

融合内容相似度与多特征计算的个性化微博推荐模型

刘宇东 孙 豪 蒋运承

华南师范大学计算机学院 广州 510631



摘 要 微博的流行导致信息过载等问题日益突出,如何帮助用户快速而准确地找到需要的微博已成为亟待解决的问题。基于协同过滤技术和基于 LDA 的微博推荐虽然能够达到一定的准确性,但并不能解决内容分类过于笼统及使用 LDA 模型处理短文本存在弊端的问题。为此,文中提出了一种融合内容相似度与多特征计算的个性化微博推荐模型。首先,从微博内容语义出发,基于 word2vec 技术计算得到用户与微博的内容相似度;然后,根据微博的时间、点赞数、评论数和转发数等特征,计算微博的保鲜度及受欢迎度;最后,综合考虑微博的内容相似度、保鲜度和受欢迎度,计算微博排序评分,从而实现用户的个性化微博推荐。该模型根据内容相似度进行推荐,从而避免了上述问题,也使得推荐结果在语义上更为精确。实验结果表明,所提推荐模型在准确率、召回率和 F 值上均具有良好的表现,尤其在准确率方面有明显的提升效果,约提升了 10%,F 值也提升了约 5%,从而证明了该模型的有效性。

关键词: 微博;word2vec;相似度;保鲜度;受欢迎度

中图法分类号 TP391

Personalized Microblog Recommendation Model Integrating Content Similarity and Multi-feature Computing

LIU Yu-dong, SUN Hao and JIANG Yun-cheng

School of Computer Science, South China Normal University, Guangzhou 510631, China

Abstract With the popularity of microblog, problems such as information overload are increasingly prominent. How to help users find the microblog they need quickly and accurately has become an urgent problem to be solved. Although microblog recommendation based on collaborative filtering technology and LDA can achieve certain accuracy, it can not solve the problems of general classification of content and the disadvantages when LDA model is used to deal with short texts. Therefore, this paper proposes a personalized microblog recommendation model integrating content similarity and multi-feature computing. Firstly, the content similarity between user and microblog is calculated based on word2vec. Then, according to the characteristics such as time, number of likes, comments and reposts, the freshness and popularity of microblog are calculated. Finally, the content similarity, freshness and popularity of microblog are comprehensively considered to calculate its ranking score, so as to realize users' personalized microblog recommendation. This model considers recommendation from the perspective of content similarity, avoiding the above problems and making the recommendation results more accurate in semantics. Experimental results show that the proposed model has good performance in accuracy, recall rate and F-measure, in particular, the accuracy has been significantly improved by about 10%, and F-Measure is increased by about 5%, and the validity of the model is proved.

Keywords Microblog, word2vec, Similarity, Freshness, Popularity

随着互联网及网络服务的快速发展,网络社交媒体已成为人们彼此之间分享成果、介绍经验、发表观点、相互沟通的重要工具和平台。微博作为网络社交媒体的杰出代表,以信息简短、传播速度快、更新时间短的优势,迅速获得了人们的青睐。截至 2019 年 3 月底,新浪微博月活跃用户数增至 4.65 亿,每天有大量的微博信息被发布和转发。以 2018 年世界杯为例,有 1 亿人在微博上参与了相关讨论,总互动量超

过 10 亿,短视频播放量超过 170 亿,话题阅读量近 1 000 亿。因此,在海量的微博信息中及时选择并推荐用户真正需要的信息已显得越来越重要。

近年来,关于微博推荐的研究已有不少,具体方法可以大致分为两种:1)基于协同过滤技术进行推荐,通常侧重于研究相似用户群体对目标用户的影响^[1-2],也有些研究主要依据用户间的社交关系^[3-5];2)从微博内容出发考虑推荐,这种方法

到稿日期:2019-07-09 返修日期:2019-09-03 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:国家自然科学基金面上项目(61772210);广州市科技计划项目(201807010043)

This work was supported by the National Natural Science Foundation of China (61772210) and Guangzhou Science and Technology Project (201807010043).

