视新浪博客的最近更新列表,下载程序间歇性地抓取了 19 427 位用户 2015 年 7 月 16 日~2016 年 8 月 17 日发布的微博,并存储在数据库中.这些数据库记录包括了博文的标签、发布时间、微博内容、微博转发数、评论数及点赞数等信息.

对实验数据进行预处理,首先过滤文本中的@用户名、地址链接、和其他无意义字符等噪声信息后,对其进行分词,去除停用词.其中,分词采用 python 开源分词组件 jieba,停用词表采用新浪提供的 1 208 个停用词.实验中随机选取 13 000 名用户,删除拥有少于 4 个词汇的微博以及拥有少于 50 篇微博的用户,得到最终的实验数据集合.数据集中有用户 10 390 名,微博 2 186 283 条,标签个数 5 897 个.新浪微博允许用户最多添加 10 个关键词对自己进行描述,表 2 统计了数据集中添加不同标签数量的用户分布,其中,48.3%的用户至少添加了一个标签,而 51.7%的用户没有为自己添加标签,这充分表明了标签扩充对进行微博推荐的必要性.

Table 2 Distribution of users with different number of tags in dataset

表 2 数据集中添加不同标签个数的用户分布

标签数量	0	1	2	3	4	5	6	7	8	9	10
用户个数	5 371	964	427	373	312	293	364	196	178	267	1 645

为了验证推荐算法的准确性,将微博数据集分为训练集和测试集:训练集用来学习推荐方法中的相关参数,测试集用来验证推荐算法的准确性.为了避免数据过拟合,本文采用十折交叉验证的方法,将每个微博用户的数据样本随机划分成 10 个大小相等的子样本集,交叉验证过程重复 10 次.每次一个样本集被保留作为测试集的验证数据,其余 9 个样本集作为训练数据,其中,训练集中有 1 967 655 条样本,测试集中有 218 628 条样本.

本文实验环境为:Windows 7 操作系统,4GB 内存,Intel Core(TM) 2 Duo CPU 2.66GHz,实验程序使用 Java1.6 语言开发,数据库为 Mysql5.0.

准确率(precision)、召回率(recall)、*F*1 值(*F*-measure)和平均正确率(average precision,简称 AP)是广泛用于信息检索和推荐领域的 4 个度量值,用来评价结果的质量.为了评估微博推荐质量,本文采用前 L 条结果的准确率 $P@L$、前 L 条结果的召回率 $R@L$、前 L 条结果的 $F1$ 值 $F1@L$ 以及前 L 条结果的平均正确率 $AP@L$ 来评价微博推荐质量.$P@L$,$R@L$,$F1@L$ 和 $AP@L$ 定义如下.

$$P@L=\frac{1}{|User|}\sum_{i=1}^{|User|}\frac{\text{前}L\text{条推荐结果中用户}u_i\text{喜欢的微博个数}}{\text{算法向用户}u_i\text{推荐的微博个数}L} \tag{26}$$

$$R@L=\frac{1}{|User|}\sum_{i=1}^{|User|}\frac{\text{前}L\text{条推荐结果中用户}u_i\text{喜欢的微博个数}}{\text{用户}u_i\text{测试数据集中的微博个数}} \tag{27}$$

$$F1@L=\frac{1}{|User|}\sum_{i=1}^{|User|}\frac{2\times P@L\times R@L}{P@L+R@L} \tag{28}$$

$$AP@L=\frac{1}{|User|}\sum_{i=1}^{|User|}\frac{\sum_{k=1}^{L}P@k\times rel(k)}{\text{前}L\text{条推荐结果中用户}u_i\text{喜欢的微博个数}} \tag{29}$$

其中,$rel(k)$是一个指示函数,当推荐返回结果的第 k 个位置为相关微博,$rel(k)=1$;否则,$rel(k)=0$.微博中没有明确表明用户喜好的数据,本文中把用户发布的微博都认为是用户喜欢的微博.

