软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(Suppl.(2)):150−161

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于分析特征与动态步长的微博排序学习算法*

周诗龙, 徐俊刚

(中国科学院大学 计算机与控制学院,北京 100190)

通讯作者: 周诗龙, E-mail: kunlong0909@163.com

摘 要: 目前,微博搜索大多应用向量空间模型计算查询词与文档间的相关程度,通常使用 TF-IDF(term frequency-inverse document frequency)统计方法来确定词的权重.然而仅使用词进行微博搜索并不能检测到某条微博的信息含量,而这些往往是查询用户所关注的问题.为此提出了一种基于分析特征与动态步长的微博排序学习算法.首先,定义了一些微博分析特征,经过统计分析获得的这些分析特征可以用来预测用户行为;其次,在此基础上,提出了以词性为单位计算微博相关度的方法,结合信息熵计算方法得到微博词性信息的含量,并用来预测该微博的信息含量;最后,在现有 ListNet 排序学习算法的基础上,引入了动态步长的概念,对步长进行了动态优化,最终形成了一种基于动态步长的微博排序学习算法——RDLS(ranking based on dynamic learning stepsize)算法.实验结果表明,无论是基于直接特征还是加入分析特征,在相同迭代轮数情况下,相比 ListNet 算法,RDLS 算法可以训练出更优的模型,在微博排序方面有更好的表现.

关键词: 微博;ListNet;动态步长;分析特征;排序学习

中文引用格式: 周诗龙,徐俊刚.基于分析特征与动态步长的微博排序学习算法.软件学报,2013,24(Suppl.(2)):150−161. http://www.jos.org.cn/1000-9825/13033.htm

英文引用格式: Zhou SL, Xu JG. Learning to rank algorithm for microblogs based on analysis features and dynamic stepsize. Ruan Jian Xue Bao/Journal of Software, 2013,24(Suppl.(2)):150−161 (in Chinese). http://www.jos.org.cn/1000-9825/13033.htm

Learning to Rank Algorithm for Microblogs Based on Analysis Features and Dynamic Stepsize

ZHOU Shi-Long, XU Jun-Gang

(School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100190, China)

Corresponding author: ZHOU Shi-Long, E-mail: kunlong0909@163.com

Abstract: Currently, most of searching methods for microblog use vector space model to calculate the relevance between the query and document. The statistical method of Term Frequency-Inverse Document Frequency (TF-IDF) is widely used to determine the weight of words. However, only using word as the unit of microblog searching is not enough to detect the whole information content of a microblog, which is usually the intent of the search users. To solve this problem, a learning to rank algorithm for microblogs based on analysis features and dynamic stepsize is proposed. Firstly, some analysis features for microblogs are defined, The features can be obtained through statistical analysis method, and used to predict user's behaviors. Secondly, a method to calculate the relevance of microblogs based on part of speech is proposed. It uses the strategy of information entropy to calculate POS information content of microblog and it can be used to predict the information content of the microblog. Finally, based on the existing ListNet algorithm, the concept of dynamic stepsize is introduced to optimize the calculation of stepsize, eventually a learning to rank algorithm for microblogs based on dynamic stepsize named Ranking based on Dynamic Learning Stepsize (RDLS) algorithm is formulated. The experimental results show that RDLS algorithm can get a more optimal training model by using either direct features or both direct and analysis features with the same iterations, and can attain better effect in microblog ranking compared with the ListNet algorithm.

* 基金项目: 国家科技支撑计划(2012BAH23B03)
收稿时间: 2013-03-15; 定稿时间: 2013-07-11

Key words: microblog; ListNet; dynamic stepsize; analysis feature; learning to rank

近年来,作为一种方便、快捷的社交网络形式,微博服务得到了快速的发展.当前国外最大的微博服务提供商是 Twitter,巴黎分析公司 Semiocast 发布消息称截至 2012 年 7 月 1 日,Twitter 用户总数为 5.17 亿[1].同时,国内微博服务也开始兴起,目前主流的互联网门户都提供了微博服务,包括新浪、腾讯和搜狐等,其中最典型的代表是新浪微博[2].新浪公司总裁曹国伟称,截止 2012 年 12 月 31 日,新浪微博的注册用户达到 5.03 亿.2012 年 12 月,日均活跃用户数在 4620 万左右[3].在微博平台上,任何用户可以就任何话题发布任何消息,微博用户还可以追随其他用户,或者向追随者推荐微博,使得微博平台成为一个巨大的社会化网络.

为了方便微博用户从海量动态更新的微博中找到感兴趣的内容,Twitter 提供了微博搜索服务.与传统的搜索引擎类似,Twitter 用户输入关键字进行搜索,就可以得到按照时间排序的微博列表[4].目前,国内主要是新浪推出了微博搜索服务,但是仅针对注册用户.

当前对微博排序算法有诸多研究.Meij 等人提出了一种基于语义链接(semantic linking)的方法[5];Efron 研究了基于标签(Hashtag)的微博搜索方法[6];Zhao 等人探索了一种加权多元素排序算法[7].上述微博排序算法都有较好的性能表现,但是仍然存在两方面不足:一是提取的特征没有创新,多是基于传统 Web 的方法提取的特征如微博长度[4]、TF 值[5]等,这些特征在研究中都具有良好的表现[4,5],但是对于微博排序发挥的作用有限,微博具有与 Web 网页不同的特性[8],需要探索针对微博自身特性的特征.另一方面,很少应用基于列表学习的排序学习算法,因为其对多特征微博海量数据搜索排序的适应性和训练出模型的良好性能,还没有得到验证.

现有的微博搜索应用存在很多不足,只提供了基于关键词的搜索接口,无法满足用户快速从海量微博中获取感兴趣和有价值信息的要求.从更深层角度分析,我们总结有两方面原因:一是因为有些微博搜索引擎是通过进行全关键词匹配方式提升微博匹配度,这样会由于关键词匹配不当而导致极少微博信息返回或者返回海量微博数据;另一方面,因为微博较短,而且用户写微博具有一定的随意性,不同微博的信息含量大相径庭,当前搜索算法返回的列表显示的搜索结果,用户感兴趣的微博没有能够被放在靠前的位置,因而给用户带来不便.

