山东农业大学学报(自然科学版),2016,47(3):469-472 VOL.47 NO.3 2016
Journal of Shandong Agricultural University (Natural Science Edition) doi:10.3969/j.issn.1000-2324.2016.03.029

基于 LDA 与 WordNet 方法的微博排序

聂 丁
湖南商学院计算机与信息工程学院, 湖南 长沙 410205

摘 要: 微博搜索排序是近年来微博研究的热点之一。对于任意一个话题，它内容的生产者很容易达到成千上万个，甚至更多，产生的微博数更是不计其数，同时，也给关键字搜索的微博排序提出了更大的挑战。因此，本文提出了基于话题的用户权威值计算方法、基于WordNet的内容语义相似度方法，以及基于LDA的方法将输入关键词和所召回微博与其所属话题相关联，使用LearningToRank监督学习方法，学习一种排序策略。在此基础上，对提出的方案在实际数据集上分别对用户话题权威性、微博内容语义相似度、以及综合排序因素进行验证。

关键词: 微博排序; 语义相似度; 特征拟合
中图法分类号: TP391.3 **文献标识码**: A **文章编号**: 1000-2324(2016)03-0469-04

Ranking Sensitive Topics in a Micro-blog Based on LDA and WordNet Method

NIE Ding
School of Computer and Information Engineering/Hunan University of Commerce, Changsha 410205, *China*

Abstract: Microblog ranking is one of the hot research area in recent years. For any one topic, it is easy to reach thousands of producers or even more, the number of micro-blogs is countless, but also it comes with a greater challenge during searching keywords in micro-blog. In view of this, we proposed to incorporate topical authority of user, content similarity based on WordNet and topical relevance based on LDA algorithm between search keywords and microblogs that recalled to enhance the performance of microblog ranking with learning to rank related algorithm. On this basis, the user's topic authority, micro-blog content semantic similarity as well as the integrated ranking factors in a proposed project were verified on the actual data set.

Keywords: Microblog ranking; semantic similarity; feature fitting

随着微博、博客、论坛等在线社交网络的应用出现及迅猛发展，使得互联网的使用方式发生了深刻变革。在微博等社交媒体中，用户具有双重身份，既是数据信息的消费者也是数据内容的生产者[1]。鉴于此，微博搜索排序，是近年来微博研究领域的热点之一。对于如何在海量数据中挖掘出和搜索关键词高相关、含有信息量大、用户真正想要看到的微博，是非常具有现实性意义的，研究微博搜索排序的算法具有实用性意义[2]。

1 用户话题权威值的计算方案

1.1 数据集

1.1.1 *数据集评分方法* 本评分分为 3 个等级，分别为 3、2、1 分，其中，3 分为最高等级，2 分次之，1 分为最低等级。对每一条微博，评分准则如下：

1)如果包含信息与查询该微博的关键字非常相关，且带有很好的信息量，可评为 3 分；

2)如果包含信息与查询该微博的关键字比较相关，且附带有部分的信息量，可评为 2 分；

3)如果它包含的信息与查询该微博的关键字相关，且基本上不包含相关信息量；或者它基本与查询该微博的关键字无关，则评为 1 分。

1.1.2 *数据集评分情况* 由于每一个数据集的数据量大，且评分会耗费巨大的人力和物力，本文只是对数据集名为 Google 和 Healthcare 的进行评分，评分情况间表 1。

表 1 数据集 Google 和 Healthcare 的评分情况
Table 1 Scores in data sets of Google and Healthcare

数据集 Data set	1 分	2 分	3 分
Google	51.9%	29.1%	19.0%
Healthcare	37.5%	34.1%	28.4%

收稿日期: 2016-03-12 **修回日期**: 2016-04-28
作者简介: 聂 丁(1975-),男,湖南长沙人,本科,工程师,主要研究方向为计算机应用、计算机网络. E-mail:diablo@hnuc.edu.cn
数字优先出版:2016-05-29 http://www.cnki.net