4.2 实验结果与相关分析

为了验证本文方法的有效性及推荐结果的准确性,本节设计了 4 个实验,对本文提出的方法进行验证并对实验结果进行分析.一是通过改变参数α和阈值μ,比较微博推荐算法的性能,从而确定最优参数值;二是在参数值确定的基础上,验证标签扩充个数对微博推荐性能的影响;三是通过比较标签扩充前后的内容,展示标签扩充方法的性能;四是本文的微博推荐算法与其他算法的比较.

4.2.1 参数设置对方法性能的影响

下面将通过实验来考察方法中涉及到的参数对算法性能的影响,它们分别是参数α和阈值μ.当测试其中一个参数值对算法的影响时,另外一个参数值保持不变.

对超边加权时,α用于调节微博时间因子和微博人气指数的比重,其值越高,意味着用户发布微博的时间对于用户兴趣的提取影响越大;其值越小,就意味着微博的评论数、转发数等对于用户兴趣提取影响提高.为了计算参数α对于推荐结果的影响,本文分别对不同α取值下算法在微博推荐个数为 L=5、L=10、L=15 及 L=20 的推荐结果进行对比.设参数μ=0.5,分别在α取不同值时,比较方法的性能,图 2 展示了实验结果.对比图 2 左、右两图,可以发现以下几点.

(1) 左图中,当 L=15 时算法的准确率 $P@L$ 达到最佳;在右图中,当 L=20 时算法的召回率 $R@L$ 达到最佳.这是由于召回率依赖于用户测试样本数目,随着推荐数目的增加,算法召回率也逐渐增加.

(2) 当α=0.7 时,算法的准确率 $P@L$ 和召回率 $R@L$ 均达到最大值,算法的性能在各个微博推荐个数上都达到最佳状态.值得注意的是,当α=1 时,算法的推荐性能在准确率和召回率上都优于α=0 时,因此,微博时间因子对用户兴趣提取准确性的影响大于微博人气指数.在接下来的实验中,设定α=0.7.

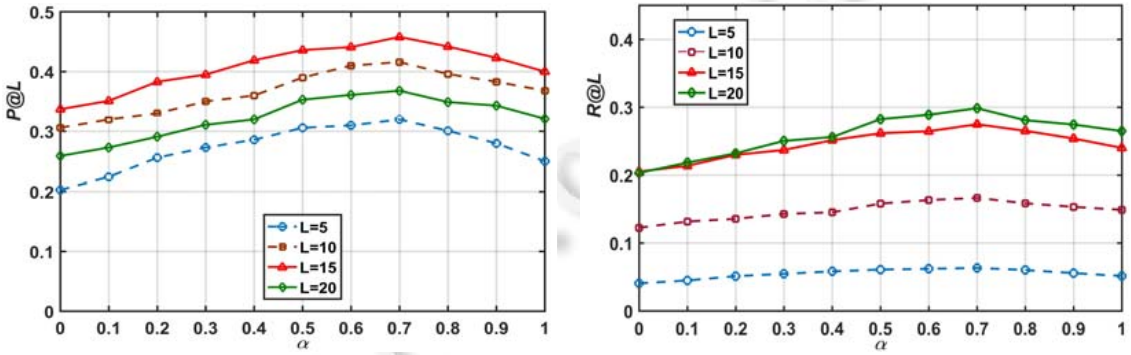


Fig.2 Impact of parameter α on recommendation algorithms

图 2 参数α对推荐算法的影响

阈值μ的大小决定了推荐方法向用户推荐微博数量的大小,阈值μ越小,则向用户推荐的微博数量越多;阈值μ越大,则向用户推荐的微博数量越少.为了清楚地了解阈值μ的取值对推荐算法的影响,令参数α=0.7,μ取不同的值,在测试数据集上计算方法取得的实验结果如图 3 所示.

