基于线性学习模型的社会媒体流排名算法

张 威 李跃新
(湖北大学计算机与信息工程学院 武汉 430064)

摘 要 在社会媒体中,对用户推荐适合的状态更新不仅降低了用户搜索信息的时间,也可以增加用户对服务的粘性。针对社会媒体中状态更新而推荐的准确性低的不足,提出了一种基于线性学习模型的状态更新排名算法。首先,根据社会媒体的性质定义了相应的偏好属性,并提出了一种基于线性模型的潜在偏好模型;其次,根据状态更新以及接收者的特征定义了相应的线性特征模型;最后,将潜在偏好模型和特征模型相结合,提出了一种引入时间效应的线性模型。通过实验验证表明,提出的算法与其它相关算法相比,算法的预测准确性更高,执行效率更快。

关键词 社会媒体流,排名算法,排名学习,线性模型
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.12.058

Learning to Rank Based on Linear Model for Social Media Streams

ZHANG Wei LI Yue-xin
(Faculty of Computer Science and Information Engineering, Hubei University, Wuhan 430064, China)

Abstract In social media, recommending suitable updates for users can not only reduce information searching time, but also improve users' stickiness for social media. In order to improve the recommendation accuracy of updates in social media, this paper proposed a linear model based learning to ranking algorithm for updates. Firstly, according to attributes of social media, we defined corresponding bias features, and proposed a linear model based latent bias model. Secondly, according to features of update and recipients, we defined corresponding linear feature model. Finally, combining the latent bias model and the feature model, we proposed a linear model with temporal effect. The experiments show that, compared with related works, the proposed algorithm has better prediction accuracy and higher execution efficiency.

Keywords Social media streaming, Ranking, Learn to rank, Linear model

1 引言

在Web环境下,社会媒体已经渗入到了人们的日常生活中。随着Facebook、Twitter、LinkedIn以及新浪微博等社会媒体的普及,数以亿计的用户通过文本、图片、音频和视频等媒体信息进行连接并进行相互间的沟通。在社会媒体中,登录到系统的任何用户可以在任何时间发布任何媒体信息,每一条信息是一个状态更新[1]。对于整个系统而言,所有用户发布的全部状态更新可以按照系统的时间轴形成一个数据流,并且这个数据流源源不断地涌入系统。当用户登录到系统时,可以观察到其关注的用户发布的实时状态更新。

在社会媒体中,状态更新所特有的属性是用户可以获取实时的信息,如好友的位置以及新闻等。然而社会媒体中用户规模常常以亿计,当用户关注的用户数量很大时,用户往往会被大量的状态更新所湮没,这将严重影响用户对社会媒体的使用体验和粘性[2]。首先,当用户面对大量来自好友的状态更新时,用户不能高效地获取重要的信息,这种现象被称为信息过载。此外,用户所能获取到的状态更新仅仅局限于他所关注的用户,而那些没有关注的用户发布的重要信息却不能有效地识别。为了获取那些非好友用户发布的重要信息,用户不得不花费大量的时间和精力进行信息的搜索与鉴别,这种现象被称为信息匮乏。为了应对社会媒体所面临的信息过载和信息匮乏,社会媒体监测系统得到了学术界和工业界的广泛关注[3]。社会媒体监测系统[4]通过对海量的状态更新进行监测和过滤,并根据设定的指标向用户推荐状态更新。

对社会媒体中的状态更新进行过滤并对用户进行个性化推荐与传统的信息检索和推荐系统相似,却又有所区别。从信息检索[5]的角度来看,状态更新的过滤与推荐是根据用户的兴趣将状态更新进行降序排名,并过滤出排名靠前的若干项推荐给用户,采用的方法包括个性化PageRank[6]、HITS[7]、ObjectRank[8]、SimRank[9]等。在信息检索中,用户需要指定若干个查询关键字,系统根据关键字进行搜索并将搜索到的结果进行排名。然而在状态更新推荐中,用户不需要输入具体的查询内容,查询内容根据一定的规则进行隐式地推导。此外,社会媒体中的用户需求具有多样性特征,即展现给用户的结果应该属于不同的范围。为了解决信息检索中的多样性问题,研究人员提出了基于排名学习的检索方法。