1.1.3 排序评价指标 指标 NDCG 是在 DCG[3]的基础上，进行的一个改进，NDCG 综合考虑微博的得分和其所处排序后的位置，适用于对不同的 Query 的排序评价后进行比较。其计算方法如下：

$$NDCG_n = Z_n \sum_{i=1}^{n} \frac{2^{G_i} - 1}{\log_2(i+1)} \tag{1}$$

其中，n 表示经过重排序后的前 n 条微博，G_i 是重排序后的微博列表的第 i 条微博的得分，Z_n 是归一化因子，它使得 NDCG 的理想值为 1。

1.2 方案概述

本研究方案的思想是通过获取用户搜索关键词信息，将用户搜索关键词划分到某个话题，然后对微博搜索引擎按照时间顺序返回来的近几天最新结果，再在该话题上对所有的用户计算话题权威值（表征该用户的话题权威性），根据此话题权威值，再一次对搜索引擎返回的搜索结果进行重排序[4]。计算步骤如图 1 所示。

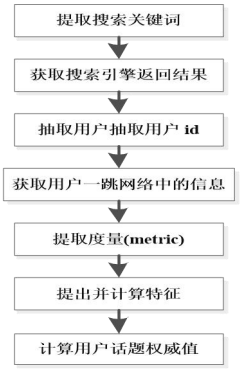

图 1 用户话题权威值计算步骤图

Fig.1 Calculation steps of user topic authority

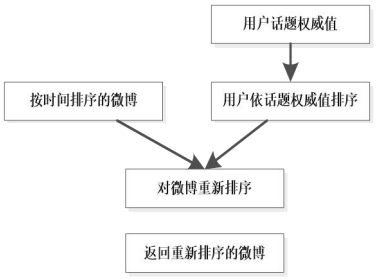

图 2 基于用户话题权威值的排序步骤

Fig.2 Ranking steps based on the user's authority

在此基础上，提出一种基于用户话题权威性的微博重排序方法，具体步骤如图 2 所示。

1.3 用户话题权威之计算方法

1.3.1 特征提取 根据用户话题权威性度量，构建 12 个相应的衡量用户话题权威性的特征，其中，TS 表示作者参与一个特定话题的程度，SS 用来衡量作者微博的原创性程度，同时也衡量作者的话题性强度[5]。另外，$\overline{CS}$ 用来衡量作者在在这个话题上发表微博的程度，以及作者从该话题跑题到会话的程度。则，

$$\overline{CS} < \frac{OT1}{OT1 + OT2} \tag{2}$$

这样，根据此不等式，有

$$\lambda < \frac{OT1}{OT1 + OT2} * \frac{CT1 + 1}{OT1 + CT1} \tag{3}$$

就求解出 λ。根据经验值，取 λ 满足 90%的用户，其中 λ 用于表示用户倾向于进入微博会话的程度。接下来，特征 RI 把作者的微博被转发的次数以及转发作者微博用户的个数考虑在内，用于衡量作者微博内容的影响力[6]。NS 综合考虑了在该话题上活跃的粉丝数与其关注的人中在该话题上活跃的数量，旨在估计在作者周围该话题的活跃程度。对于 OT21、OT41，是用来计算超链接以及 Hashtag 在作者原创微博中的出现的比率。OT3 用于计算作者在其所有的 n 条（包括该话题上以及该话题外）微博中，所使用的单词的重复度，其中，对于两个单词的集合，其相似度被定义为：

$$S(s_i, s_j) = \frac{\left| s_i \bigcap s_j \right|}{\left| s_i \right|} \tag{4}$$

其中，s_i，s_j 是由作者的第 i，第 j 条微博中通过去掉停用词以及做 Stem 之后得到的单词的集合，且在计算特征 OT3 之前，所有微博先按照时间排序，即 $times(s_i)<times(s_j):\forall_i<j$。观上来讲，对于一个特定的话题领域，在话题上用户粉丝的比率越大，该用户在该话题上的影响力就越大[7]。