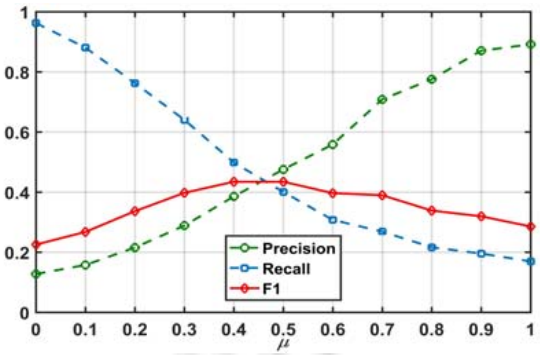


Fig.3 Impact of threshold μ on recommendation algorithms

图 3 阈值μ对推荐算法的影响

从图中可以看出,随着阈值μ的增大,算法的准确率逐渐上升,而算法的召回率却逐渐减小.这是因为随着阈值μ的增大,方法要求所推荐微博与用户的相似度也在不断增大,因而推荐给用户的微博越来越少.当μ=0.4 或 0.5 时,算法在 $F1$ 这一评价指标上都达到了最佳.因此,本文又在综合考虑μ=0.4 或 0.5 时算法的准确率和召回率这两个评价指标后,确定μ=0.45.

4.2.2 不同标签扩充个数对推荐算法的影响

为了验证标签扩充个数 Top-Q 对微博推荐方法的影响,分别选取{1,3,5,7,9,10}个关键词对用户标签进行扩充,计算在不同标签扩充个数的情况下本文算法的准确率,进而确定标签扩充个数 Top-Q 的值,如图 4 所示.从图中可以看出,随着标签扩充个数的增加,算法的准确率 $P@L$ 也逐渐增加.当标签扩充个数 P>9 时,算法的准确率 $P@L$ 不增反降.这是由于随着标签扩充个数的增大,一些排名靠后的关键词也被扩充到用户标签集合中,这部分标签并不能很好地表征用户的兴趣爱好.从图中可以看出,当 Q=7 或 9 时,算法的准确率 $P@L$ 并无明显增加,算法的性能在各个微博推荐个数上都达到最佳状态.因此,取 7 和 9 的均值 8 作为标签扩充个数 Q 的值.

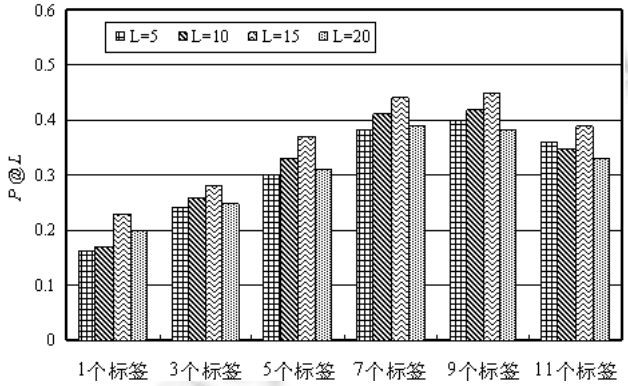


Fig.4 Impact of the number of tag extensions on recommendation algorithms

图 4 标签扩充个数对推荐算法的影响

4.2.3 标签扩充前后用户标签详情对比

限于篇幅,本部分只展示了 10 位用户自己添加的标签以及经过超图随机游走算法为用户添加的标签的情况,见表 3.

Table 3 Comparison of user tag before and after the tag expansion

表 3 标签扩充前后用户标签详情对比

微博用户	扩充前用户自己添加的标签	扩充后由推荐算法给用户添加的标签(Q=8)
User-1	淘宝控;摄影;睡觉;宅;旅游	马云;优惠券;音乐;90 后;美食;单反;时尚;大学生
User-2	美食;国际;娱乐;人生;时尚	成都;舌尖上的中国;旅游;电影;听歌;看书;自由;华人
User-3	艺术;收藏;历史	故宫;拍卖会;时尚;新闻;媒体;电影;财经;养生
User-4	信息化;科技;社会	计算机;手机;互联网+;创新;财经;创业;电子商务
User-5	文学;摄影;新媒体;旅行;电影;新闻资讯;普利策	音乐;时尚;奋斗;美食;诺贝尔;媒体人;科技;创新
User-6	设计;建筑	旅游;电影;音乐;美食;文学;摄影;历史;财经
User-7	美食;电影;音乐;90 后;奋斗	综艺;旅游;电子控;游泳;宅;大学生;优惠券;红包
User-8	历史;财经;健康;养生;保健	心灵鸡汤;平常心;情感;摄影;文学;新闻;教育;防诈骗
User-9	旅游;电影;音乐;时尚;自由;游泳;摄影;美食;90 后;旅游	大学生;舌尖上的中国;张艺谋;新媒体;韩剧;创业;奋斗;科技
User-10	互联网;创新;电子控	电子商务;财经;创业;互联网+;科技;奋斗;青年;新闻