从推荐系统[10]的角度来看,状态更新的过滤与推荐是将

到稿日期:2014-10-29 返修日期:2014-12-09 本文受国家自然科学基金项目(61170306),湖北省科技支撑项目(2014BAA089)资助。
张 威(1978-),男,硕士,讲师,主要研究方向为传感器技术以及计算机应用技术;李跃新(1958-),男,博士,教授,主要研究方向为人工智能与知识工程、智能控制系统、嵌入式技术(通信作者)。

状态更新作为推荐系统中的项推荐给用户。推荐系统中常用的推荐方法包括基于内容的推荐[11]、基于协同过滤的推荐[12]、基于知识的推荐[13]，以及基于模型的推荐方法[14]等。状态更新推荐与传统的推荐系统的区别在于：社会媒体是动态的，系统中每时每刻都可能出现大量的状态更新。这些最新出现的状态更新相当于推荐系统中的新项，这将导致状态更新推荐面临大量的冷启动问题，因此传统的推荐算法并不能被直接采用。

本文基于线性学习模型研究了社会媒体中的状态更新排名问题。将状态更新的发送者和接收者的特征描述成特征模型，将接收者的偏好描述成潜在偏好模型，在对上述两个模型进行合并后又引入了时间效应，提出了一种基于线性学习模型的社会媒体流排名方法。

2 线性学习排序模型

在社会媒体中，当系统中出现一个状态更新后，用户可以选择是否对该状态更新进行回应。为了便于讨论，本文假设所有的用户回应都是对该状态更新的点击事件，没有点击表示不回应，因此状态更新的用户响应为二值的回应。

在社会媒体中，当用户关注的好友数量过多时，其好友的状态更新可能过多而无法显示在用户的首页内。本文在后续的讨论中不考虑单个用户的首页，而考虑所有用户首页的并集，用 y 表示所有用户的首页内对所有状态更新的响应向量。由于只考虑对状态更新的响应，因此 y 中元素的顺序并不重要。用 y_i 表示整个数据集中对第 i 个状态更新的响应，f_i 表示根据模型对 y_i 进行估计的估计值，R 表示接收者集合，S 表示发送者集合，T 表示类型的集合，I 表示状态更新的集合。此外，本文定义如下辅助函数：$r(i)$ 表示状态更新 i 的接收者；$s(i)$ 表示状态更新 i 的发送者；$t(i)$ 表示状态更新 i 的类型；$c(i)$ 表示发送者 i 的类型。

2.1 潜在偏好模型

在 y 中，用户对状态更新的响应包含用户的偏好行为，本文通过线性模型描述这种偏好。

假设 μ 为所有用户对所有状态更新的平均响应率，那么一个新的状态更新 i 被用户点击的平均概率如式(1)所示。式(1)表示所有用户的平均值，并没有考虑每个用户的不同偏好。

$$f_i=\mu \tag{1}$$

本文对式(1)进行扩展，考虑了如下偏好属性。

- b_i：状态更新 i 可以用向量表示，不同的状态更新可以用不同的向量来表示；
- $b_{t(i)}$：用来区别不同状态更新的类型，不同类型的状态更新可能有着不同的重要性。例如，好友工作职位改变的状态更新的重要性相对较高。
- $b_{r(i)}$：表示接收者的偏好。
- $b_{c(i)}$：用来区别不同发送者的类型。例如，企业对员工发送的通知不同于好友的留言信息。
- $b_{s(i)}$：表示发送者的偏好。

用 b 来表示偏好，b 的下标表示不同类型的偏好，那么包含偏好的用户对状态更新 i 的响应如式(2)所示。由于这些偏好是未知的，可以将它们视为数据集中包含的潜在变量。

$$f_i^{(1)}=\mu+b_i+b_{r(i)}+b_{s(i)}+b_{t(i)}+b_{c(i)} \tag{2}$$

2.2 特征模型

对于每一个状态更新，本文用线性模型来描述并识别其相应的特征。给定状态更新 i，用 ϕ_i 表示 i 的特征向量，用 $\phi_{r(i)}$ 表示 i 的接收者的特征向量。对 ϕ_i 和 $\phi_{r(i)}$ 进行线性组合，可以得到用户对 i 的相应预测值，即

$$f_i^{(2)}=\beta_{r(i)}^{\mathrm{T}}\phi_{r(i)}+\alpha_{r(i)}^{\mathrm{T}}\phi_i \tag{3}$$