1.3.2 计算方法 使用基于累积概率分布来计算每一个用户在该话题上的权威值，即 CDF_10 或 CDF_12 方法。对于用户 x_i，其话题权威值计算公式如下：

$$AS(x_i)=\prod_{f=1}^{m}F_f(x_i^f;\theta_f) \tag{5}$$

其中，其中 x_i 表示第 i 个用户，x_i^f 表示用户 i 在第 f 个特征上的值（f 取值范围为 1~12），F_f 表示参数为 θ_f 的第 f 个特征的累积概率分布函数在 x_i^f 处的 CDF 值，m 表示所用到的特征的个数。为了更好的逼近真实话题特征值，在以上话题权威值计算公式的基础上又提出了一种基于加权的计算公式，即 CDF_weighted 方法，其话题权威值计算公式如下：

$$AS(x_i)=\left[\prod_{f=1}^{11}F_f(x_i^f;\theta_f)\right]^{\beta}\left[F_{12}(x_i^{12};\theta_{12})\right]^{(1-\beta)} \tag{6}$$

根据微博用户权威值对微博重排序的具体流程如下：首先根据前面计算出的用户话题权威值按照从大到小的顺序对用户排序；其次根据用户的排名顺序对搜索引擎返回的按照时间顺序排列的微博进行重新排序，对于一个用户多条微博的情况，微博之间按照时间先后排序；最后将重新排序的微博结果返回给用户。

2　基于微博内容的语义相似度计算方案

2.1　方案概述

本研究方案的特点是，考虑微博的语义信息，并通过使用 WordNet 词典来计算两个单词之间基于语义的相似度，并在此基础上考虑单词的重要程度，即计算 TFIDF 值。基于语义的内容相似度计算方法示意图为：

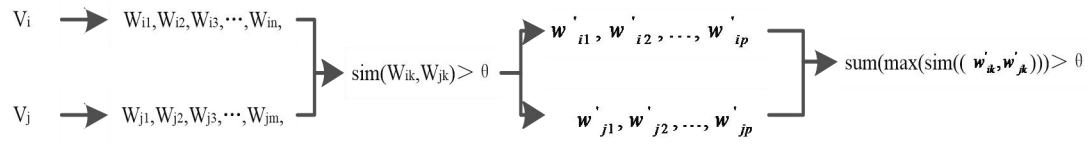

图 3 基于语义内容相似度计算方法示意图

Fig.3 Calculation method based on semantic content similarity

2.2　语义相似度计算方案

基于语义相似度的计算方法主要分为两个步骤，即首先计算出每一对微博的语义相似度，然后每一条微博与其他所有微博的相似度由它与其他微博相似度和来表示。对于 v_i 和 v_j 两条微博，即对于微博 v_i 中的每一个单词 w：

1)在微博 v_j 中找出一个单词的集合 Set，该集合中的每一个单词 u 都满足，它与单词 w 的语义相似度大于给定阈值，即使得 $sim(w,u)>\emptyset$，且 $w\in v_i$，$u\in v_j$；

2)取步骤 1)中集合 Set 中的一个使得 $sim(w,u)$取最大值的 u；

3)在步骤 2）的基础上，使用 TFIDF 值做权重，来计算两个单词之间的相似度 $v(w,v_i)*v(u,v_j)*sim(w,u)$；

4)将得到的两个单词之间的相似度相加，得到 $SIM(v_i,v_j)$。

由于以上得到的这个度量不具有自反性，即 $SIM(v_i,v_j)\neq SIM(v_j,v_i)$，采取以下操作：

$$SIM(v_i,v_j)=\frac{SIM(v_i,v_j)+SIM(v_j,v_i)}{2} \tag{7}$$

2.3　结果分析

实验结果表明，在没有考虑单词重要性的前提下，仅仅考虑语义信息，总体上来讲排序性能是不理想的。同时，可以得出一个结论，基于内容相似度的且不考虑语义信息的排序方法性能都不理想。综上所述，本研究课题提出的基于微博主题重要性的语义相似度计算方案在排序性能上总体优于仅仅考虑一方面即只考虑单词的重要性即基于 TFIDF 的方法，或者只考虑语义信息的排序方案。