可以看到,少数的高频词出现在相当多的微博用户标签中,这些热门标签的内容多是大众性的兴趣爱好的描述,如“音乐”“电影”“美食”等;或者是对一些常见人群的描述,如“大学生”“90 后”“宅”.这些标签之所以被频繁使用,一是因为这其中的一些标签在用户添加标签的页面作为系统推荐选项出现,因此有更大的概率被用户看到和选中,而不用手动输入;二是此类标签对于新浪微博用户具有普适性,即很多微博用户都会发现这样的标签在某种程度上符合对自己的描述.例如在实验数据集中,有 52.7%的用户是大学生, “奋斗”“90 后”两个标签非常符合对这些用户的描述,因此成为高频标签.

4.2.4 不同微博算法的性能比较

为了验证该推荐算法的有效性,比较本文提出的 LeALpc 算法与基于标签关联关系推荐算法(label

correlation,简称 LC)[13]、基于标签概率相关性推荐算法(label probability correlation,简称 LPC)[14]、融合标签关系与用户关系推荐算法(label correlation and user social relation,简称 ILCAUSR)[15]、协同个性化微博推荐(collaborative personalized tweet recommendation,简称 CTR)[3]、基于用户嵌入的学术微博推荐(user embedding for scholarly microblog recommendation,简称 UEMR)[4]和基于背景和内容的微博推荐(microblog recommendation based on profile and content,简称 BPACMR)[22]的预测效果.选择以上 6 种算法作为对比算法是基于以下几点考虑.

(1) 本文算法是在 LPC 算法的基础上改进而来,LPC 算法与本文的算法最相似.

(2) LC 算法、ILCAUSR 算法、LPC 算法以及本文的算法都是基于标签进行微博推荐的.

(3) ILCAUSR 算法已被证明在微博推荐算法上优于其他算法.

(4) 由于前面 3 种比较算法都是从标签角度出发的微博推荐算法,为了更好地验证本文方法的有效性,采用从其他角度出发且具有较好性能的微博推荐算法(CTR 算法、UEMR 算法和 BPACMR 算法)进行对比.

利用不同微博推荐列表长度 L=5、L=10、L=15 及 L=20 对以上算法进行实验,比较在不同推荐列表长度的情形下,几种推荐算法的准确率 $P@L$、$F1$ 值 $F1@L$ 以及平均正确率 $AP@L$,结果见表 4.

Table 4 Comparison of different recommendation algorithms

表 4 不同推荐算法比较

名称	L=5		L=10		L=15		L=20	
	P	*AP*	*P*	*AP*	*P*	*AP*	*P*	*AP*
LC	0.281	0.42	0.316	0.386	0.334	0.524	0.231	0.442
LPC	0.283	0.4	0.359	0.452	0.379	0.465	0.226	0.469
ILCAUSR	**0.315**	0.52	0.427	0.5	0.436	0.58	**0.255**	0.542
CTR	0.295	0.448	0.396	0.524	0.418	0.563	0.247	0.52
UEMR	0.299	0.432	0.403	0.469	0.413	0.549	0.25	0.488
BPACMR	0.293	0.54	0.388	0.488	0.408	0.517	0.246	0.506
LeALpc	0.308	**0.559**	**0.428**	**0.586**	**0.439**	**0.643**	0.253	**0.587**

从表中可以看出,本文提出的 LeALpc 算法与从内容角度出发的 CTR 算法、UEMR 算法和 BPACMR 算法以及从标签角度出发的 LPC 算法、算法 LC 和 ILCAUSR 算法在平均准确率方面相比都更优异.这是由于这些算法过分关注用户的整体兴趣而忽视了用户的个性化兴趣,导致推荐列表前几位的命中率低.而 LeALpc 算法结合了微博文本和用户标签这两个体现用户兴趣的重要方面,它更能展现用户的个性化兴趣.在实际应用中,推荐正确的次序尤其重要,因为用户不可能耐心浏览完所有推荐的微博.在其他评价指标上,LeALpc 算法明显高于除 ILCAUSR 算法之外的 5 种算法,但是与 ILCAUSR 算法相比并没有明显优势.这是由于 ILCAUSR 算法将用户间社交关系融入到微博推荐算法中,较为准确地表示出了用户的兴趣.而本文尚未考虑,这也将是本文今后继续研究的方向.