当前的微博搜索基本上是传统网页搜索的简单克隆,缺乏对微博本身特征的深度挖掘并应用合适的微博排序算法.本文针对这一问题,将微博中的特征根据分析的深度分为直接特征和分析特征,并研究分析了分析特征中词性信息量的计算方式;将排序学习(learning to rank)的算法应用在微博排序中,并对排序算法进行改进.本文主要贡献如下:

(1) 定义并提取了微博中的直接特征.基于当前微博排序研究以及对微博的进一步观察提取了一些直接特征,每一个直接特征都会在一定程度上影响排序结果.

(2) 定义并提取了微博中的分析特征.分析特征是隐含在微博中的与搜索结果相关度有关系的特征,需要通过统计分析和计算得到,每一个分析特征对排序结果也会产生重要影响.

(3) 将文档排序算法 ListNet 应用到微博搜索中,并进行了改进.针对 ListNet 算法存在的问题,创新性地引入动态步长策略,使其经过很少的迭代轮数即可训练出高精度的模型.

1 相关工作

1.1 微博特征提取

当前的研究中,对 twitter 和中文微博的特征提取主要在于直接特征提取.用户的权威度[7]可以作为微博的特征之一,当前大部分的研究通过用户的粉丝数量来判断用户的权威度,用户拥有更多的粉丝就代表该用户更加受到人们的重视,更加受到人们的欢迎.另一种计算方式是根据粉丝数量和好友数量之间的比率来判断用户的权威度.Kwak 等人研究了 Twitter 的拓补结构和作为一个新平台的的影响力[9].为了区分 Twitter 上的不同影响力,他们根据用户粉丝的数量并利用 PageRank 算法对用户进行排序.他们发现这两种计算方式是相似的,然而转发数量的结果与这两种算法又不相同,这就意味着单纯的粉丝数量不是用户影响力的唯一决定因素.

Cha 等人探索了 3 种影响因素:入度、转发数和提及数[10].入度影响因素指的是一个用户的粉丝数量,转发

数因素就是微博的被转发数量,提及数是提到某人的名字的次数.

此外,在研究中涉及到的直接特征还有微博的词项长度[5]、微博中是否包括@[8]、微博中是否包括标签#[11]、用户发博文数量等.这些直接特征对于提取历史微博数据来讲作用是很大的,但是这些特征还不够,为了获取更加满意的微博排序结果,需要定义更多的直接特征.同时微博是一个实时性非常强的平台,在短时间内一篇信息量大并引起用户兴趣微博的转发量和评论数量都不会达到很大的规模.针对这一问题,我们提出了微博分析特征的概念,它通过数据统计方法来计算微博信息含量的大小.

1.2 排序学习算法

目前,信息检索常用的排序模型有BM25[12],语言模型(language model,简称LM)[13]和PageRank[14]等.这些模型都包含多个参数,对于测试未知查询时会出现过度拟合的问题.而机器学习方法在自动调整参数和避免过度拟合方面具有一定的优势,因此可以利用机器学习技术自动地建立有效的排序模型[15],这类方法通常称为排序学习方法.

现有的排序学习方法分为3类:点方法(point-wise)、对方法(pair-wise)、列表方法(list-wise).点方法采用一个查询对应的单一文档作为训练样例,而不考虑此文档与该查询对应的其他文档之间的关系.为了改进这一点,对方法被提出来.对方法采用查询对应的文档对作为训练样例,模型训练的目的是正确的找到输入实体对等级的差异.该方法训练出的模型比较复杂,训练时间较长,为此提出了列表方法.列表方法采用查询对应的文档序列作为训练样本,理论上讲这种方法能够获得最优的排序性能,因为对训练数据的优化目标完全符合人们对检索系统结果优劣的判断.

排序学习的最大优势是能够集合大量的特征和判别训练.而商业搜索引擎急需这样的能力,因为只用少量的特征无法满足网络用户复杂多变的搜索需求,而且判别训练能够从用户的这些反馈信息中进行学习以改进排序机制.因此,本文我们将排序学习方法应用到微博搜索中,应用微博特有的特征,将会获得更好的排序结果,满足微博搜索用户的搜索要求.

1.3 ListNet算法

ListNet算法[16]是由Cao等人提出的基于列表方法的一种典型的排序学习方法,它将损失函数构造为用户衡量预测的文档序列与实际最好的文档序列之间的差异,并基于神经网络,利用梯度下降原理进行模型建立和应用.文献[16]中的实验表明,ListNet算法比之前提出的基于对方法的RankNet,Ranking SVM和RankBoost算法具有更好的表现.

ListNet使用两个概率模型来计算损失函数,这两个模型分别是序列概率(permutation probability)和top k概率(top k probability).

序列概率假设已经排好序的样本集被标上序号1,2,3,…,n.这些样本的一个序列π定义为{1,2,3,…,n}对它自己的双射.序列π可以表示为$\pi=\langle\pi(1),\pi(2),...,\pi(n)\rangle$.在这里,$\pi(j)$表示序列中排在第$j$位的样本.$n$个样本所有可能序列的集合表示成$\Omega_n$.

假设π是一个具有n个样本的序列,$\Phi(\cdot)$是一个递增并且严格正的函数.则对于给定的列表得分s的序列π的概率定义为

$$p_s(\pi)=\prod_{j=1}^{n}\frac{\Phi(s_{\pi(j)})}{\sum_{k=j}^{n}\Phi(s_{\pi(k)})} \tag{1}$$

其中$s_{\pi(j)}$表示样本在序列π第j个位置上的得分.