3　对话题敏感的排序方案实现及分析

3.1　话题区分方法

本话题区分方法使用 LDA 文档主题模型实现，实现思想是首先判定微博搜索词所属话题，然后再根据微博搜索词所属话题将搜索引擎召回的微博在该话题上的分布情况进行判定，其实现主要步骤为：

1)随机选取数据集中 3/4 的数据作为训练集用于训练 LDA 模型；

2)使用 LDA 训练模型对搜索关键词和其余 1/4 的数据集进行推断（即 Inference）操作，得到搜索关键词及测试集中每一条微博在所有话题上的概率分布；

3)求搜索关键词在所有话题上分布的最大值，并将其话题编号求出；

4)根据步骤 3)所求得的话题编号，对所有被搜索引擎召回的微博取其在该话题上的分布概率值，并进行归一化。

将上述步骤最终得到的向量作为微博排序的一个特征，将其称为话题关联特征，作为微博排序的一个因素。

3.2　排序方案

本文采用基于统计的Learning To Rank即排序学习方法对微博进行排序，其中使用9个特征，即<用户话题权威值，微博内容相似度，时间相近性，转发次数，微博长度，超链接数量，标签数量，@数量，关键词与微博的话题关联性>。这些特征主要可以分为三个维度，即用户维度、微博内容维度以及微博自身维度，如图4所示。

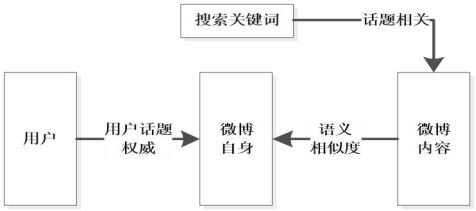

图 4 排序方案图解

Fig.4 Ranking steps

4　结 论

针对现有微博搜索排序的不足，本文针对用户权威值的计算方面，提出基于话题的用户权威值计算方法。综合考虑用户话题权威性以及传统的权威度量，提出了最终的话题权威值计算公式。针对现有微博内容相似度计算方案的不足，本文提出了基于单词重要性的语义相似度计算方案。该语义相似度计算方案首先考虑单词之间的语义相似性，在此基础上再考虑单词的重要性。对于提出的以上两个研究方案，本文对它们分别从理论可行性和现实可行性两方面进行了分析，并与已有的研究方案进行比较，从实验的角度证明了所提方案的排序性能的高效性。最后，考虑到搜索关键词和所召回微博之间的话题相关性，提出基于LDA的话题区分方法，将搜索关键词及微博关联起来，并作为一个排序特征，应用到基于排序学习的框架中，通过与基准的排序方案相比较，从实验的角度证实了所提特征的有效性。

参考文献

[1] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003(3):993-1022

[2] Griffiths TL, Steyvers M. Finding scientific topics[J]. Proceedings of the National academy of Sciences of the United States of America, 2004,101(S1):5228-5235

[3] 王 晟,王子琪,张 铭.个性化微博推荐算法[J].计算机科学与探索,2012,6(10):895-902

[4] Friedman JH. Stochastic gradient boosting[J]. Computational Statistics & Data Analysis, 2002,38(4):367-378

[5] Mahinthan V, Rutagemwa H, Mark JW, *et al.* Cross-layer performance study of cooperative diversity system with ARQ[J]. IEEE Transactions on Vehicular Technology, 2009,58(2):705-719

[6] Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language[J]. Sensor Fusion & Decentralized Control in Robotic Systems III, 2011,11(1):95-130

[7] 时晓飞.从最小省力原则来看微博[J].才智,2014(1):309