接着,为了进一步比较 LeALpc 算法和其他 6 种算法的推荐性能,从表 4 中选取推荐性能(正确率 $P@L$)最好情况(L=15)和最坏情况的(L=20)的正确率和平均正确率展开分析,分别是 L=15 和 L=20 时,7 种算法在不同评价指标上 10 次交叉验证所得结果的分布情况,如图 5 所示.箱线图是一种数据样本统计图,它可以看出数据是否具有对称性以及数据分布的分散程度等信息.因此,从图 5 可以看出,

- 在评价指标 $AP@L$ 上,无论是最好情况(L=15)还是最坏情况(L=20),本文提出的 LeALpc 算法不但具有较高的平均值,而且 10 次所得结果也比较稳定.
- 在评价指标 $P@L$ 上,在最好情况(L=15)下,本文提出的算法与 ILCAUSR 算法相比,在平均值上虽然并没有明显优势,但是在结果分布上要优于其他算法.

这更加验证了根据微博用户以往发布的微博内容对其标签进行扩充以及根据标签概率相关性对用户进行微博推荐这一方法的有效性.

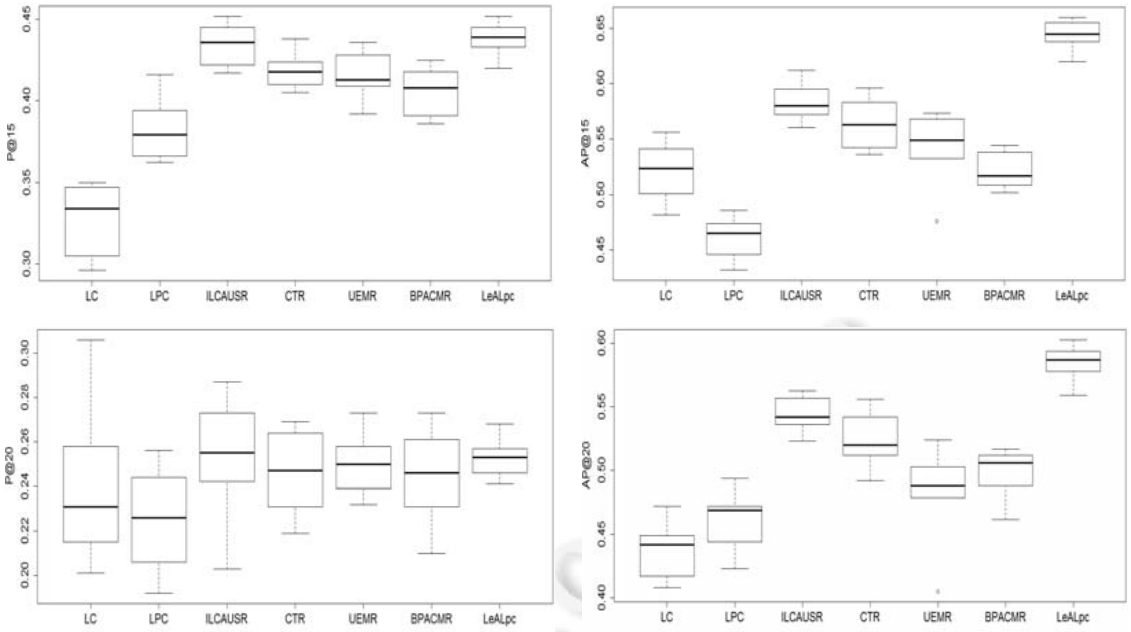


Fig.5 Performance comparison among LeALpc algorithm and other baselines on L=15 and L=20