如果只是应用序列概率,则复杂性会达到$O(n!)$,这在实际应用中不具有实用性.为了解决这个问题,Cao等人在文献[16]中采用了top k概率.样本$(j_1,j_2,\ldots,j_k)$的top k概率代表的是给出所有样本的得分时,它们排在第k位的概率.top k概率计算公式如下:

$$p_s(\vartheta_k(j_1,j_2,...,j_k))=\prod_{j=1}^{k}\frac{\Phi(s_{\pi(j)})}{\sum_{k=j}^{n}\Phi(s_{\pi(k)})} \tag{2}$$

其中 $\vartheta_k(j_1,j_2,...,j_k)$ 包含了前 k 个对象的所有的排序可能,$\vartheta_k(j_1,j_2,...,j_k)=\{\pi\in\Omega_n\mid\pi(t)=j_t,\forall t=1,2,...,k\}$.在集合 ϑ_k 中共有 $\frac{n!}{(n-k)!}$ 个元素,这比 Ω_n 中的元素数目要少的多.

对于给定的查询 $q^{(i)}$,排序函数 f_ω 能够生成一个得分列表 $z^{(i)}(f_\omega)=(f_\omega(x_1^{(i)}),f_\omega(x_2^{(i)}),...,f_\omega(x_{n^{(i)}}^{(i)}))$,$x_k^{(i)}$ 表示第 i 个查询的第 k 个特征得分.然后文档 $(d_{j_1}^{(i)},d_{j_2}^{(i)},...,d_{j_k}^{(i)})$ 的 top k 概率可以计算:

$$P_{z^{(i)}(f_\omega)}(\vartheta(j_1,j_2,...,j_k))=\prod_{i=1}^{k}\frac{\exp(f_\omega(x_{j_t}^{(i)}))}{\sum_{l=t}^{n^{(i)}}\exp(f_\omega(x_{j_t}^{(i)}))} \tag{3}$$

加入交叉熵 $L(y^{(i)},z^{(i)}(f_\omega))$ 关于 ω 的导数计算公式导出式(4):

$$\Delta\omega=\frac{\partial L(y^{(i)},z^{(i)}(f_\omega))}{\partial\omega}=-\sum_{\forall g\in\vartheta_k}\frac{\partial P_{z^{(i)}(f_\omega)}(g)}{\partial\omega}\frac{P_{y^{(i)}}(g)}{P_{z^{(i)}(f_\omega)}(g)} \tag{4}$$

2 微博特征提取

本文将微博特征分为直接特征和分析特征,这里对提取的两类特征进行定义和简要说明.两者是根据获取方式不同进行分类的,但是起到的作用是相同的.

2.1 直接特征提取

直接特征指的是直接能够从微博中获取到,或者只是经过简单的计算能够得到的特征.当前已有的研究中提升排序结果准确率只是关注算法的改进而没有注意提升提取特征的特性,因而提取到的特征大部分为直接特征.

综合现有文献[5−11]提到的直接特征,我们结合中文微博特点进行了整理,总结并补充得到微博以下一些直接特征.

(1) 用户的权威度[7]

用户的权威度根据粉丝数量和好友数量之间的比率来判断用户的权威度.对于明星用户来说,粉丝数量会较多而好友数量会较少,这样其权威度计算值就会较大.

(2) 微博的转发数量[10]

如果微博内容有人转发,则说明该微博具有较高的兴趣度或者信息中含有让大家感兴趣的内容.该篇微博被转发的数量越多,代表对其感兴趣的用户越多.

(3) 搜索的关键词是否是话题[11]

若返回微博是一篇话题微博且关键词出现在话题里,则该微博内容是针对相应关键词的评论信息,具有一定的相关性.据 Miles Efron 的实验[6]结果表明,话题关键词搜索比普通词项搜索具有更好的性能.

(4) 用户互相关注的数量[11]

通过用户互相关注的数量,能够看出该用户的好友数量,如果互相关注的人数比较多,则从一定程度上可以说明该用户微博具有一定的影响力.与仅拥有相同数量单向关注用户的用户来说,该用户所发微博更能引起他人的注意.

(5) 微博发布的时间[11]

搜索信息具有一定时效性.人们搜索微博的时候都希望先看到最新的消息,然后再看之前的消息.

(6) 微博的长度[5]

非常短的微博被认为是有歧义和无重点的.所以较长微博的相关信息含量大的概率要高一些.

(7) 指定用户的数量(即@用户的数量)[8]

如果一篇微博中有@标签,则说明博主是想让指定的人看到该微博并进行相应的评论.则该微博对大众和@用户来说,重要程度就不一样.对大众来说,该类微博的兴趣度会较低,而对@用户来说,该微博的兴趣度会较高.

(8) 用户是否是加 V 用户

用户加 V 表示该用户已经被新浪认证,经过认证的用户在发信息的时候会更加注意自己的形象,因而微博内容会更加认真和让人感兴趣一些.

(9) 微博内容中包含的表情数量

微博内容中包含的表情数量会影响微博的信息含量.有人发微博只发表情,有人发微薄在文字中夹杂表情,这些会影响到微博信息对大众的兴趣度.

(10) 关键词在微博中的位置

美国的 EE.Baxendale 的调查结果显示:段落的论题是段落首句的概率为 85%,是段落末句的概率为 7%.因此,有必要提高处于微博特殊位置的句子权重,以此来区分不同相关度微博.

当然,上述直接特征还是不完整的,这需要今后根据不同的用户需求不断地补充和完善.

