第 43 卷 第 11 期 计 算 机 工 程 2017 年 11 月
Vol. 43 No. 11 Computer Engineering November 2017

·人工智能及识别技术· 文章编号：1000-3428(2017)11-0152-09 文献标志码：A 中图分类号：TP18

基于排序学习模型的微博多样性检索问题研究

王 莹，罗准辰，于 洋

（中国国防科技信息中心，北京 100142）

摘 要： 多样性检索主要用于解决传统信息检索中面临的查询词歧义问题。为此，研究微博中的多样性检索，提出一种新的微博多样性检索方法，将多样性排序学习方法应用到微博多样性检索。开发一系列社交媒体特征和子话题分布特征，采用查询短语与博文间相关性特征和博文与博文间文本多样性特征模型作为基准，分别加入上述特征，检验其对微博多样性的影响。实验结果表明，多样性排序学习方法能有效解决微博多样性检索问题，明显提高微博检索的效果。

关键词： 机器学习；信息检索；多样性检索；排序学习；社交媒体

中文引用格式： 王 莹，罗准辰，于 洋. 基于排序学习模型的微博多样性检索问题研究[J]. 计算机工程，2017，43(11)：152-160.

英文引用格式： WANG Ying，LUO Zhunchen，YU Yang. Research on Microblog Diversification Retrieval Problem Based on Rank Learning Model[J]. Computer Engineering，2017，43(11)：152-160.

Research on Microblog Diversification Retrieval Problem Based on Rank Learning Model

WANG Ying，LUO Zhunchen，YU Yang

（China Defense Science and Technology Information Center，Beijing 100142，China）

【Abstract】 Diversification retrieval is used to solve users' information needs，which typically described by query phrase are often ambiguous and have more than one interpretation. This paper researches microblog diversification retrieval，and proposes a novel microblog diversification retrieval method，diversification learning to rank method is applied to microblog diversification retrieval. It develops a series of social media features considering the characteristics of microblog and subtopics distribution，and adds these features one by one to the baseline model which only considering the relational features and the text diversity feature to verify the effectiveness of them. Experimental results show that diversification learning to rank approach can solve microblog diversification retrieval problem，and improve the effectiveness of microblog retrieval.

【Key words】 machine learning；information retrieval；diversification retrieval；rank learning；social media

DOI： 10.3969/j.issn.1000-3428.2017.11.025

0 概述

微博是近年来发展迅速且影响深远的社交媒体平台。随着微博数据剧烈增长，从大量微博数据中检索用户感兴趣的博文成为微博检索的重要任务。但是，用户输入的查询短语通常不明确，一方面，查询短语可能存在歧义；另一方面，查询所对应的内容可能多元化。例如，在 Twitter（微博的代表性网站）中，对于查询短语“dreamliner battery”，可以检索到以下 3 条内容相关的博文：

1）$tweet_1$：Boeing 787 battery fire was difficult to control：An investigation of a battery fire aboard a Boeing 787。

2）$tweet_2$：Boeing 787 Dreamliner battery was miswired，Japan says-CTV News。

3）$tweet_3$：Rockford-Area News Boeing proposes battery fix to FAA for 787 Dreamliner planes。

其中，$tweet_1$ 介绍梦幻客机电池事故；$tweet_2$ 分析电池事故原因；$tweet_3$ 则提出事故的解决方案。这 3 条博文表达了查询短语“dreamliner battery”的不同子话题，仅仅从查询短语无法确定用户对哪一方面感兴趣。

多样性检索是一种新兴的技术趋势，其目的是为了解决传统信息检索中查询词歧义或不确定性问

基金项目： 国家自然科学基金-青年科学基金(61602490)。

作者简介： 王 莹(1992—)，男，硕士研究生，主研方向为自然语言处理、信息检索；罗准辰，工程师、博士；于 洋，高级工程师、博士。

收稿日期： 2016-09-07 **修回日期：** 2016-12-02 **E-mail：** suneony@gmail.com

题[1]。多样性检索是在传统的相关性检索的基础上考虑了文档与文档间关系,核心思想是在有限的位置空间呈现内容多样的检索结果,降低信息冗余,使得不同背景的用户都能在检索结果中快速找到满足其需求的内容。

传统网页中多样性检索是基于查询短语短小问题提出来的,对于微博检索来讲,查询短语更短(平均仅包含 1.64 个词[2]),因此在微博检索中考虑多样性问题更有必要。但是,与传统网页相比,微博文本简短,书写随意,噪声大且实时性强,导致微博多样性检索要比传统网页多样性检索更加困难。而现有网页多样性检索方法没有针对微博的这些特点进行优化,简单地将其直接应用到微博检索中,效果一般[3]。

微博本身具有丰富的社交媒体信息,如发布时间、主题词、提及、超链接及发布者信息等。这些信息从不同的角度反映博文的属性,进而反映博文间关系,能够帮助微博多样性检索问题的解决。例如,针对同一查询短语,相同时间窗口的 2 条博文更可能与同一子话题相关。本文将研究这些社交媒体信息在微博多样性检索中的作用。

本文提出利用多样性排序学习模型[4]解决微博多样性检索问题,并根据微博的特点提出一系列特征,包括文本特征、社交媒体特征以及子话题分布特征。通过多样性排序学习模型,能方便地集成这些特征,并根据训练数据自动估算其权值。最终通过数据集 Tweets2013 进行方法有效性评价实验[5]。