通信作者:刘宇东(cnlyd@163.com)

其实是协调过滤技术的延续和发展,如 An 等提出基于内容的热门微话题个性化推荐^[6]。后来,不少学者引入 LDA 模型,Zhao 等提出了一种适用于短文本的 Twitter-LDA 模型,并提出了“单微博-单主题”的主题构建思想^[7]。Ben-Lhachemi 等提出通过使用 tweets 的语义嵌入表示,帮助用户实时为自己的帖子选择相关的推文话题,然后捕获 tweets 之间的语义相似性或关联性,从而实现推文话题的推荐^[8]。

另外,有学者在研究微博推荐时综合考虑了用户特征、网络社交结构以及微博博文内容。Chang 等将用户的博文信息与网络结构结合,设计了一套社交推荐系统架构,用于发现用户感兴趣的商品^[9]。Wang 等结合网络拓扑结构与用户的属性信息,提出一种社交圈检测算法,并根据用户间社交圈的相似性实现好友的推荐^[10]。Li 等从最受欢迎、最值得信赖和最相似的链接或微博内容等多个方面进行考虑,提出综合信任和社会网络关系的推荐模型,并应用 LDA 主题模型及矩阵分解技术推断微博的主题分布和用户的兴趣取向,实现微博的个性化推荐^[11]。Wang 等结合使用用户关系和互动行为信息,提出了一种针对非名人用户的新型关注推荐算法^[12]。

以上研究通常存在两个问题:1)往往将新用户或者沉默用户(即最近一段时间内没有任何操作的用户)与其他用户无区别对待,导致推荐的信息没有根据,较为盲目;2)在内容上大多根据用户兴趣并按照通用的分类进行推荐,但这些分类比较笼统,如某用户喜爱 NBA,则往往将其归类于篮球运动乃至整个体育运动大类,结果自然不甚理想。有些学者通过引入 LDA 模型来解决该问题,但 LDA 模型在处理微博短文本方面存在先天不足,以致效果不甚理想。

针对以上问题,本文提出一种融合内容相似性与多特征计算的个性化微博推荐模型。该模型运用 word2vec 技术从内容相似度的角度进行推荐,从而避免了通用分类过于笼统以及使用 LDA 模型处理短文本存在弊端的问题,也使得推荐结果在语义上更为精确。同时,该模型综合多特征计算得到的微博保鲜度及受欢迎度(包括微博的时间、点赞数、评论数和转发数等特征),实现了个性化的推荐。另外,该模型结合了用户的地理位置、年龄、性别等特征,有效解决了新用户或者沉默用户的推荐问题。

1 word2vec 简介

word2vec 是 Google 在 2013 年提出的将词表示为实数值向量的技术,通过训练可以把对文本内容的处理简化为给定维数向量的运算,而向量空间上的相似度可以用来表示文本语义上的相似度^[13]。

word2vec 技术训练生成词向量的基本思想来自于 Ben-gio 提出的神经网络语言模型(Neural Network Language Model, NNLM)。然而,NNLM 存在一个严重问题,即训练速度太慢。对此,Mikolov 等对其进行了针对性改进和优化,例如,去掉了隐藏层,将输出层改用 Huffman 树等。

word2vec 技术中,词向量作为神经概率语言模型的输入,其本身是神经概率模型的副产品,是为了通过神经网络学习某个语言模型(“CBOW”或“Skip-gram”)而产生的中间结果。该学习过程中可以选取两种降低复杂度的近似方法:

Hierarchical Softmax 或 Negative Sampling。

使用 word2vec 训练学习得到一个文档中的词向量后,可以进一步获得该文档的向量^[14]。另外,Google 在 2014 年提出了专门进行文档向量化的工具 doc2vec,其训练和预测过程与 word2vec 模型相似,区别在于前者加入了文档 ID 向量,在一个文档的训练过程中,文档 ID 保持不变,共享着同一个文档向量,这相当于在预测单词的概率时,都利用了真实句子的语义。一般而言,当训练数据集是普通文本时,使用 doc2vec 的效果较好;但如果遇到诸如微博、短信等短文本时,由于语义环境不明显,其效果通常不如词向量累加或者取平均值的方法^[15]。

