boldsymbol{P}$ 被表示为向量：

$$\boldsymbol{P}=<w_1, w_2,\cdots, w_n> \quad (4)$$

其中，$w_j, j=1,2,\cdots,n$表示文章$\boldsymbol{P}$属于类别C_j的概率。

假定该作者在某一时刻T_i发表的文章总数为t，本文引入兴趣矩阵$\boldsymbol{IM}$对他的兴趣特征进行量化描述：

$$\boldsymbol{IM}=\begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1j} & \cdots & m_{1n} \\ m_{21} & m_{22} & \cdots & m_{2j} & \cdots & m_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ m_{i1} & m_{i2} & \cdots & m_{ij} & \cdots & m_{in} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ m_{t1} & m_{t2} & \cdots & m_{tj} & \cdots & m_{tn} \end{bmatrix} \quad (5)$$

其中，矩阵的行i表示某个时刻T_i；列j代表博客作者对相应类别C_j的兴趣特征，因此，矩阵中的值m_{ij}表示的是在时刻T_i对类别C_j的兴趣，即该时刻发表的文章属于类别C_j的概率。

最后，给定一个具体描述了在不同时刻用户兴趣的兴趣矩阵 $\boldsymbol{IM}$，本文提出 2 种方法对该用户的总体兴趣特征进行建模。第 1 种简单的建模方法就是仅仅将矩阵的所有行相加，用所得的结果向量对用户总体的兴趣特征进行量化描述，该向量即为用户的兴趣特征模型 M，其中，每项维度值的计算公式如下：

$$M_j=\sum_{i=1}^{t}m_{ij} \quad (6)$$

但考虑到人记忆的自然遗忘规律，本文引入衰减因子的

概念，即博客文章对用户总体兴趣特征的权值将随着时间的流逝逐渐变小。由此，本文提出第 2 种建模方法，即带有记忆衰减因子的博客作者兴趣建模方法。衰减因子$F(t_k)$用于描述发表于时刻t_k的文章的衰减程度，其公式表述如下：

$$F(t_k)=\mathrm{e}^{\frac{\mathrm{lb}(t_0-t_k)}{hl}} \tag{7}$$

其中，t_0表示当前时刻；hl表示半衰期，即经过hl天后用户的兴趣遗忘一半。原始的用户兴趣模型M修正如下：

$$M_j^{'}=\sum_{i=1}^{t}F(t_i)m_{ij} \tag{8}$$

本文在引入了记忆衰减因子后，模型 M'能比模型 M 更准确地描述博客作者的兴趣特征。

4 实验与结果讨论

为比较本文提出的排序算法与传统算法的性能优劣，本文实现了一个基于上述算法的博客检索原型系统，邀请志愿者对检索结果进行比较。

4.1 实验策略

由于没有标准的测试语料库，因此对本文所提出的排序算法进行性能评估就显得比较困难，查准率、查全率和 $F1$ 值等信息检索和分类中常见的评价标准不再适用。因此，本文采用对比完成具体任务所需要的时间的方式来评估算法的性能。

本文实验采取的策略为：给定若干任务，邀请一定数量的志愿者通过搜索来寻找相应的答案，完成任务。全体志愿者均分成 2 组：一组使用本文设计的系统，另一组则使用 Google 博客搜索。所有结果将以统一格式输出，确保系统对志愿者的透明性，以保证实验的公正性。随后，记录志愿者完成任务的时间，每组完成一项任务的平均完成时间作为衡量算法性能的依据。志愿者返回的答案的正确性由另外 2 位评判人员判定。

需要说明的是，本实验中的原型系统不可能抓取与 Google 相同级别的博客数据，因此，本文的系统是建立在 Google 的搜索结果之上的。在实验中，预先设置好每个任务的查询（志愿者只能使用这些规定的查询进行检索），然后将 Google 对这些查询的搜索结果作为博客数据保存起来，随后以每个博客作者为单位抓取该作者所有的文章，基于此对博客作者进行兴趣建模。这一策略保证了实验对比是在相同的环境下进行。另一个需要关注的问题是志愿者有可能在搜索答案的过程中出现错误，在这种情况下，小组完成某任务的平均完成时间将在原平均完成时间的基础上加上该小组成员中完成该任务时最长的消耗时间，以此作为惩罚。