1　相关工作

本文从 3 个方面介绍相关工作:微博检索,网页中的多样性检索和微博中的多样性检索。

1.1　微博检索

微博检索已经取得了一系列研究成果。文献[6]通过贝叶斯网络对查询短语与文档间关系进行建模,构建贝叶斯网络时,考虑了文档的时间与文本特征。文献[7]在检索模型中,通过将文档长度归一化来解决微博文本的稀疏性问题。文献[8]在 Learning To Rank 框架下对微博检索进行研究,他们考虑多种可能影响检索结果排序的特征,包括博文中的超链接、主题词、提及、微博用户的信息(关注数和好友数等)以及微博的转发数和回复数等。文献[9-10]提出博文信息结构化的表示方法,这种方法能通过文本结构对博文进行聚类,而每一类别都有特定的信息特征,这些特征能够提高微博中的检索效果。除了针对微博文本的检索研究外,还有对微博用户信息分析的检索方法研究。文献[11]使用了一种类似于 PageRank 的方法来计算用户的权威度,这种方法在计算用户影响力时不仅使用 PageRank 方法考虑用户关系网结果,还考虑了用户之间所关注的话题相似性。此外,微博检索中还包含了对博文中主题词的检索[12]。在微博中通常用“#”开头的词表示一个热门主题词。而对主题词的检索就是对带有这样标记的词进行检索。用户通常会引入一个主题词来表达一条博文的话题,因此对主题词的检索能够直接获取内容相关的博文。文献[13]对社交媒体中存在的噪音和冗余信息进行过滤和筛选,获取高质量的信息,提出基于核主分析和小波变换的高质量微博提取框架并设计一种基于多特征融合的高质量信息的提取算法。文献[14]结合微博中的会话、转发和话题标签,将微博划分为用户兴趣、用户互动和话题微博 3 类,提出基于作者主题模型(ATM)的话题标签主题模型 HC-ATM,使用 Gibbs 抽样法对模型进行推导,获取微博主题结构。

文本检索会议 TREC(Text Retrieval Conference)是文本检索领域最权威的评测会议。TREC 在 2011 年第一次举办了微博检索任务,任务要求在给定的 Twitter 数据集上进行实时检索,对于每个查询返回一个按微博创建时间逆序排列的列表,这个列表不但要与查询短语内容相关还要满足实时性的要求。许多院校和研究机构围绕这个问题开展研究[15]。

上述微博检索都是简单的 ad-hoc 检索方法,只考虑了博文与查询短语间的关系,并没有从多样性检索任务的角度考虑博文与博文间的关系。

1.2　网页多样性检索

最近几年,网页中的多样性检索方法的研究逐渐深入。大体上这些方法可以分为两类:隐式多样性检索方法与显式多样性检索方法。

隐式多样性检索方法假设内容相似的文档一般包含相似内容子部分,这些子部分可以满足不同用户的需求。隐式多样性检索方法在选择返回结果文档时,需要考虑被选的文档是否与已经选择过的检索结果中的文档内容相似,如果不相似,则将其纳入检索结果中[16],反之亦然。在度量内容相似的方法上,可以考虑余弦夹角方法或 KL(Kullback Leiler)距离[17]。文献[18]利用 MMR(Maximal Marginal Relevance)方法减少文档之间的数据冗余,这个方法利用迭代的思想将文档添加到检索结果集合中,在添加文档的过程中,尽量使得被选文档与查询短语内容相关,降低文档与结果集合中已选文档的内容相关性。文献[19]提出一种利用负反馈来尽量扩大检索结果多样性的方法。文献[20]提出子话题检索模型以此满足不同用户的需求,该模型主要考虑了被检索文档之间的内容相关性,以此尽量覆盖给定查询短语话题关联的子话题。文献[21]同样考虑了被检索文档之间的内容相关性,他们的研究同样发现,降低文档之间的内容相关性,能在一定程度上满

足不同用户的检索需求。文献[22]提出了一种支持搜索结果多样化的数据融合排名算法 combSumDiv，并与两种具有代表性的支持检索结果多样化的算法 xQuAD[16] 和 PM-2[23] 进行比较，实验结果表明 combSumDiv 在性能上优于 xQuAD[16] 与 PM-2[23]。

显式多样性检索方法则直接从查询短语进行分析，找到用户需求的各个子部分，检索出的文档集合在内容上尽量覆盖各个子部分，以此满足不同用户的需求。文献[1]将一个分类系统分别用于查询短语与检索文档，以此解决多样性检索的问题，他们的方法是将两个归到一类的文档视为相似文档，而检索文档集合尽量包含不同类别的文档，类别则首先基于查询短语的不同子部分类别确定。文献[24]提出了一种概率模型，这种模型最大限度使检索到的文档集合覆盖查询短语的各个子部分，而查询短语的各个子部分则通过初始检索中排序靠前的文档集合分析得到。文献[25]也提出了一种过滤检索文档集合的方法，使得文档集的分布能与查询短语各个子部分最大程度地相关，他们的方法是从商业搜索引擎的搜索日志中分析查询短语的各个子部分。另外，比较有效的显式方法还包括 RxQuAD[26]，IA-select[1]，PM-2[23]，DSPApprox[27] 等。文献[28]针对用户的查询意图会随着时间的推移发生改变问题，根据时间点击图挖掘原理提出一种查询建议方法。对原始的查询日志文件进行预处理，生成时间点击图。对时间点击图进行非连通子图检测和图的合并操作，以降低或消除图的非连通性。采用基于随机游走模型的图挖掘算法，生成给定查询的查询建议集。

文献[4]提出了一种基于机器学习的多样性排序学习方法 R-LTR(Relational Learning To Rank)。R-LTR 是对传统的相关性排序学习方法的改进，相关性排序学习方法只考虑查询短语与文档间的关系，而 R-LTR 在此基础上还考虑了文档与文档间的关系。R-LTR 中排序函数依据上述两方面进行定义，是对查询短语与文档间相关性得分与文档与文档间多样性得分的组合。

以上多样性检索方法都是针对传统网页的检索，并没有针对微博的特点进行微博多样性检索方面的研究。

1.3 微博多样性检索

文献[5]构建了一个微博多样性检索评测数据集 Tweets2013。数据集包括微博数据和一系列话题。每条微博数据是否与某个子话题相关都进行了人工判定。为了构建该数据集，作者通过 Twitter API 收集了 2013 年 2 月 1 日—2013 年 3 月 31 日间的 Twitter 数据，共包含 291 922 201 条博文。通过 Wikipedia's Current Events Portal[5] 获取了 2013 年 2 月1 日—2013 年 3 月 31 日间的新闻事件并选择其中 47 个作为查询话题。通过该数据集，可以对微博多样性检索方法进行评价。