图 5 LeALpc 算法与对比算法在 L=15 和 L=20 时的性能比较

5 结论与展望

随着移动互联网的快速发展以及社交网络规模的不断增大,个性化的信息推荐越来越受到信息接收者的青睐.为了提升用户浏览信息的体验度,面对海量复杂的微博消息,实现内容精准推荐.本文从微博用户标签入手,针对绝大多数微博用户没有给自己加注标签或标签较少的问题,提出一种结合标签扩充与标签概率相关性的微博推荐方法.首先,该方法将微博视为超边,微博中的词视为超点来构建超图,并以一定的加权策略对超边和超点进行加权,通过在超图上随机游走得到一定数量的关键词对微博用户标签进行扩充;然后,采用相关性标签权重加权方案,构建用户-标签矩阵,利用标签间的概率相关性,构造标签相似性矩阵,对用户-标签矩阵进行更新,更新后的用户标签矩阵不仅稀疏性得到了有效缓解,而且还包含了丰富的标签关联关系;最后,依据构建的兴趣模型对用户进行信息推荐.在未来的工作中,将进一步对用户与用户之间的社交属性进行研究,提升用户模型的准确度,实现更加精准的推荐.

References:

[1] Chen Y, Cheng XQ, Yang S. Finding high quality threads in Web forums. Ruan Jian Xue Bao/Journal of Software, 2011,22(8): 1785−1804 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3857.htm [doi: 10.3724/SP.J.1001.2011.03857]

[2] Zhang D. Research on microblog recommendation method based on user social behavior [Ph.D. Thesis]. Lanzhou: Northwest Normal University, 2018.

[3] Chen KL, Chen TQ, Zheng GQ, Jin O, Yao EP, Yu Y. Collaborative personalized tweet recommendation. In: Proc. of the 35th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Portland: ACM Press, 2012. 661−670.

[4] Yang Y, Wan XJ, Zhou XJ. User embedding for scholarly microblog recommendation. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016. 449−453.

[5] Gao M, Jin CQ, Qian WN, Wang XL, Zhou AY. Real-time and personalized recommendation on microblogging systems. Chinese Journal of Computers, 2014,37(4):963−975 (in Chinese with English abstract).

[6] Sun A. Short text classification using very few words. In: Proc. of the 35th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Portland: ACM Press, 2012. 1145−1146.

[7] Meng XW, Liu SD, Zhang YJ, Hu X. Research on social recommender systems. Ruan Jian Xue Bao/Journal of Software, 2015, 26(6):1356−1372 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4831.htm [doi: 10.13328/j.cnki.jos.004831]

[8] Ramage D, Dumais ST, Liebling DJ. Characterizing microblogs with topic models. In: Proc. of the Int'l AAAI Conf. on Weblogs and Social Media. Washington: AAAI Press, 2010. 130−137.

[9] Liu WY, Quan XJ, Feng M, Qiu B. A short text modeling method combining semantic and statistical information. Information Sciences, 2010,180(20):4031−4041.

[10] Weng JS, Lim EP, Jiang J, He Q. TwitterRank: Finding topic-sensitive influential twitterers. In: Proc. of the 3rd ACM Int'l Conf. on Web Search and Data Mining. New York: ACM Press, 2010. 261−270.

[11] Zhang B, Zhang Y, Gao KN, Guo PW, Sun DM. Combining relation and content analysis for social tagging recommendation. Ruan Jian Xue Bao/Journal of Software, 2012,23(3):476−488 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4001.htm [doi: 10.3724/SP.J.1001.2012.04001]

[12] Xing QL, Liu L, Liu YQ, Zhang M, Ma SP. Study on user tags in Weibo. Ruan Jian Xue Bao/Journal of Software, 2015,26(7): 1626−1637 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4655.htm [doi: 10.13328/j.cnki.jos.004655]

[13] Ma HF, Jia MHZ, Xie M, Lin XH. A microblog recommendation algorithm based on multi-tag correlation. In: Proc. of the 8th Int'l Conf. on Knowledge Science, Engineering and Management. Chongqing: Springer-Verlag, 2015. 483−488.

[14] Zhang D, Ma HF, Jia JJ, Yu L. A microblog recommendation method based on label probability correlation. Computer Engineering and Science, 2017,39(9):1742−1748 (in Chinese with English abstract).