2 融合内容相似性与多特征计算的个性化微博推荐模型

2.1 推荐模型

本文提出的融合内容相似性与多特征计算的个性化微博推荐模型如图 1 所示。

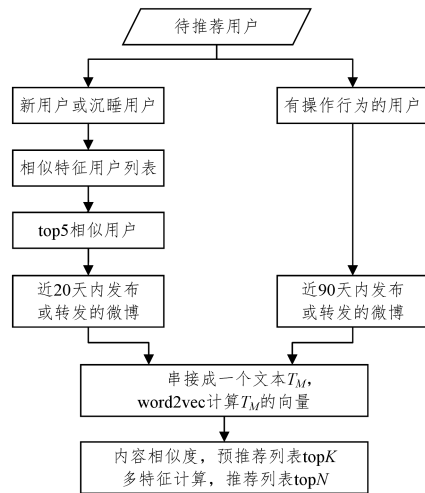


图 1 推荐模型的流程

Fig. 1 Flow chart of recommended model

算法 1 给出了推荐模型的具体过程。

算法 1 融合内容相似性与多特征计算的个性化微博推荐算法

Input: 待推荐用户 u

Output: 待推荐用户 u 的 topN 微博推荐列表

Step1 如果用户 u 是新用户或沉默用户(参考文献[16],以 90 天内是否存在操作行为作为判断依据),则根据 u 的特征集 $\{l(u), a(u), g(u)\}$,选择与 u 相似度大且存在操作行为的前 5 位用户。选择原则:如果用户相似度大小相同,则随机截取这部分用户中近期(20 天内)存在操作行为的 5 位。

如果该用户 u 是近期(90 天内)有发布博文操作行为的用户,则转到 Step3。

Step2 获取这 5 位用户 20 天内发布或转发的微博,记这些微博组成的集合为 $M = \{m_1, m_2, \dots\}$ 。

Step3 如果 u 是新用户或沉默用户,将 Step2 得到的集合 M 的所有微博博文串接为一个文本 T_M ,再进行分词、去停用词、提取词干等预处理操作。

如果 u 是非沉默用户,同样将其 90 天内发布或转发的微博组

成集合 $M = \{m_1, m_2, \dots\}$, 并将所有博文串接为一个文本 T_M 。

然后, 利用 word2vec 计算 T_M 的文本向量 $\vec{T_M}$ 。

Step4 利用 word2vec 计算待推荐微博 m 的文本向量, 并计算 m 与 T_M 的文本向量 $\vec{T_M}$ 的余弦相似度, 从而得到用户 u 与微博 m 的内容相似度。据此, 取 topK 组成预推荐列表。

Step5 在 topK 预推荐列表中, 计算排序分数, 从而得到待推荐用户 u 的 topN 微博推荐列表。

2.2 内容相似度

微博文本经过数据清理、分词、去停词、提取词干等预处理后, 就可以使用 word2vec 工具包进行处理, 从而得到表征每个词语的给定维数的向量。然后进一步计算微博文本的向量, 为简单起见, 用微博文本中词语向量的平均值来表征该微博文本的向量^[15]。

给定微博文本 T , 对其进行分词等预处理后, 记 $T = \{t_1, t_2, \dots, t_i, \dots, t_K\}$, 对于 $\forall t_i \in T$, 假定其经过 word2vec 处理后得到的对应词向量为 $\vec{t_i}$, 则 T 的文本向量为:

$$\vec{T} = AVERAGE(\vec{t_1}, \vec{t_2}, \dots, \vec{t_K}) \quad (1)$$

定义 1 假设微博 m 与 n 的博文经分词、去停词、提取词干等预处理后, 通过 word2vec 工具包处理并计算得到的给定维数的文本向量分别为 $\vec{m}$ 和 $\vec{n}$, 则定义微博 m 与 n 的内容相似度 $sim(m, n)$ 为向量 $\vec{m}$ 和向量 $\vec{n}$ 的余弦值, 即 $sim(m, n) = \cos(\vec{m}, \vec{n})$ 。

定义 2 将用户 u 在最近给定时间内发布或转发的微博博文串接成一个文本, 记为 T_u 。假设对于任意微博 m 和 T_u , 通过 word2vec 工具包处理并计算得到给定维数的文本向量分别为 $\vec{m}$ 和 $\vec{T_u}$, 则定义微博 m 与用户 u 的内容相似度 $sim(u, m)$ 为向量 $\vec{m}$ 和向量 $\vec{T_u}$ 的余弦值, 即 $sim(u, m) = \cos(\vec{m}, \vec{T_u})$ 。