目前针对微博多样性检索的研究还比较少，现有的方法都是直接将传统网页多样性检索方法应用到微博检索中。文献[3]对这些传统网页多样性检索方法进行了比较实验，这些方法包括：LexRank[29]，Biased LeaRank[30]，MMR[18]，Max-Sum[3]，Max-Min[3] 以及 xQuAD[16] 等。上述研究发现，将传统网页多样性检索方法直接应用到微博多样性检索中，隐式多样性检索方法要优于显式多样性检索方法。这是由于对查询短语进行显式建模时，由于微博文本简短，用于描述查询多种意图的词汇并没有出现在博文中所造成的。

由于 Twitter 的开放性和数据获取的便捷性，本文选取 Twitter 平台作为研究对象，因此文中提及到的微博等同于 Twitter，博文等同于 tweet。

本文通过数据集 Tweets2013 进行方法有效性评价实验。该数据集是文献[5]针对微博多样性检索问题构建。实验结果表明，多样性排序学习方法在解决微博多样性检索任务中与传统的多样性方法相比效果相当。与现有方法相比，α-nDCG@20 和 ST-Recall@20 指标值分别提高了 6.2% 和 5.8%，其他指标值相当。另外，在微博检索中，引入博文间的多样性特征是有效的，与不加入多样性特征相比，各个指标都有显著提升，其中 Precsion-IA@10 与 Precsion-IA@20 指标值提高明显，分别提高了 80.0% 和 72.1%。最后，本文发现子话题分布特征和社交媒体特征(特别是发布时间特征、主题词特征和发布者的地理位置特征)能有效帮助微博多样性检索问题的解决。引入子话题分布特征后，α-nDCG@10 与 α-nDCG@20 指标值分别提高了 8.8% 和 10.9%；引入时间特征后，α-nDCG@10与 α-nDCG@20 指标值分别提高了 4.6% 和 8.3%；引入主题词特征后，α-nDCG@10 与 α-nDCG@20 指标值分别提高了 8.0% 和 12.2%；引入地理位置特征后，α-nDCG@10 与 α-nDCG@20指标值分别提高了 3.5% 和 3.6%。

本文的主要贡献如下：

1)提出了一种新的微博多样性检索方法。

2)将多样性排序学习方法[4]应用到微博多样性检索任务中。

3)开发了一系列社交媒体特征和子话题分布特征，并将其应用到多样性排序学习方法中，帮助微博多样性检索问题的解决。

4)实验结果表明，将多样性排序学习方法应用到微博多样性检索任务中，检索效果与传统的多样性方法效果相当；在微博检索中，引入多样性特征是有效的；子话题分布特征和社交媒体特征能有效帮助解决微博多样性检索问题。

2　问题定义

微博多样性检索可以看作是对初步检索结果再排序的过程,其目标是用最小的结果集合最大限度地覆盖查询的各个子话题。因此,该问题可以定义为:在微博检索中,给定查询短语 q 和与 q 相关博文集合 S,查询短语 q 的所有意图集合为 A。博文集合 S 的初始排序为 R。对集合 S 重新排序得到新的排序 N,使得 N 中前 k 个结果 T 最大限度覆盖查询 q 的各个意图,即最大化 $|U_{d\in T}a|$,其中,a 为博文 d 对应的查询意图且 $a\in A$;此外,还要使 T 中博文间的冗余度最低。

3　本文方法

本文采用多样性排序学习框架解决微博多样性检索问题并提取一系列特征。

3.1　多样性排序学习框架

传统的微博检索已经可以通过相关性排序学习方法很好地解决。相关性排序学习方法是一种将相关性特征有效地整合到排序模型的机器学习算法。而在多样性检索中,在考虑博文与查询词间相关性的基础上还要考量博文间的相似性。本文采用多样性排序学习方法解决微博中的多样性检索问题,通过机器学习的方式,将多种影响博文与查询词相关性和博文间相似性的因素,转换为特征进行训练与测试,将特征按照不同的方式进行组合,通过排序效果验证特征的有效性。微博中的多样性排序学习框架如图 1 所示。