2.2 分析特征提取

分析特征指的是需要采用适当的模型或者统计分析方法来获得的特征.这些特征对于微博搜索结果具有更加重要的影响,这里,我们定义了以下一些分析特征:

(1) 微博词性(Part of Speech,简称 POS)信息量

汉语言中,能标识文本特性的往往是文本中的实词,如名词、动词、形容词等.而文本中的一些虚词,如感叹词、介词、连词等,对于标识文本的类别特性并没有贡献,也就是对确定文本类别没有意义的词.在提取文本特征时,应首先考虑剔除对文本分类没有用处的虚词,这可以通过停用词去除来实现,而在实词中,又以名词和动词对于文本的类别特性的表现力最强,所以可以对名词和动词进行细化分类,引入二级分类,统计计算得到词性的权重 $(w_{pos(1)}, w_{pos(2)},..., w_{pos(m)})$.微博词性得分计算公式如下:

$$Score_{pos} = n_{pos(1)}w_{pos(1)} + n_{pos(2)}w_{pos(2)}...+n_{pos(m)}w_{pos(m)} \tag{5}$$

(2) 微博的分析长度

比较短的微博被认为是有歧义和无重点的.我们认为微博内容达到平均长度是比较合适的,微博长度得分 $Score_{mi}$ 可以计算如下:

$$Score_{mi} = \frac{1}{1+\left|avg_{mi} - tl_{mi}\right|} \tag{6}$$

(3) 语言模型

语言模型[17]最早是由 Ponte 和 Croft 提出,它是信息检索中一种非常通用的形式化方法,它在实现时有多个变形.信息检索中最早使用也是最基本的语言模型是查询似然模型(query likelihood,简称 QL).在这个模型中,需要对微博集中的每篇微博构建对应的语言模型 M_d,然后将微博按照其与查询相关的似然 $P(d \mid q)$ 来进行排序.使用线性插值平滑方法的似然函数可以表示如下:

$$P(d \mid q) \propto P(d)\prod[(1-\lambda)P(t \mid M_c) + \lambda P(t \mid M_d)] \tag{7}$$

这里,我们给出了 3 个微博分析特征,今后可以根据需要再进行补充和添加.

3 基于动态步长的微博排序学习算法

ListNet 算法的步长是给定的,利用的是梯度下降法,这样会使得训练收敛速度较慢.本文中创新性地引入 Armijo 步长准则[18]来动态计算梯度下降的步长,在相同的迭代轮数下,本算法可以训练出更好的模型.为了方便表示,我们将改进后的 ListNet 算法称为 RDLS(ranking based on dynamic learning stepsize)算法.

RDLS 算法的训练数据集是一个查询集合,每一个查询 $q^{(i)}$ 都有一个相关的文档列表

$$d^{(i)}=\{d_1^{(i)},d_2^{(i)},...,d_{n^{(i)}}^{(i)}\},$$

其中,$d_j^{(i)}$ 表示第 i 个查询的第 j 个文档,$n^{(i)}$ 表示第 i 个查询返回的文档个数.每一个文档列表 $d^{(i)}$ 都与一个标记得分集合 $y^{(i)}=\{y_1^{(i)},y_2^{(i)},...,y_{n^{(i)}}^{(i)}\}$ 相关联,即每个文档 $d_j^{(i)}$ 都有一个对应的相关度标记得分 $y_j^{(i)}$.

每个查询文档对 $(q^{(i)},d_j^{(i)})$ 都有一个特征向量 $x_j^{(i)}=\psi(q^{(i)},d_j^{(i)}),i=1,2,...,m;j=1,2,...,n^{(i)}$,每个文档列表的特征 $x^{(i)}=\{x_1^{(i)},x_2^{(i)},...,x_{n^{(i)}}^{(i)}\}$ 与对应的相关度标记得分列表 $y^{(i)}=\{y_1^{(i)},y_2^{(i)},...,y_{n^{(i)}}^{(i)}\}$ 形成一个“实例”,因此训练集可以写成 $\tau=\{(x^{(i)},y^{(i)})\}_{i=1}^{m}$.然后创建一个排序函数 f,对于每个特征向量 $x_j^{(i)}$ (对应于文档 $d_j^{(i)}$),它输出一个得分 $f(x_j^{(i)})$.对于一个列表的特征向量 $x^{(i)}$,得到一个列表的得分 $z^{(i)}=(f(x_1^{(i)}),...,f(x_{n^{(i)}}^{(i)}))$.

Armijo 准则是许多非线性规划(nonlinear programming)算法求步长时都必须执行的步骤.Armijo 准则是指给定 $\beta\in(0,1),\sigma\in(0,0.5)$,令步长因子 $\alpha_k=\beta^{m_k}$,其中,m_k 是满足下列不等式的最小非负整数:

$$f(x_k+\beta^m d_k)\leqslant f(x_k)+\sigma\beta^m g_k^T d_k \tag{8}$$

g_k 是函数 $f(x)$ 在当前迭代点 x_k 处的梯度函数,d_k 是当前迭代点 x_k 处的搜索方向,可以证明 $f(x)$ 若是连续可微的且满足 $g_k^T d_k<0$,则准则是有限终止的,即存在正数 σ,使得对于充分大的正整数 m,上式成立.

由此结合 ListNet 算法,我们可以得到权重更新公式:

$$\begin{aligned}\omega_{k+1}&=\omega_k+\alpha_k\Delta\omega\\&=\omega_k-\alpha_k\sum\frac{\partial P_{z^{(i)}(f_\omega)}(g)}{\partial\omega}\frac{P_{y^{(i)}}(g)}{P_{z^{(i)}(f_\omega)}(g)}\end{aligned} \tag{9}$$

RDLS 算法的伪代码见算法 1.

算法 1. RDLS 算法.

输入:训练数据 $\{(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),...,(x^{(m)},y^{(m)})\}$.

参数:迭代次数 T 和 $\beta\in(0,1),\sigma\in(0,0.5)$

初始化 ω

for t=1 to T do

for i=1 to m do

输入查询 $q(i)$的 $x(i)$到神经网络,并用当前的ω来计算得分列表;

用公式(4)来计算梯度 $\Delta\omega$

利用 Armijo 算法计算步长 α_k

更新 $\omega_{k+1}=\omega_k+\alpha_k\Delta\omega$

end for

end for

输出神经网络模型

4 实 验

4.1 实验环境

实验平台为一台 Dell 个人电脑,具体配置为:Intel Core i5 CPU,4G 内存,32 位 Windows 7 操作系统.微博分词采用了张华平博士 2013 年最新推出的 NLPIR(natural language processing & information retrieval)中文分词工具.RDLS 算法基于 RankLib 库实现,并使用了它的评价函数.实验中的其他处理,包括数据的爬取、处理、数据标记工具的创建,都由作者开发完成.