2.3 多特征计算

2.3.1 用户相似度的计算

定义 3 用户 u 的特征集是一个三元组 (l, a, g) , 其中 l 表示用户所在地理位置, 取值是所在的省份和城市; $a \in A = \{a_1, a_2, a_3, a_4, a_5\}$ 表示年龄, 根据微博用户的实际情况, 年龄取值分别代表以下年龄段: 13~18 岁, 19~24 岁, 25~35 岁, 35~50 岁, 50 岁以上; $g \in G = \{g_1, g_2\}$ 表示性别, 取值为男、女。

研究表明, 特征相同或相似的用户往往存在相似的内容偏好。比如, 相同地点的用户可能有比较相近的选择; 大部分年轻女性可能会偏好化妆、健身、购物等类型的信息。

定义 4 用户相似度用于反映用户之间特征的相似程度。用户 u 和用户 v 的相似度用 $Sim(u, v)$ 表示:

$$Sim(u, v) = r_1 l(u, v) + r_2 a(u, v) + r_3 g(u, v) \quad (2)$$

其中, r_1, r_2 和 r_3 为权值, $r_1 + r_2 + r_3 = 1$; $l(u, v)$ 是用户地理位置相似度, 如果 u 和 v 的省份相同则取值为 0.6, 如果城市相同则取值为 1, 省份和城市都不同则取值为 0; $a(u, v)$ 是用户年龄相似度, 如果 u 和 v 在相同年龄段则取值为 1, 否则取值为 0; $g(u, v)$ 是用户性别相似度, 如果 u 和 v 性别相同则取值为 1, 否则取值为 0。

2.3.2 微博保鲜度

定义 5 微博的保鲜度定义为一个时间效用函数:

$$f(t) = (1 + \alpha)^{-t} \quad (3)$$

其中, $f(t)$ 的取值范围为 $(0, 1]$, $\alpha (\alpha > 0)$ 是调整因子, t 为微博的发布时间与当前时间的间隔天数。

式(3)中的调整因子 α 是一个表征微博保鲜度变化趋势的量, 其值越小, 保鲜度变化趋势越小, 意味着微博信息的发布时间对微博保鲜度的影响越小; 其值越大, 保鲜度变化趋势越大, 意味着微博信息的发布时间对保鲜度的影响越大。特别地, 当 $\alpha = 0$ 时, $f(t) = 1$, 表示微博信息发布时间的先后完全不影响保鲜度, 即任何时候发布的微博的保鲜度都一样^[17]。

研究认为, 微博的保鲜时间为 16.8 天, 所以 α 一般取值为 0.1~0.3, 本文中取 $\alpha = 0.2$ 。

2.3.3 微博受欢迎度

用户对微博的操作行为包括浏览、点赞、评论、转发等。用户对微博的喜好程度可以通过他们对微博的操作行为进行分析。例如, 用户转发了一条微博, 会被认为比点赞行为存在着更浓厚的兴趣。由于数据集并没有记录用户是否浏览过某条微博的信息, 因此主要从点赞、评论和转发这 3 种操作行为进行分析。

定义 6 微博操作值 B 是反映所有用户对微博的操作行为的量, 通过用户对微博的点赞、评论、转发这 3 个行为来衡量。微博 m 的微博操作值记为:

$$B(m) = sum(m) = \lambda_1 a + \lambda_2 b + \lambda_3 c \quad (4)$$

其中, a, b, c 分别表示微博 m 的点赞数、评论数和转发数; $\lambda_1, \lambda_2, \lambda_3$ 是权重系数。本文主要参考文献[18], 取 $\lambda_1 = 1, \lambda_2 = 2, \lambda_3 = 3$ 。

定义 7 有微博集合 $M = \{m_1, m_2, \dots, m_i, \dots, m_K\}$, 假设微博按 $B(m_i)$ 降序排列后为 $m'_1, m'_2, \dots, m'_j, \dots, m'_K$ (如果几个微博的操作值相同, 那么它们在降序排列后的下标相同, 即假如 $B(m_i) = B(m_s)$, 则它们排序后为 $m'_j(m_i), m'_j(m_s)$), 则微博受欢迎度定义为:

$$\begin{cases} Q(m'_1) = 1 \\ Q(m'_j) = Q(m'_{j-1}) - \frac{(B(m'_{j-1}) - B(m'_j))}{B(m'_1)}, j = 2, 3, \dots \end{cases} \quad (5)$$