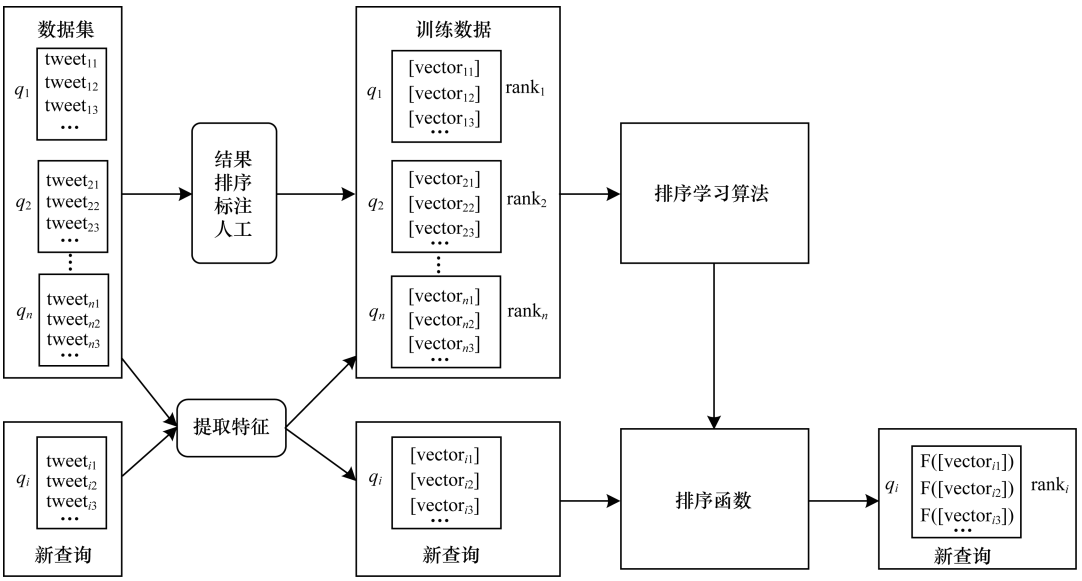

图 1　多样性排序学习框架

首先给定查询词集合 $Q=\{q_1,q_2,\cdots,q_n\}$,每个查询词 q_i 都有对应一个博文集合 $T_i=\{\text{tweet}_{i1},\text{tweet}_{i2},\cdots,\text{tweet}_{im}\}$。每条博文 tweet_{ij} 都人工标注其是否与对应的查询词 q_i 相关及其对应的子话题(查询意图),通过这些标注信息生成博文集合 T_i 排序的标准答案 rank_i。每个查询词 q_i 都有对应一个博文集合 $T_i=\{\text{tweet}_{i1},\text{tweet}_{i1},\cdots,\text{tweet}_{im}\}$。一系列用于衡量博文集合 T_i 中的每条博文 tweet_{ij} 与查询词 q_i 间相关性以及衡量博文集合 T_i 内每条博文间相似性的特征被设计和提取,形成特征向量 vector_{ij}。假设多样性排序中的排序函数为 f,多样性排序学习的目标就是通过人工标注的训练数据得到最优的排序函数 f。

对于一个新的查询 q 和其对应的博文集合 T,抽取同样的特征形成特征向量 vector,然后利用排序函数 f 生成最终的多样性检索结果。多样性排序学习中的排序函数定义式(1)所示。

$$f(x_j^{(i)},\boldsymbol{r}_j^{(i)})=\boldsymbol{w}_r^{\mathrm{T}}x_j^{(i)}+\boldsymbol{w}_d^{\mathrm{T}}h(\boldsymbol{r}_j^{(i)}) \tag{1}$$

其中,$\boldsymbol{w}_j^{(i)}$ 表示查询 q_i 与对应的博文 tweet_{ij} 间的相关性特征向量,$\boldsymbol{r}_j^{(i)}$ 表示博文 tweet_{ij} 与排在其前面的博文集合的相似性特征矩阵,函数 h 通过将每个特征对应的多个值求平均值的方式将相似性特征矩阵转化为特征向量。$\boldsymbol{w}_r^{\mathrm{T}}$ 和 $\boldsymbol{w}_d^{\mathrm{T}}$ 分别表示相关性特征向量和相似性特征向量的权重。

3.2　博文相似性特征

从上文可以看出,多样性检索的核心是博文与查询间相关性和博文间相似性的度量,针对微博检索的研究已经提出了多种相关性特征,本文着重研究博文间相似性度量。针对微博多样性检索问题的特点,本文从以下角度设计博文间的相似性特征。

1)文本特征:描述博文间文本相似度。

2)子话题分布特征:采用话题模型发现子话题,描述博文在子话题分布上的相似度。

3)社交媒体特征:描述微博特有的社交媒体信息相似度。

4 特征描述

4.1 文本特征

关于同一主题的两篇博文的文本相似性越高,则更可能两篇博文涉及同一子话题。因此,考虑博文间的文本特征。计算该特征时,首先采用空间向量模型对博文进行建模,所谓空间向量模型是指一个博文可以表示为由博文中的词项及其权重组成的向量,每个词项表示向量空间的一个维度,词项的权重表示博文向量在该维度上的取值。通过空间向量模型对博文建模后,博文间的相似性可以用博文向量的余弦相似度表达。特征的计算方式如式(2)所示。

$$TT_1 = 1 - \frac{\boldsymbol{t}_i \cdot \boldsymbol{t}_j}{|\boldsymbol{t}_i| \cdot |\boldsymbol{t}_j|} \tag{2}$$

其中,$\boldsymbol{t}_i$和$\boldsymbol{t}_j$分别为博文文本的向量化表示。

4.2 子话题分布特征

多样性检索的核心思想是使检索结果尽可能覆盖查询短语所对应的所有子话题。子话题模型能够检测关于同一主题的所有博文的隐式子话题并估算每条博文在其上的概率分布。

例如,在所有与查询短语“hillary clinton resign”相关的博文中,可以检测到不同的子话题(通常用一系列的词汇表示),如“last, politician, popular, obama, clinton”和“look, woman, india, resign, speech”。博文在不同子话题上的概率分布反映了博文间的关系。因此,在微博多样性检索中考虑博文的子话题特征。采用HDP(Hierarchical Dirichlet Processes)模型检测与话题相关的博文集合中的隐式子话题,并采用HAC模型计算每条博文在子话题上的概率分布。

表1所示是在与查询短语“hillary clinton resign”相关的博文集合中检测到的子话题以及博文tweet_4“Hillary Clinton tops Obama as most popular US politician”和博文tweet_5“Hillary Clinton Endorses Gay Marriage”在这些子话题上的概率分布。子话题特征的计算公式如式(3)所示。

$$TT_2 = \sqrt{\sum_{k=1}^{m} (p(z_k|\boldsymbol{t}_i) - p(z_k|\boldsymbol{t}_j))^2} \tag{3}$$

其中,$p(z_k|t_i)$是指博文tweet_i在子话题z_k上的概率。

表1 子话题及博文关于子话题的概率分布

子话题	子话题描述	tweet_4	tweet_5
SubTopic1	last,politician,popular,obama,clinton	0.980 7	0.072 1
SubTopic2	look,woman,india,resign,speech	0.003 0	0.100 0
SubTopic3	run,go,want,hat,snap	0.002 6	0.116 3
SubTopic4	justin,send,bieber,gonna,lift	0.003 4	0.128 3
SubTopic5	clinon,dick,give,email,check	0.001 7	0.181 0
SubTopic6	better,miss,bank,wemen,bill	0.003 2	0.105 5
SubTopic7	first,john,visit,iraq,kabul	0.000 2	0.158 6
SubTopic8	week,sworn,post,know,vote	0.002 0	0.106 5
SubTopic9	support,former,samesex,syria,hurrah	0.002 7	0.317 9