4.2 实验数据集

本文研究针对的对象是中文微博,首先利用新浪微博搜索功能对关键词进行搜索,爬写每个搜索的前 100 条返回微博的 ID.再利用新浪提供的 API 爬取详细的微博和用户数据.实验数据集概况见表 1.

表 1 实验数据集

搜索时间	微博数量(条)	查询数目(个)	去重
2013-03-01 22时	2000	20	是
2013-03-07 16时	6200	62	是

相关度标记方法有多种,一种熟知的方法是二类标记方法,分为相关和不相关两种.然而,在实际搜索引擎中,对于用户的信息需求,文档相关度不能简单的分为两类[19].我们将相关度标记分为 5 个等级,分别为:好、较好、一般、较差、极差.为了能够让标记者尽量客观评价每条微博,作者开发了微博标记软件,在用户标记的时候,微博分类显示的信息较全,包括颜色区分和图片信息等.为了不误导用户,特意将微博数据进行混序操作,即打乱微博的顺序.

爬取的微博数据经过预处理,生成用于训练的数据格式,见表 2.

表 2 预处理后的微博数据格式

⟨line⟩ .=. ⟨relevance⟩ qid: ⟨qid⟩ ⟨feature⟩:⟨value⟩ ... ⟨feature⟩:⟨value⟩
⟨relevance⟩ .=. 0 | 1 | 2 | 3 | 4
⟨qid⟩ .=. ⟨positive integer⟩
⟨feature⟩ .=. ⟨positive integer⟩
⟨value⟩ .=. ⟨float⟩

4.3 评价指标

搜索结果的评价指标有很多,较为常见的有 P@n(Precision at Position),MAP(mean average precison)、NDCG(normalized discount cumulative gain),ERR(expected reciprocal rank).这里采用 NDCG[20,21]和 ERR[22]作为评价指标,因为 NDCG 和 ERR 适合计算多相关度问题[23],而 P@n 和 MAP 是为二相关度问题设计的.

NDCG 对传统的评价标准进行了改进,高度相关的文档比部分相关的文档更有价值,在评价中应该赋予更大的权值.微博在搜索结果序列中的位置越靠后,这篇微博的价值越小,从用户的角度考虑,由于时间、精力以及已经从阅读过的微博得到了信息等原因,用户可能根本不会去看排序靠后的微博.给定微博搜索结果列表的 DCG 值,计算公式为

$$DCG@T=\sum_{i=1}^{T}\frac{2^{l_i}-1}{\log(1+i)} \tag{10}$$

其中 T 为截断水平(例如:如果我们只关注返回结果的前 10 篇微博,则我们会取 T=10), l_i 是第 i 篇微博的标记得分.这里取值为 $l_i\in\{0,1,2,3,4\}$.NDCG 是标准化的 DCG,计算公式为

$$NDCG@T=\frac{DCG@T}{\max DCG@T} \tag{11}$$

ERR 与 NDCG 不同.NDCG 具有潜在的条件独立性假设,而 ERR 在计算用户对当前文档的满意度时,则需要考虑之前文档的相关性以及用户找到该相关文档的努力[24].ERR 的计算公式如公式(12)[22]所示.

$$\begin{aligned}ERR&=\sum_{i=1}^{n}\frac{1}{i}P(\text{user stops at } i)\\&=\sum_{i=1}^{n}\frac{1}{i}R(y_i)\prod_{j=1}^{i-1}(1-R(y_j))\end{aligned} \tag{12}$$

其中 $R(y)=\frac{2^{y}-1}{2^{y_{\max}}}, y\in\{0,...,y_{\max}\}$,由此可见,随着 n 值增大,ERR 值变化逐渐变小,ERR 曲线趋于平缓.

4.4 基于直接特征的微博排序实验

本实验通过提取微博的直接特征,计算每篇微博的直接特征得分,用 ListNet 算法和 RDLS 算法进行训练和

验证.ListNet 算法和 RDLS 算法的参数设置见表 3,其中经过多次实验发现,当β取 0.2,σ取 0.5 时,RDLS 算法会得到较好的模型.

表 3 ListNet 算法和 RDLS 算法参数设置

设置	值
训练的轮数 T	1 000
数据标准化方式	归 1 标准化
优化训练数据的标准	NDCG@n
测试训练数据的标准	NDCG@n
β(RDLS)	0.2
σ(RDLS)	0.5
训练数据数量	3 500
验证数据数量	1 400
测试数据数量	3 300

分别计算新浪微博搜索、ListNet 算法训练模型、RDLS 算法训练模型的 NDCG 和 ERR 值,结果如图 1 和图 2 所示,图中横轴为取出的微博排序返回结果的位置,纵轴分别为 NDCG 和 ERR 值.由于微博本身是短文本,用户浏览查询结果比较快,浏览结果会比较多,因而在本实验中计算了前 70 条结果的 NDCG 值和 ERR 值.