2.4 微博排序评分

在融合内容相似度与多特征计算的个性化微博推荐模型中, 综合考虑微博博文的内容相似度、微博保鲜度和微博受欢迎度计算排序评分, 进而进行推荐。对待推荐用户 u 而言, 微博 m 的排序评分按式(6)进行计算:

$$Sort(m) = \omega_1 sim(u, m) + \omega_2 f(t) + \omega_3 Q(m) \quad (6)$$

其中, ω_1, ω_2 和 ω_3 为权重系数, $\omega_1 + \omega_2 + \omega_3 = 1$ 。

3 实验与结果分析

本文使用新浪微博提供的 API 采集的 2014 年 2 月 10 日至 2014 年 5 月 12 日的微博数据进行实验, 数据包含 6 万多名用户的 84168 条微博信息。该数据集按用户 ID 被随机分为 80% 的训练集和 20% 的测试集。

实验中的 word2vec 工具包选用 Gensim, 这是一款开源的第三方 Python 工具包。它支持包括 word2vec 在内的多种

模型算法,支持流式训练,并提供了诸如相似度计算、信息检索等一些常用任务的 API 接口。

分词工具则使用自然语言工具箱 JIEBA,它是一个基于 Python 语言的类库。在进行中文语言处理研究和应用时,恰当地利用 JIEBA 中提供的函数可以大幅度地提高效率。JIEBA 主要用来对中文文本进行分词、去停用词、提取词干等预处理操作。

实验采用的硬件环境为 Core i7-8700 CPU,8 GB DDR4 内存,1 TB 硬盘,128 G SSD;软件环境为 Windows 10 操作系统, Mysql 5.7 数据库, Python 3.7, Gensim, JIEBA 开发平台。

3.1 评价指标

微博推荐系统一般使用准确率(Precision)、召回率(Recall)、F 值(F-Measure)来衡量推荐结果的有效性。

对于待推荐用户 u ,假设 RL 为模型推荐的微博列表中 u 喜好的微博个数, N 为推荐列表长度, $L(u)$ 为用户 u 喜好的微博个数,则其推荐准确率、召回率和 F 值分别为^[19]:

$$P = \frac{RL}{N}$$
(7)

$$R = \frac{RL}{L(u)}$$
(8)

$$F = \frac{2PR}{P+R}$$
(9)

微博中没有明确表明用户喜好的数据,只能从用户所发表微博的内容来判断用户对该微博是否喜好。本文把推荐给某用户的微博列表中内容与其所发表的微博集吻合的都认为是用户喜好的微博。

3.2 评估结果

融合内容相似度与多特征计算的个性化微博推荐模型在计算用户相似度和排序评分时含有 r_1, r_2, r_3 及 $\omega_1, \omega_2, \omega_3$ 两组权重参数,参数的取值将影响模型推荐结果的质量。

3.2.1 计算用户相似度的特征权值

式(2)中的 r_1, r_2 和 r_3 为特征权值, $r_1 + r_2 + r_3 = 1$ 。

根据特征计算用户相似度时,各特征权值的确定通常以特征的不同取值占特征出现次数的比例作为权衡标准,即特征的不同取值越多,该特征在相似度计算时的贡献越大,权值也就越大^[20-21]。例如,微博用户的性别特征有 2 种取值,年龄特征有 5 种取值,因此年龄相较于性别对用户分类的重要性更大,相应的权值也更大;而地理位置特征的取值则更多,在本文使用的数据集中其取值达 550 多种,故地理位置权值应最大。于是,在此基础上选取若干种权值组合情况(如表 1 所列)进行实验。对于将用户相似度计算应用于对新用户或沉默用户的微博推荐,已有文献大多没有相关的实验验证描述,即使有提及,一般也都使用人工标注的方法。本文实验从测试集中选取 20 名沉默用户,通过本文模型计算出推荐结果后,根据这些沉默用户的特征分别找到对应特征的真实用户(大学本科生、研究生和教职工等),由真实用户对所推荐的微博进行喜好与否的标示。统计这 20 名用户的平均数据,结果如图 2 所示。从图中可以看出,设定 2(即 $r_1 = 0.6, r_2 = 0.25, r_3 = 0.15$)的推荐效果最好,此时准确率为 72.35%,召回率为 77.52%, F 值为 74.85%。该结果同时显示了所提模型在解决新用户或沉默用户的微博推荐问题上表现还不错。