[15] Ma HF, Jia MHZ, Zhang D, Lin XH. Combining tag correlation and user social relation for microblog recommendation. Information Sciences, 2017,385:325−337.

[16] Hua W, Wang ZY, Wang HX, Zheng K, Zhou XF. Short text understanding through lexical-semantic analysis. In: Proc. of the 31st Int'l Conf. on Data Engineering. Seoul: IEEE Press, 2015. 495−506.

[17] Bellaachia A, Al-Dhelaan M. HG-rank: A hypergraph-based keyphrase extraction for short documents in dynamic genre. In: Proc. of the 4th Workshop on Making Sense of Microposts. Seoul: CEUR Workshop, 2014. 42−49.

[18] Liu Q, Li ZG, Lui JCS, Cheng JF. PowerWalk: Scalable personalized pagerank via random walks with vertex-centric decomposition. In: Proc. of the 25th ACM Int'l Conf. on Information and Knowledge Management. Indianapolis: ACM Press, 2016. 195−204.

[19] Tu NN, Kanhabua N, Zhu X. A time-aware random walk model for finding important documents in web archives. In: Proc. of the 38th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Santiago: ACM Press, 2015. 915−918.

[20] Song SX, Zhu H, Chen L. Probabilistic correlation-based similarity measure on text records. Information Sciences, 2014,289: 8−24.

[21] Zhou XK, Wu S, Chen C, Chen G, Ying SS. Real-time recommendation for microblogs. Information Sciences, 2014,279:301−325.

[22] Zhong ZM, Guan Y, Hu Y, Li CH. Mining user interests on microblog based on profile and content. Ruan Jian Xue Bao/Journal of Software, 2017,28(2):278−291 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5030.htm [doi: 10.13328/j.cnki.jos.005030]

附中文参考文献:

[1] 陈友,程学旗,杨森.面向网络论坛的高质量主题发现.软件学报,2011,22(8):1785−1804. http://www.jos.org.cn/1000-9825/3857.htm [doi: 10.3724/SP.J.1001.2011.03857]

[2] 张迪.基于用户社交行为的微博推荐方法研究[硕士学位论文].兰州:西北师范大学,2018.

[5] 高明,金澈清,钱卫宁,王晓玲,周傲英.面向微博系统的实时个性化推荐.计算机学报,2014,37(4):963−975.

[7] 孟祥武,刘树栋,张玉洁,胡勋.社会化推荐系统研究.软件学报,2015,26(6):1356−1372. http://www.jos.org.cn/1000-9825/4831.htm [doi: 10.13328/j.cnki.jos.004831]

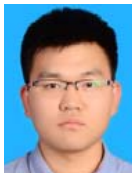
[11] 张斌,张引,高克宁,郭朋伟,孙达明.融合关系与内容分析的社会标签推荐.软件学报,2012,23(3):476−488. http://www.jos.org.cn/1000-9825/4001.htm [doi: 10.3724/SP.J.1001.2012.04001]

[12] 邢千里,刘列,刘奕群,张敏,马少平.微博中用户标签的研究.软件学报,2015,26(7):1626−1637. http://www.jos.org.cn/1000-9825/4655.htm [doi: 10.13328/j.cnki.jos.004655]

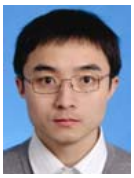
[14] 张迪,马慧芳,贾俊杰,余丽.一种基于标签概率相关性的微博推荐方法.计算机工程与科学,2017,39(9):1742−1748.

[22] 仲兆满,管燕,胡云,李存华.基于背景和内容的微博用户兴趣挖掘.软件学报,2017,28(2):278–291. http://www.jos.org.cn/1000-9825/5030.htm [doi: 10.13328/j.cnki.jos.005030]

马慧芳(1981－),女,甘肃兰州人,博士,教授,CCF 专业会员,主要研究领域为机器学习,数据挖掘.



张迪(1992－),男,硕士,主要研究领域为互联网数据挖掘.



赵卫中(1981－),男,博士,副教授,CCF 专业会员,主要研究领域为机器学习,数据挖掘,算法分析与设计.

史忠植(1941－),男,研究员,博士生导师,CCF 会士,主要研究领域为人工智能,机器学习,神经计算,认知科学.