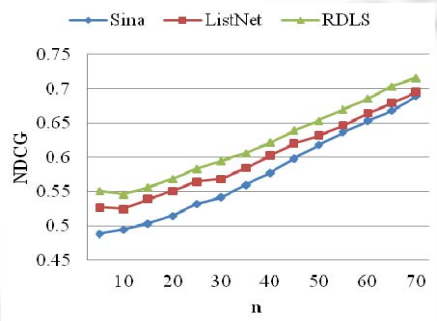


图 1 3 种基于直接特征的训练模型 NDCG 对比

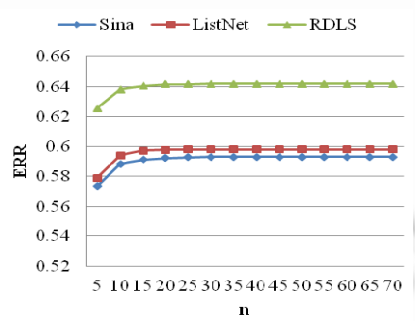
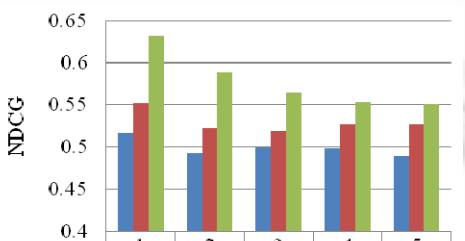


图 2 3 种基于直接特征的训练模型 ERR 值对比

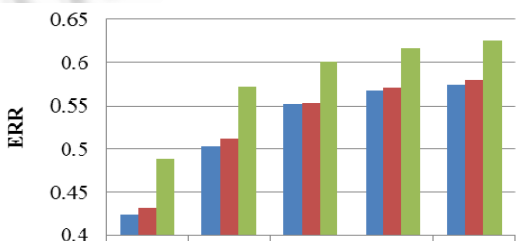
通过图 1 和图 2 可以看出 ListNet 模型和 RDLS 模型都取得比新浪微博搜索算法更好的结果.RDLS 模型也明显优于 ListNet 模型.

排序算法需要将最相关的微博排在最前面.图 3 和图 4 显示了前 5 个微博的 NDCG 值和 ERR 值.由此可见,在基于直接特征训练的模型下,ListNet 算法和 RDLS 算法在测试集上前 5 个微博的排序情况都比新浪本身算法要好.而且在训练相同轮数下,RDLS 算法训练的模型比 ListNet 算法训练的模型具有更好的表现.



	1	2	3	4	5
Sina	0.5172	0.4929	0.4991	0.4981	0.4888
ListNet	0.5518	0.5227	0.5189	0.5274	0.5272
RDLS	0.632	0.5891	0.5648	0.5533	0.551

图 3 3 个基于直接特征的训练模型前 5 个微博的 NDCG 值对比



	1	2	3	4	5
Sina	0.4242	0.5035	0.5527	0.567	0.5733
ListNet	0.4318	0.5117	0.5541	0.5706	0.5789
RDLS	0.4886	0.572	0.6012	0.6158	0.6254

图 4 3 个基于直接特征的训练模型前 5 个微博的 ERR 值对比

4.5 微博分析特征统计实验

本实验针对对本文提出的微博词性信息量进行实验,利用标记数据针对微博词性比例进行统计分析得到相应的权重和得分,并统计了不同相关度情况下微博的平均长度.

(1) 微博词性信息量

对应标记微博数据中最高两级相关度的微博数据进行统计,同时进行了去停用词处理,前 10 个词性进行对比,其中 n 代表名词,v 代表动词,m 代表数词,d 代表副词,nr 代表人名,ng 代表名词性语素,t 代表时间词,a 代表形容词,q 代表量词,x 代表字符串,f 代表方位词,b 代表区别词,url 代表网址,vg 代表动词性语素,vi 代表不及物动词.如图 5 所示.

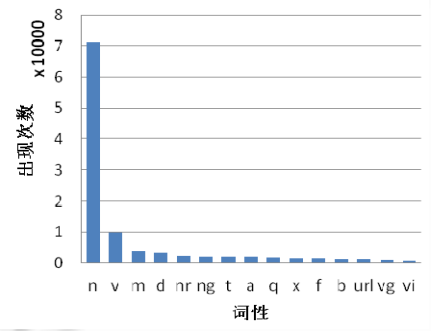


图 5 相关度前两级微博各个词性统计

从图 5 可以看出,名词和动词在整体上占有很大的比例,这对于几乎每一篇微博都包含这两种词性的情况下,这两种词性就会变得不好区分.为此,可以通过计算逆文档频率来降低这些出现次数过多词性在相关性计算中的重要性.计算见公式(13).

$$idf_t = \log\frac{N}{df_t} \tag{13}$$

其中 df_t 代表文档频率,即词性在不同微博中出现的频率. idf_t 代表逆文档频率,N 代表总的微博数量.通过计算,可得每个词性在相关度前两级的微博中的出现频率以及经 idf 计算后前两级相关微博中的词性权重,分别如图 6 和图 7 所示.

从图 6 可以看出,在标记为前两个相关度等级的微博中,名词和动词的出现频率仍然占有较大比例.从图 7 可得,名词权重为 0,则说明其在每一篇高相关微博中都出现了,通过计算得到的 idf 值就为 0,从这里看出,单纯的名词堆砌的文本我们认为是不含有信息量的.最后通过求解信息熵可以得到微博中的词性信息量,见公式(14):所示

$$H(x) = -\sum_{x} p(x)\log(p(x)) \tag{14}$$

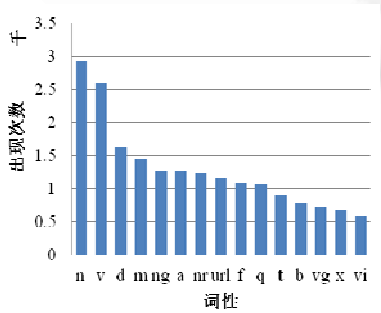


图 6 统计每个词性在相关度前两级的文档频率

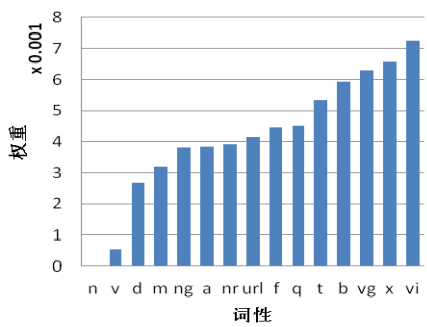


图 7 经 idf 计算后前两级相关微博中词性权重

(2) 微博的分析长度

通过对标记微博数据进行统计,可以得到微博分析长度与相关度的关系,见表 4.