表 1 用户特征权值的设定

Table 1 User feature weight setting

	r_1	r_2	r_3
设定 1	0.60	0.30	0.10
设定 2	0.60	0.25	0.15
设定 3	0.50	0.30	0.20
设定 4	0.50	0.04	0.10
设定 5	0.40	0.35	0.25

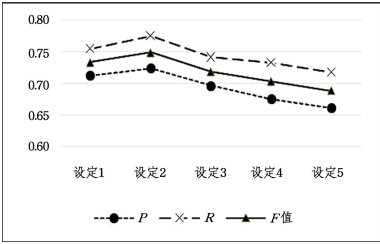


图 2 不同用户特征权值下的 P, R 和 F 值

Fig. 2 Precision, Recall and F-Measure under different user feature weights

3.2.2 微博排序评分系数权值

式(6)中的 ω_1, ω_2 和 ω_3 为权重系数, $\omega_1 + \omega_2 + \omega_3 = 1$ 。

微博排序分数综合了语义相似度、微博新鲜度和微博受欢迎度。由于本文提出的推荐模型主要考虑微博语义内容,因此语义相似度在排序时应当起主要作用;而考虑新鲜度和受欢迎度则保证了推荐结果在时间及传播热度上的领先性。为此,选取几种典型权值组合情况(如表 2 所列)进行实验,从测试集中选取 50 名用户,分别计算出推荐结果后,统计这 50 名用户的平均数据,结果如图 3 所示。可以发现,设定 3(即 $\omega_1 = 0.5, \omega_2 = 0.3, \omega_3 = 0.2$)下模型的推荐效果最好。

表 2 微博排序评分系数权值

Table 2 Microblog ranking score coefficient weights

	ω_1	ω_2	ω_3
设定 1	0.60	0.25	0.15
设定 2	0.60	0.30	0.10
设定 3	0.50	0.30	0.20
设定 4	0.50	0.20	0.30
设定 5	0.40	0.30	0.30

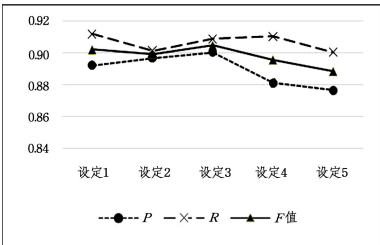


图 3 不同微博排序评分系数权值下的 P, R 和 F 值

Fig. 3 Precision, Recall and F-Measure under different microblog ranking score coefficient weights

3.3 与现有方法的对比

绝大多数已有研究所使用的微博数据集的内容(包括时间、字段等)都不相同,各自计算准确率、召回率和 F 值的方法也不相同,其中还有不少研究采用了人工标注是否喜好微博的方法。实验从测试集中选取 100 名用户,根据给这 100 名用户的推荐结果计算准确率、召回率和 F 值并取平均值。

将 He 等^[2]提出的推荐模型和 Li 等^[11]提出的推荐方法作为对比方法,实验结果如图 4 所示。

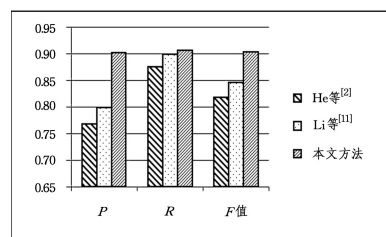


图 4 实验结果

Fig. 4 Experimental results

实验结果表明,本文方法在准确率、召回率和 F 值方面均有提升,尤其在准确率方面提升明显,约提升了 10 个百分点, F 值提升了约 5 个百分点。同时,推荐列表 topN 结果显示,本文方法的推荐结果能够较好地体现最新的各 topic 的微博及当前反应热烈的微博。

结束语 本文提出一种融合内容相似度与多特征计算的个性化微博推荐模型。该模型先基于 word2vec 技术进行一系列计算得到微博文本向量及相关的内容相似度,然后根据内容相似度及多特征计算得到的微博保鲜度、受欢迎度获取待推荐用户的微博推荐 topN 列表。另外,本文由用户的特征集导出相似用户集合,并从相似用户的偏好出发考虑新用户或沉默用户的微博推荐。实验表明,所提模型具有较好的表现。然而,微博博文作为短文本,存在信息含量少、稀疏、用词不规范等问题,需要在以后有针对性地进行相应的处理;同时,在短文本向量表示上,需要寻找更适合微博的方法,以进一步提高模型的有效性。