4.3 社交媒体特征

微博特有的社交媒体特征从侧面反映了博文间的关系。本文考虑微博的社交媒体特征。

1)时间特征

与话题相关的同一子事件往往发生在同一窗口内,所以2条博文在同一话题下,发布时间越接近,则其可能涉及相同的子话题。因此,考虑博文间的时间特征。时间特征的计算基于2个时间归一化后的时间戳(采用Min-Max归一化),计算方式如式(4)所示。

$$TT_3 = |t_{\text{norm}}(\boldsymbol{t}_i) - t_{\text{norm}}(\boldsymbol{t}_j)| \tag{4}$$

其中,$t_{\text{norm}}(\boldsymbol{t}_i)$和$t_{\text{norm}}(\boldsymbol{t}_j)$分别表示两篇微博发布时间的归一化表示,例如,有最小时间戳“Fri Feb 01 00:09:29 +0000 2013”和最大时间戳“Sun Mar 31 23:57:58 +0000 2013”,对时间戳“Tue Mar 25 14:45:00 +0000 2008”归一化后为0.387 101。

2)主题词特征

主题词是指博文中以“#”开头的词汇,通常用主题词来表达微博的主题。如果2篇博文包含相同主题词,则说明2条博文涉及的子话题可能相同。因此,考虑博文间主题词特征。本文采用Jaccard方法来计算两篇博文涉及的主题词的相似性,计算方法如式(5)所示。

$$TT_4 = 1 - \frac{|Term(\boldsymbol{t}_i) \cap Term(\boldsymbol{t}_j)|}{|Term(\boldsymbol{t}_i) \cup Term(\boldsymbol{t}_j)|} \tag{5}$$

3)提及特征

在微博中,用户通常在用户名前面加上“@”来提及其他用户。如果2条关于同一主题的博文提及到相同的用户,则2条博文的发布者可能对同一用户谈及相同子话题。因此,考虑博文与博文间提及特征并采用二元特征来表达2条博文是否提及相同的用户。

4)超链接特征

超链接在微博中很常见,通常包含超链接的博文都是对链接内容的简介。如果2条关于同一话题的博文包含相同的超链接,则2条博文可能涉及相同的子话题。因此,考虑博文与博文间超链接特征并采用二元特征来表达2条博文是否包含相同的超链接。

5)发布者用户特征

微博作为一个典型的社交媒体平台,其丰富的用户信息可能帮助解决微博多样性检索问题。因此,考虑博文发布者间的用户特征。博文发布者的用户特征包括用户的地理位置(location)、用户是否认证(verified)、用户语言(language)、用户发布的博文数量(statuses)、用户的好友数量(friends)、用户的关注者数量(followers)、用户被其他用户分组次数(listed)。当某个子事件发生在某一地区时,相同地区的人们往往都会讨论这个子事件,因

此,考虑用户的地理位置信息。采用二元特征表达用户的地理位置信息。直观来看,2 个使用相同语言的用户比使用不用语言的用户更可能会关注相同的子话题,因此,考虑用户的语言特征。采用二元特征来表达用户的语言信息。用户的其他属性也可能反映用户关注的话题间的关系,比如用户是否为认证用户,用户发布的博文数量、好友数量、关注数量和被分组次数等,因此考虑用户的这些特征。采用二元特征表达用户的认证信息,如果 2 个用户都通过认证,则 verified 为 0,否则为 1。计算其余 3 个特征时,将其归一化到区间[0,1],计算归一化后数值间的差异。

5　实验

5.1　数据集与实验设定

文采用 Tweet2013[5] 数据集对多样性检索方法进行评测,该数据集是针对多样性检索问题构建。数据集收集了 2013 年 2 月 1 日—2013 年 3 月 31 日间的 Twitter 数据,共包含了 47 个查询主题,每个查询主题平均包含 9 个子话题。

采用文献[4]提出的多样性排序学习模型进行实验,与传统的相关性排序学习方法不同,多样性排序学习方法在考虑博文与查询短语间的相关性的基础上还考虑了博文与博文间的相似性。计算博文间的相似性采用前文提出的多种特征,计算博文与查询短语间的相关性,也提出多种特征,如表 2 所示。

表 2　相关性特征

特征	特征描述
Rel_content	查询短语与博文间文本相似度
Rel_hashtag	查询短语与主题词相似度
Rel_mention	博文是否包含提及
Rel_url	博文是否包含超链接
Rel_verified	用户是否为认证用户
Rel_statuses	用户发布的博文数目
Rel_friends	用户的好友数目
Rel_followers	用户的关注数目
Rel_listed	用户被分组次数

本文关注 3 个子问题:

1)多样性排序学习方法在解决微博多样性检索问题时是否有效(见 5.3 节)。

2)微博间的多样性特征能否提高微博检索的效果(见 5.4 节)。

3)子话题分布特征和社交媒体特征对微博多样性检索的影响(见 5.5 节)。

本文通过 3 个实验分别说明上述 3 个子问题。实验采用 5 折交叉验证的方式,将数据按照 3∶1∶1 的比例分为训练数据、验证数据和测试数据。

5.2　评测标准

研究人员提出一系列用于评价网页多样性检索系统性能的指标,本文采用这些标对微博多样性检索方法进行评测。

1)α-nDCG[31]。TREC 多样性检索任务中采用了这个评价指标。α-nDCG 基于 nDCG(Normalized Discounted Cumulative Gain)并在其基础上考虑了文档对排在其前面的文档的依赖度。