表 4 微博的分析长度与相关度关系

相关度等级	平均长度
完美	39.41388174807198
较好	37.34917940011319
一般	31.25068332682546
较差	29.057845263919017
极差	29.625352112676055
总平均长度	33.33938837032114

从表 4 可以看出,随着相关度等级的下降,微博的平均长度也呈下降趋势.这说明从平均角度来讲查询返回的内容较长的微博对于用户来说其相关度较大.

4.6 加入分析特征的微博排序实验

本实验与第 1 个实验基于相同的训练集、测试集和验证集,在相同的算法设置参数下,加入了分析特征.实验结果如图 8 和图 9 所示,其中 RDLS-A 表示加入分析特征的 RDLS 算法,RDLS-D 表示没有加入分析特征的 RDLS 算法,ListNet-A 表示加入分析特征的 ListNet 算法,ListNet-D 表示没有加入分析特征的 ListNet 算法.

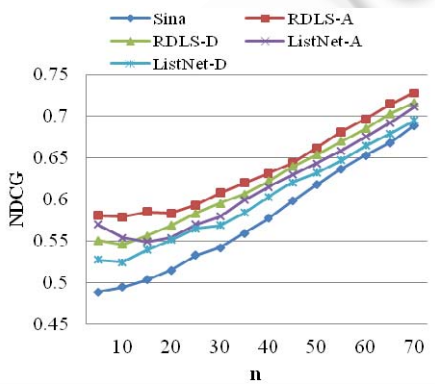


图 8 加入分析特征和不加入分析特征的 3 种算法的 NDCG 值比较(n 从 5~70 每 5 个计算 1 次)

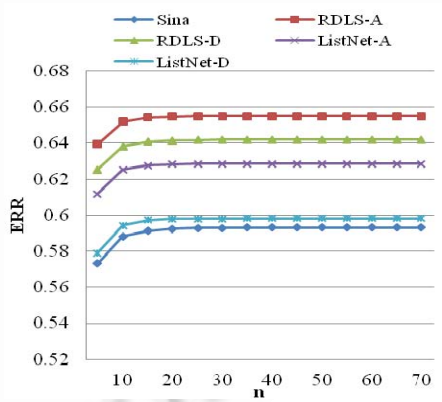


图 9 加入分析特征和不加入分析特征的 3 种算法的 ERR 值比较(n 从 5~70 每 5 个计算 1 次)

从图 8 和图 9 可以看出,从总体上来看,本文提出的算法 RDLS 在加入分析特征(RDLS-A)的情况下训练的模型具有最优的测试结果.ListNet 在加入分析特征(ListNet-A)的起始阶段具有比 RDLS 算法在直接特征下(RDLS-D)有较高的 NDCG 值.但从整体上来看 RDLS-D 仍具有较高的 NDCG 值,可见 RDLS 算法模型有较好的表现.而对于分析特征,从两种评价结果中不难看出,加入分析特征对于提高排序结果的用户满意度具有很明显的作用.

图 10 和图 11 显示了加入分析特征的情况下前 5 个微博的 NDCG 值和 ERR 值.由此可见,在加入分析特征的情况下,ListNet 算法和 RDLS 算法在测试集上前 5 个微博的排序情况都比新浪本身算法要好,而且 RDLS-A 算法比 ListNet-A 算法具有更好的表现.

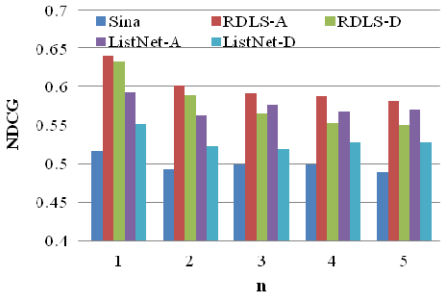


图 10 加入分析特征和不加入分析特征的 3 种算法 NDCG 值比较(*n* 取前 5 个计算)

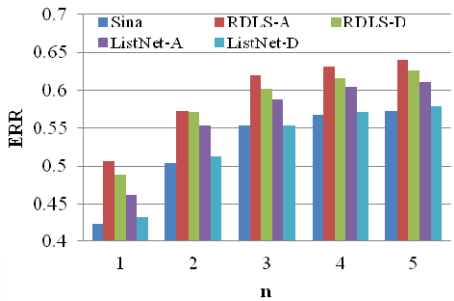


图 11 加入分析特征和不加入分析特征的 3 种算法 ERR 值比较(*n* 取前 5 个计算)

5 总 结

本文针对微博搜索精度不高这一问题,在已有微博直接特征的基础上,创新性地提出了分析特征的概念,提出了基于词性比例来计算微博信息量的思想.同时,本文基于 ListNet 算法,引入了动态步长的策略,结合 Armijo 步长准则提出了 ListNet 优化算法—RDLS 算法,这样在相同的迭代轮数下可以得到更好的模型.实验结果表明,无论是基于直接特征还是加入分析特征,RDLS 算法比 ListNet 算法和新浪微博搜索算法都有更好的表现.而且针对同一种算法(ListNet 或 RDLS),加入分析特征之后,比仅基于直接特征时,都能够取得更好的结果.

致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其要对中国科学院大学何苯副教授、北京理工大学张华平副研究员表示感谢.

References:

[1] Sina Tech. The numbers of registered users of twitter is over 500 million:rank only second to Facebook. (in Chinese) http://tech.sina.com.cn/i/ m/2012-07-31/00387445367.shtml

[2] Liu XH, Wei FR, Duan YJ, Zhou M. Semantic search of microblogs. Journal of Shandong University (Natural Science), 2012, 47(5):39−42.(in Chinese with English abstract)

[3] PhoenixNet. The numbers of registered users of sina microblog is nearly 500 million, 75% of active users login in with mobile terminals. (in Chinese) http://tech.sina.com.cn/i/2012-02-06/15246687778.shtml.