参 考 文 献

- [1] KARKADA U H. Friend recommender system for social networks[R]. SI583 Term Paper, School of Information, University of Michigan, 2009.
- [2] HE Y, TAN J. Study on SINA micro-blog personalized recommendation based on semantic network [J]. Expert Systems with Applications, 2015, 42(10): 4797-4804.
- [3] LIBEN-NOWELL D, KLEINBERG J. The link-prediction problem for social networks [J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031.
- [4] GUIMERA R, SALES-PARDO M. Missing and spurious interactions and the reconstruction of complex networks [J]. Proceedings of the National Academy of Sciences, 2009, 106(52): 22073-22078.
- [5] GUO L, MA J, CHEN Z M, et al. Incorporating Item Relations for Social Recommendation [J]. Chinese Journal of Computers, 2014, 37(1): 219-228.
- [6] AN Y, LI B, YANG R T, et al. Content-based Personalized Recommendation on Popular Micro-topic [J]. Journal of Intelligence, 2014(2): 155-160.
- [7] ZHAO W X, JIANG J, WENG J, et al. Comparing twitter and traditional media using topic models [M]// Advances in Information Retrieval. Springer Berlin Heidelberg, 2011: 338-349.
- [8] BEN-LHACHEMI N, NFAOUI E H. Using Tweets Embed-

dings For Hashtag Recommendation in Twitter [J]. Procedia Computer Science, 2018, 127: 7-15.

- [9] CHANG P S, TING I H, WANG S L. Towards social recommendation system based on the data from microblogs [C]// 2011 International Conference on ASONAM 2. 11: Advances in Social Networks Analysis and Mining. IEEE, 2011: 672-677.
- [10] WANG Y, GAO L. Social Circle-Based Algorithm for Friend Recommendation in Online Social Networks [J]. Chinese Journal of Computers, 2014, 37(4): 801-808.
- [11] LI H, MA X P, SHI J, et al. Microblog Recommendation by Trust and Social Relationship [J]. Journal of Chinese Information Processing, 2017, 31(2): 146-153.
- [12] WANG M J, HE Z M, ZHENG J. Microblog followee recommendation algorithm combining with trust and user relationship [J]. Application Research of Computers, 2018, 35(12): 46-49.
- [13] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space [C]// ICLR. 2013.
- [14] TANG M, ZHU L, ZOU X C. Document Vector Representation Based on Word2Vec [J]. Computer Science, 2016, 43(6): 214-217.
- [15] DUAN X L, ZHANG Y S, SUN Y Z. Research on Sentence Vector Representation and Similarity Calculation Method About Microblog Texts [J]. Computer Engineering, 2017, 43(5): 143-148.
- [16] TANG X B, LIANG M J. Research of Silent User Interest Modeling in Microblog Based on the Features of Structure and Content [J]. Journal of the China Society for Scientific and Technical Information, 2015(11): 1214-1224.
- [17] DING Y, XUE L. Time weight collaborative filtering [C]// Proceedings of the 14th ACM International Conference on Information and Knowledge Management. Bremen, Germany: ACM, 2005.
- [18] QI C, CHEN H C, YU H T. Method of evaluating micro-blog users' influence based on comprehensive analysis of user behavior [J]. Application Research of Computers, 2014, 31(7): 2004-2007.
- [19] SHI L, TAO Y C, LI J Y, et al. Personalized and Real-time Recommendation Model for Microblogs [J]. Journal of Chinese Computer Systems, 2016, 37(9): 1910-1914.
- [20] ZHENG Z Y, JIA C Y, WANG Z F, et al. Computing Research of User Similarity Based on Micro-blog [J]. Computer Science, 2017(2): 262-266.
- [21] YU X S, SUN S. Research on Personalized Recommendation Model Based on Network Users' Information Behavior [J]. Journal of Chongqing University of Technology (Natural Science), 2013, 27(1): 47-50.



LIU Yu-dong, born in 1974, master, lecturer. His main research interests include natural language processing and machine learning.