2)Precision-IA[1]。衡量一个话题对应的前 k 个文档与子话题的相关度。

3)Subtopic-Recall[5]。衡量一个话题对应的前 k 个文档对子话题的覆盖度。取值范围为 0 ~ 1,值越大说明对子话题的覆盖度越高。

5.3　多样性排序学习方法实验

为了验证多样性排序学习方法在微博多样性检索中的有效性,本文将 RLTR 方法与现有的微博多样性检索效果最好的方法进行比较。

1)MMR

MMR(Max Marginal Relevance)[18] 是一个经典隐式多样性检索方法,利用迭代的思想将文档添加到检索结果集合中,在添加文档的过程中,尽量使被选文档与查询短语相关,降低文档与已选文档集合的相似性。文献[3]将 MMR 方法应用于微博多样性检索中。

2)xQuAD

xQuAD(eXplicit Query Aspect Diversification)[16] 是一种显式多样性检索方法。该方法直接对查询短语进行建模,找到查询短语各个子部分,在检索时使文档尽量覆盖各个子部分。文献[3]将 xQuAD 方法应用到微博多样性检索中。

3)Simple Yet

文献[5]提出一种检测博文间冗余度的方法,并以此定义了一系列基于博文文本的特征。该方法遍历博文的初始列表,如果某一个博文与排在其前面的博文相似度高于设定的阈值,则将该博文从列表中删除。

本文将上述 3 种方法作为基准系统来验证多样性排序方法在微博多样性检索中的有效性。文献[3]在实现上述 3 种方法时,计算博文与查询短语的相关性只采用了查询短语与博文的文本特征,计算博文间相似性采用了博文与博文间文本特征、博文与博文间主题词特征和博文与博文间发布时间特征,因此本文也选取了上述特征引入多样性排序学习方法,标记为 RLTR,实验结果如表 3 所示。从表 3 中可以看出,RLTR 方法的 α-nDCG@20 与现有方法中效果最好的 Sy 相比提高了 8.9%,而其他指标比较差异不大,说明多样性排序学习方法与其他多样性检索方法在检索效果上相当。因此,将多样性排序学习方法应用到微博多样性检索中是有效的。

表 3 多样性排序学习方法指标对比

方法	α-nDCG		Prec-IA		ST-Recall	
	@10	@20	@10	@20	@10	@20
MMR	0.341	0.374	0.066	0.056	0.417	0.539
xQuAD	0.235	0.263	0.050	0.041	0.302	0.419
Sy	0.384	0.383	0.083	0.069	0.419	0.542
RLTR	0.382	0.417	0.058	0.051	0.351	0.483

5.4 多样性特征实验

微博相关性检索只考虑查询短语与博文间的相关性,多样性检索在此基础上考虑了博文与博文间的相似性。为了验证多样性特征在微博检索中的有效性,本文进行对比实验。实验采用多样性排序学习方法 RLTR。第 1 个实验中只考虑博文与查询短语间的相关性特征,如表 1 所示,标记为 $RLTR_{REL}$。第 2 个实验在第 1 个实验的基础上,加入了博文与博文间相似性特征,标记为 $RLTR_{REL+DIV}$。实验结果如表 4 所示,粗体表示与基准相比显著提高($p<0.05$)。从表 4 可以看出,加入多样性特征后,各项指标都有显著提升,其中,Prec-IA@10 与 Prec-IA@20 指标值提高最明显,分别提高了 64.4% 和60.4%;α-nDCG@10 与 α-nDCG@20指标值分别提高了 71.4% 和 58.4%。上述结果说明,在微博检索中引入博文与博文间的多样性特征是有效的。

表 4 多样性特征指标对比

方法	α-nDCG		Prec-IA		ST-Recall	
	@10	@20	@10	@20	@10	@20
$RLTR_{REL}$	0.282	0.322	0.045	0.043	0.273	0.445
$RLTR_{REL+DIV}$	**0.491**	**0.510**	**0.074**	**0.069**	**0.441**	**0.578**

5.5 社交媒体特征实验

针对微博多样性检索,本文设计了一系列社交媒体特征和子话题分布特征,并通过实验验证这些特征对微博多样性检索的影响。采用加入了查询短语与博文间相关性特征(如表 1 所示)和博文与博文间文本多样性特征(content)模型作为基准,记为 $RLTR_{BASE}$。在其基础上,分别加入子话题特征和社交媒体特征,检验其对微博多样性影响。实验结果如表 5 所示,粗体表示与基准相比显著提高($p<0.05$)。实验结果表明,社交媒体特征能帮助微博多样性检索任务的解决。其中博文与博文间时间特征、博文与博文间主题词特征、博文与博文发布者地理位置特征和子话题分布特征对多样性检索效果有显著提升。加入博文与博文间时间特征后 α-nDCG@10 与 α-nDCG@20 指标值与 $RLTR_{BASE}$ 相比分别提高了6.7% 和 8.2%。对博文与博文间时间特征来讲,如果关于同一主题的两条博文在发布时间上接近,则其可能涉及同一子话题。例如,在 Tweets2013 中,与查询短语“hillary clinton resign”相关的博文如下:

1)$tweet_4$。DTN China:Secretary of State Hillary Clinton formally resigns:Her resignation is effective upon the swearing[created_at:Fri Feb 01 20:39:04 +0000 2013]。

2)$tweet_5$。Secretary of State Hillary Clinton formally resigns:Her resignation is effective upon the swearing-in of John[created_at:Fri Feb 01 20:43:07 +0000 2013]。

3)$tweet_6$。Hillary Clinton:As Hillary Clinton leaves office after four years,John Kerry prepares to take over[created_at:Fri Feb 01 10:00:14 +0000 2013]。

其中,$tweet_4$ 和 $tweet_5$ 都是关于希拉里辞职子话题,而 $tweet_6$ 则关于希拉里的继任者子话题。可以看出,$tweet_4$ 和 $tweet_5$ 发布时间接近,而 $tweet_6$ 则与前两条博文发布时间相距较远。因此,博文与博文间发布时间特征能有效帮助微博多样性检索任务的解决。