[4] Nagmoti R, Teredesai A, De Cock M. Ranking approaches for microblog search. In: Proc. of the 2010 IEEE/ACM Int'l Conf. on Web Intelligence-Intelligent Agent Technology (WI-IAT). New York: IEEE Press, 2010. 153−157.

[5] Meij E, Weerkamp W, Rijke MD. Adding semantics to microblog posts. In: Adar E, Teevan J, Agichtein E, Maarek Y, eds. Proc. of the 5th ACM Int'l Conf. on Web Search and Data Mining. New York: ACM Press, 2012. 563−572.

[6] Efron M. Hashtag retrieval in a microblogging environment. In: Crestani F, Maillet SM, Chen HH, Efthimiadis EN, Savoy J, eds. Proc. of the 5th ACM Int'l Conf. on Web Search and Data Mining. New York: ACM Press, 2012. 563−572.

[7] Zhao LL, Zeng Y, Zhong N. A weighted multi-factor algorithm for microblog search. In: Zhong N, Callaghan V, Ghorbani AA, Hu B, eds. Proc. of the 7th Int'l Conf. of AMT 2011. Berlin: Springer-Verlag, 2011. 153−161.

[8] Teevan J, Ramage D, Morris MR. #TwitterSearch: A comparison of microblog search and Web search. In: King I, Nejdl W, Li H, eds. Proc. of the 4th ACM Int'l Conf. on Web Search and Data Mining. New York: ACM Press, 2011. 35−44.

[9] Kwak H, Lee C, Park H, Moon S. What's Twitter, a social network or a news media. In: Rappa M, Jones P, Freire J, Chakrabarti S, eds. Proc. of the 19th Int'l Conf. on World Wide Web. New York: ACM Press, 2010. 591−600.

[10] Cha M, Haddadi H, Benevenuto F, Gummadi KP. Measuring user influence in Twitter: The million follower fallacy. In: Proc. of the 4th Int'l AAAI Conf. on Weblogs and Social Media. Washington: AAAI Press, 2010. 10−17.

[11] Efron M. Information search and retrieval in microblogs. Journal of the American Society for Information Science and Technology, 2011,62(6):996−1008.

[12] Stephen R, Hugo Z, Michael T. Simple BM25extension to multiple weighted fields. In: Grossman D, Gravano L, Zhai CX, Herzog O, Evans DA, eds. Proc. of the 13th ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2004. 42−49.

[13] Gao JF, Nie JY, Wu GY, Cao GH. Dependence language model for information retrieval. In: Sanderson M, Jarvelin K, Allan J, Bruza P, eds. Proc. of the 27th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2004. 170−177.

[14] Lawrence P, Sergey B, Rajeev M, Terry W. The PageRank citation ranking: Bringing order to the web. Technical Report, Stanford University, 1999.

[15] Qin T, Liu TY, Xu J, Li H. LETOR: A benchmark collection for research on learning to rank for information retrieval. Information Retrieval, 2010,13(4):346−374.

[16] Cao Z, Qin T, Liu TY, Tsai MF, Li H. Learning to rank: From pairwise approach to listwise approach. In: Ghahramani Z, ed. Proc. of the 24th Int'l Conf. on Machine Learning. New York: ACM Press, 2007. 129−136.

[17] Ponte JM, Croft WB. A language modeling approach to information retrieval. In: Croft WB, Moffat A, Rijsbergen CJV, Wilkinson R, Zobel J, eds. Proc. of the 21st Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 1998. 275−281.

[18] Armijo L. Minimization of functions having Lipschitz continuous first partial derivatives. Pacific Journal of Mathematics, 1966, 16(1):1−3.

[19] Tsai MF, Liu TY, Qin T, Chen HH, Ma WY. FRank: A ranking method with fidelity loss. In: Kraaij W, Vris APD, Clarke CLA, Fuhr N, Kando N, eds. Proc. of the 30th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2007. 383−390.

[20] Jarvelin K, Kekalainen J. IR evaluation methods for retrieving highly relevant documents. In: Yannakoudakis E, Belkin NJ, Leong MK, Ingwersen P, eds. Proc. of the 23rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2000. 41−48.

[21] Jarvelin K, Kekalainen J. Cumulated gain-based evaluation of IR techniques. Jounal of ACM Trans. on Information Systems, 2002, 20(4):422−446.

[22] Chapelle O, Metzler D, Zhang Y, Grinspan P. Expected reciprocal rank for graded relevance. In: Cheung D, Song IY, Chu W, Hu XH, Lin J, eds. Proc. of the 18th ACM Conf. on Information and Knowledge Management. New York: ACM Press, 2009. 621−630.

[23] Burges CJC. From RankNet to LambdaRank to LambdaMART: An overview. Technical Report, MSR-TR-2010-82, Microsoft Research, 2010.

[24] Niu SZ, Guo JF, Lan YY, Cheng XQ. Top-*k* learning to rank: Labeling, ranking and evaluation. In: Hersh W, Callan J, Maarek Y, Sanderson M, eds. Proc. of. the 35th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2012. 751−760.

附中文参考文献:

[1] 新浪科技.Twitter 注册用户量超 5 亿:仅次于 Facebook.http://tech.sina.com.cn/i/m/2012-07-31/00387445367.shtml

[2] 刘晓华,韦福如,段亚娟,周明.基于语义分析的微博搜索.山东大学学报(理学版),2012,47(5):39−42.

[3] 凤凰网.新浪微博注册用户达 5 亿,75%活跃用户从移动端登陆.http://tech.sina.com.cn/i/2012-02-06/15246687778.shtml

周诗龙(1964－),男,山东蓬莱人,硕士生,主要研究领域为网络挖掘.
E-mail: kunlong0909@163.com

徐俊刚(1972－),男,博士,副教授,主要研究领域为云计算,信息检索,网络挖掘.
E-mail: xujg@ucas.ac.cn