实验所设定的任务包括：寻找一位对运动感兴趣的博客作者；寻找一篇包含对 ipod 评述的博客文章；寻找一篇包含对 ipod 描述的博客文章；寻找一篇最近的包含对 ipod 评述的博客文章。评述与描述的不同在于前者是带有个人主观色彩的评论而后者只是对事物的客观介绍。

4.2 参数估计和模型选择

在本排序算法中有 2 个参数 α 和 β 需要估计，实验中设两者之和为 1，因而只需考虑一个参数的估计即可(实验中考虑了 α)。为评估 α 值在 0~1 之间变化时算法性能的变化，实验采用上述相似的策略，邀请志愿者使用采用不同参数的实验原型系统来完成上述的 4 项任务；与此同时，实验也比较了在不同兴趣模型下算法的性能表现。图 1 是针对任务 1 的结果。

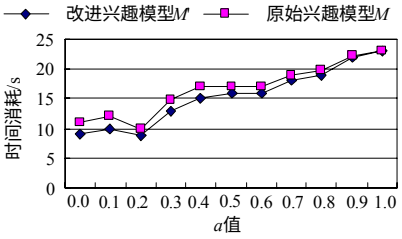

图 1　不同兴趣模型和不同 α 值情况下完成任务 1 的时间消耗

通过对图 1 的观察可知，改进的兴趣模型 M'的性能优于原始的兴趣模型 M，这说明考虑兴趣衰减是有效的。随着 α 的递增，2 条曲线开始合并，这是因为随着博客作者兴趣在整个排序算法中的权重的减少，算法开始更多地受文章相关性的影响，这使得用户兴趣衰减对整体性能的影响减弱。在图 1 中可以观察到，当 α 取 0.2 时，2 条折线均达到了最小值，此时的查询类型和兴趣模型之间的一致性是排序算法中的主要影响因素。这也表明，在博客文章排序中，查询类型和博客作者的兴趣模型的一致性要比文章的相关性、时效性更为重要。

实验中其他 3 项任务的结果同样也表明改进的兴趣模型 M'较原始的兴趣模型更适合博客排序，但是 4 项任务的时间消耗最小值出现时的参数 α 的值各不相同，之所以出现这种情况，是因为每一项任务所强调的博客特征各不相同。例如，任务 1 强调了博客作者的兴趣爱好，而任务 4 则强调了时效性。本文为将上述 4 个不同类型的任务纳入到同一框架中，最终取了每个任务中最小平均时间消耗出现时所对应的 4 个 α 值的平均值作为 α 的最终赋值。

4.3 实验结果与讨论

图 2 为使用不同搜索引擎的 2 个小组在完成每项任务时的平均时间消耗。实验结果表明，在任务 1、任务 2 和任务 4 中，采用本文算法的时间消耗要低于 Google 博客搜索服务。原因在于这些任务都是与博客的特征密切相关的，而 Google 博客搜索并未将此类因素考虑在内。尤其是在任务 2 和任务 4 中，针对一件产品的个人评述更有可能出现在酷爱此类产品的用户博客文章中，而本文的算法在设计之初就考虑了这个因素，因此在完成任务 2 和任务 4 时表现较好。此外，任务 4 的优异表现同样表明了算法考虑文章时效性是有效的。实验结果中唯一的例外是在任务 3 中，本文的算法略逊于 Google，这是因为本算法所采用的相关性度量方式过于简单。不过，这类搜索任务在博客搜索中仅占所有搜索任务的一小部分，因此对搜索服务的整体性能表现影响有限。

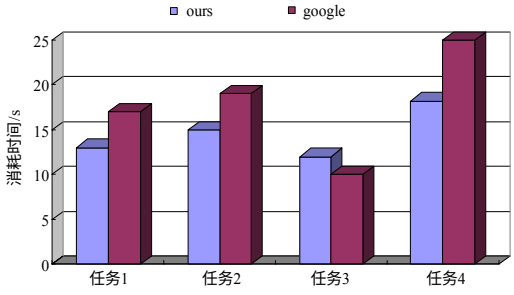

图 2　本文算法与 Google 博客搜索服务在完成不同任务时的时间消耗对比

5 结束语

在博客搜索中，由于博客系统结构和博客搜索用户的信

(下转第 52 页)