表 5 子话题与社交媒体特征实验

方法	α-nDCG		Prec-IA		ST-Recall	
	@10	@20	@10	@20	@10	@20
$RLTR_{BASE}$	0.417	0.428	0.056	0.049	0.349	0.465
+ subtopic	**0.454**	**0.475**	**0.067**	**0.061**	**0.421**	**0.574**
+ time	**0.445**	**0.463**	**0.064**	**0.057**	**0.393**	**0.536**
+ hashtag	**0.431**	**0.459**	**0.063**	**0.059**	**0.401**	**0.519**
+ location	**0.433**	**0.455**	**0.060**	**0.051**	**0.387**	**0.518**
+ verified	0402	0.421	0.056	0.048	0.341	0.471
+ language	0413	0.420	0.054	0.049	0.339	0.466
+ statuses	0.420	0.431	0.058	0.051	0.355	0.470
+ friends	0.417	0.426	0.053	0.049	0.346	0.462
+ followers	0.399	0.405	0.051	0.047	0.374	0.453
+ listed	0.401	0.409	0.050	0.049	0.353	0.475
Best	**0.499**	**0.514**	**0.075**	**0.068**	**0.447**	**0.580**

加入博文与博文间主题词特征后,α-nDCG@10 与 α-nDCG@20 指标值与 $RLTR_{BASE}$ 相比分别提高了 3.4% 和 9.6%。对博文与博文间主题词特征来讲,如果关于同一主题的 3 条博文主题词中有相同的部分,则其可能涉及同一子话题。例如,在 Tweets2013 中,与查询短语“syria civil war”相关博文如下:

1)$tweet_7$。RT @ SyriaDayofRage:(02-21-13) # Damascus # Syria 1 Rebels Move Closer to Central Damascus as Clashes continue and Rebel Rockets hitting。

2)$tweet_8$。So far today 17 martyrs were reported in #Damascus and its suburbs,3 in #Idlib,3 in #Hama,3 in #Aleppo,2 in #Homs。

3)$tweet_9$。RT @ lysdeschamps:# Syria 13 ‘02 Killed:231.20 children,98 women,59 rebels. Aleppo 50 Idlib 33 Damascus 30 Daraa 21 Homs 18 Deir Azzo。

其中,$tweet_7$ 和 $tweet_8$ 都是关于叙利亚内战发生城市的信息,而 $tweet_9$ 则关于战争伤亡状况。可以看出,$tweet_7$ 和 $tweet_8$ 都包含主题词"#Damascus",而 $tweet_9$ 则没有。因此,博文与博文间主题词特征能有效帮助微博多样性检索任务的解决。

加入博文发布者地理位置特征后 α-nDCG@10 与 α-nDCG@20 指标值与 $RLTR_{BASE}$ 相比分别提高了 3.8% 和 6.3%。当某个子事件发生在某一地区时,相同地区的人们往往都会讨论这个子事件。因此,博文与博文间发布者地理位置特征能有效帮助微博多样性问题的解决。

加入博文子话题特征后,α-nDCG@10 与 α-nDCG@20指标值与 $RLTR_{BASE}$ 相比分别提高了 8.8% 和 10.9%。对博文的子话题特征来讲,2 篇关于子话题分布类似的博文可能涉及到相同的子事件。

最后将上述显著提高多样性排序效果特征全部加入到框架中,得到最优模型,包括时间特征、主题词特征、发布者地理位置特征、子话题分布特征,结果如表 5 中 Best 所示。

6　结束语

本文采用多样性排序学习方法解决微博中多样性检索的问题,并针对微博的特点,设计一系列社交媒体特征。通过多样性排序学习模型,能方便地集成这些特征,并根据训练数据自动估算其权值。实验结果表明,多样性排序学习方法对于微博多样性检索是有效的,在微博检索中考虑多样性特征能够提高微博多样性检索的效果,微博的子话题特征和社交媒体特征能够帮助微博多样性问题的解决,其中博文与博文间的时间特征、博文与博文间的主题词特征和博文发布者间的地理位置特征作用明显。

参考文献

[1] AGRAWAL R, GOLLAPUDI S, HALVERSON A, et al. Diversifying Search Results[C]//Proceedings of the 2nd ACM International Conference on Web Search and Data Mining. New York, USA: ACM Press, 2009: 5-14.

[2] TEEVAN J, RAMAGE D, MORRIS M R. # Twitter Search: A Comparison of Microblog Search and Web Search[C]// Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York, USA: ACM Press, 2011: 35-44.

[3] OZSOY M G, ONAL K D, ALTINGOVDE I S. Result Diversification for Tweet Search[M]. Berlin, Germany, 2014: 78-89.

[4] ZHU Y, LAN Y, GUO J, et al. Learning for Search Result Diversification[C]//Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. New York, USA: ACM Press, 2014: 293-302.

[5] TAO K, HAUFF C, HOUBEN G J. Building a Microblog Corpus for Search Result Diversification[M]. Berlin, Germany, 2013: 251-262.

[6] JABEUR L B, TAMINE L, BOUGHANEM M. Uprising Microblogs: A Bayesian Network Retrieval Model for Tweet Search[C]//Proceedings of the 27th Annual ACM Symposium on Applied Computing. New York, USA: ACM Press, 2012: 943-948.

[7] NAVEED N, GOTTRON T, KUNEGIS J, et al. Searching Microblogs: Coping with Sparsity and Document Quality[C]//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2011: 183-188.

[8] ZHANG X, HE B, LUO T, et al. Query-biased Learning to Rank for Real-time Twitter Search[C]//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2012: 1915-1919.

[9] LUO Z, OSBORNE M, PETRROVIC S, et al. Improving Twitter Retrieval by Exploiting Structural Information[C]// Proceedings of AAAI'12. Washington D. C., USA: IEEE Press, 2012: 214-223.

[10] LUO Z, YU Y, OSBORNE M, et al. Structuring Tweets for Improving Twitter Search[J]. Journal of the Association for Information Science and Technology, 2015, 66(12): 2522-2539

[11] WENG J, LIM E P, JIANG J, et al. Twitterrank: Finding Topic-sensitive Influential Twitterers[C]//Proceedings of the Third ACM International Conference on Web Search and Data Mining. New York, USA: ACM Press, 2010: 261-270.

[12] EFRON M. Hashtag Retrieval in a Microblogging Environment[C]//Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2010: 787-788.

[13] 彭　敏,傅　慧,黄济民,等. 基于核主成分分析与小波变换的高质量微博提取[J]. 计算机工程, 2016, 42(1): 180-186.

[14] 李　敬,印　鉴,刘少鹏,等. 基于话题标签的微博主题挖掘[J]. 计算机工程, 2015(4): 30-35.

[15] OUNIS I, MACDONALD C, LIN J, et al. Overview of the Trec-2011 Microblog Track[C]//Proceedings of the 20th Text Retrieval Conference. Washington D. C., USA: IEEE Press, 2011: 367-376.

[16] SANTOS R L T, MACDONALD C, OUNIS I. Exploiting Query Reformulations for Web Search Result Diversification[C]//Proceedings of the 19th International Conference on World Wide Web. Washington D. C., USA: IEEE Press, 2010: 881-890.

[17] LIANG S, REN Z, WEERKAMP W, et al. Time-aware Rank Aggregation for Microblog Search[C]//Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. New York, USA: ACM Press, 2014: 989-998.

[18] CARBONELL J, GOLDSTEIN J. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries[C]//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 1998: 335-336.

[19] CHEN H, KARGER D R. Less is More: Probabilistic Models for Retrieving Fewer Relevant Documents[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2006:429-436.

[20] ZHAI C X, COHEN W W, LAFFERTY J. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval[C]//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2003:10-17.

[21] WANG J, ZHU J. Portfolio Theory of Information Retrieval[C]//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2009:115-122.

[22] 陈婷婷,黄春兰,吴胜利. 支持搜索结果多样化的排名算法比较研究[J]. 计算机工程,2016,42(10):45-50.

[23] DANG V, CROFT W B. Diversity by Proportionality: An Eelection-based Approach to Search Result Diversification[C]//Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2012:65-74.

[24] CARTERETTE B, CHANDAR P. Probabilistic Models of Ranking Novel Documents for Faceted Topic Retrieval[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York, USA: ACM Press, 2009:1287-1296.

[25] RADLINSKI F, DUMAIS S. Improving Personalized Web Search Using Result Diversification[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2006:691-692.

[26] VARGAS S, CASTELLS P, VALLET D. Explicit Relevance Models in Intent-oriented Information Retrieval Diversification[C]//Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2012:75-84.

[27] DANG V, CROFT B W. Term Level Search Result Diversification[C]//Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2013:603-612.

[28] 张乃洲. 基于时间点击图挖掘的查询建议方法[J]. 计算机工程,2015,41(5):191-196.

[29] ERKAN G, RADEV D R. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization[J]. Journal of Artificial Intelligence Research, 2004:22:457-479.

[30] OTTERBACHER J, ERKAN G, RADEV D R. Biased LexRank: Passage Retrieval Using Random Walks with Question-based Priors[J]. Information Processing & Management, 2009, 45(1):42-54.

[31] CLARKE C L A, KOLLA M, CORMACK G V, et al. Novelty and Diversity in Information Retrieval Evaluation[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2008:659-666.

编辑 索书志

(上接第 151 页)

参考文献

[1] 王 璐. 基于本体的个性化推荐系统[D]. 成都:电子科技大学,2013.

[2] 张艳梅,王 璐. 适应用户兴趣变化的社会化标签推荐算法研究[J]. 计算机工程,2014,40(11):318-321.

[3] 孙慧峰. 基于协同过滤的个性化 Web 推荐[D]. 北京:北京邮电大学,2012.

[4] LI Ruimin, LIN Hongfei, YAN Jun. Mining Latent Semantic on User-tag-item for Personalized Music Recommendation[J]. Journal of Computer Research & Development, 2014, 51(10):2270-2276.

[5] GUO C, WANG H. Improved Collaborative Filtering Algorithm Based on Tags[J]. Computer Engineering & Applications, 2016, 52(8):56-61.

[6] 赵 艳,王亚民,刘怀亮. 基于标签网络聚类的个性化资源推荐模型研究[J]. 情报杂志,2014(4):179-183.

[7] 万元元. 社会性标签系统的个性化资源推荐[D]. 天津:天津大学,2011.

[8] 夏平平,帅建梅. 基于相似度拓展与兴趣度缩放的协同过滤算法[J]. 计算机工程,2016,42(1):199-202.

[9] 王娅丹,李 鹏,金 瑜,等. 标签共现的标签聚类算法研究[J]. 计算机工程与应用,2015,51(2):146-150.

[10] HAN M, TANG C, DUAN L, et al. TF-IDF Similarity Based Method for Tag Clustering[J]. Journal of Frontiers of Computer Science & Technology, 2010, 4(3): 240-246.

[11] SYMEONIDIS P, NANOPOULOS A, MANOLOPOULOS Y. A Unified Framework for Providing Recommendations in Social Tagging Systems Based on Ternary Semantic Analysis[J]. IEEE Transactions on Knowledge & Data Engineering, 2010, 22(2):179-192.

[12] 吕 刚,郑 诚,胡春玲. 基于标签与深度本体的 Web 推荐方法研究[J]. 计算机工程,2015,41(12):156-160.

[13] 任看看,钱雪忠. 协同过滤算法中的用户相似性度量方法研究[J]. 计算机工程,2015,41(8):18-22.

[14] KIM H N, ALKHALDI A, SADDIK A E, et al. Collaborative User Modeling with User-generated Tags for Social Recommender Systems[J]. Expert Systems with Applications, 2011, 38(7):8488-8496.

[15] 张付志,常俊风,周全强. 基于模糊 C 均值聚类的环境感知推荐算法[J]. 计算机研究与发展,2013,50(10):2185-2194.

编辑 刘 冰