其中，参数 $\beta_{r(i)}^{\mathrm{T}}$ 和 $\alpha_{r(i)}^{\mathrm{T}}$ 需要根据数据集中的数据进行学习得到，并且不同的用户 u 具有不同的 β 和 α 取值。式(3)对用户的特征和状态更新的特征进行了线性组合，并且不同的用户具有不同的参数取值。

在潜在偏好模型和特征模型的基础上，对二者进行结合，得到的结果仍为线性模型，结果如式(4)所示。在该模型中，不同的变量取值为不同的用户行为进行了不同的解释。

$$f_i^{(3)}=f_i^{(1)}+f_i^{(2)} \tag{4}$$

2.3 引入时间效应的线性模型

在社会媒体中，状态更新在本质上是时序敏感的，用户往往忽略那些过时的信息而关注那些最新的信息。本文通过简单的时序特征 $t_{recency}=t_{imp}-t_{upt}$ 来描述状态更新的时序特征，其中 t_{upt} 表示状态更新的发布时间，t_{imp} 表示用户在首页观察到该状态更新的时间。对于不同的状态更新，其时序特征是不同的。对于相同的状态更新，由于不同用户观察到该状态更新的时间 t_{imp} 是不同的，因此时序特征也是不同的。在引入了时序特征后，用户对状态更新 i 的相应预测值如式(5)所示：

$$f_i^{(4)}=f_i^{(*)}+\zeta\times t_{recency} \tag{5}$$

其中，$f_i^{(*)}$ 为未引入时序特征的预测模型，可以为 $f_i^{(1)}$，$f_i^{(2)}$ 或者 $f_i^{(3)}$，参数 ζ 表示时序特征的重要性。ζ 作为预测值的个性化参数可以从训练数据集中学习得到。本文对 ζ 的取值进行了限制，通过如下两个因素手动调整 ζ 的取值。

1)由于并不是所有的用户都频繁地与社会媒体流进行交互，并且新用户不断加入到系统中，需要为这些用户合理地推荐状态更新。在这种情况下，显示的特征可能不存在，并且从训练集合中学习到的偏好也可能不可靠。

2)现有的社会媒体主要基于时序进行排名，并且大多数用户熟悉这种排名模式。在这种情况下，我们不希望改变用户的习惯，因此需要对其进行调整。

由于用户对状态更新的响应是二值(点击与否)的，本文在正确值 y_i 和预测值 f_i 之间引入对数误差来生成逻辑回归的学习过程。对于给定的状态更新 i，通过最小化如下目标函数对线性模型的参数进行学习。

$$l_1(y_i,f_i^{(*)})=\log[1+\exp(-y_i f_i^{(*)})] \tag{6}$$

其中，$y_i\in\{+1,-1\}$ 表示用户对 i 的真实响应，$f_i^{(*)}$ 为采用上述公式得到的预测值。

实际上，为了避免训练数据集的过拟合现象，采用 L_2 范式对上述参数进行规范化。于是，式(5)的目标函数可表示为：

$$\begin{aligned}L_1=&\sum_i l_1(y_i,f_i^{(4)})+\lambda_1\Big(\sum_i\|b_i\|^2+\sum_{t(i)}\|b_{t(i)}\|^2+\sum_{c(i)}\|b_{c(i)}\|^2+\sum_{r(i)}\|b_{r(i)}\|^2+\sum_{s(i)}\|b_{s(i)}\|^2\Big)+\lambda_2\Big(\sum_u\|\beta_u\|_F^2+\sum_u\|\alpha_u\|_F^2\Big)\end{aligned} \tag{7}$$

其中，λ_1 和 λ_2 为需要手动调整的规范化参数。本文采用随机梯度下降法对式(7)进行求解，其梯度计算公式如下。

$$\frac{\partial L_1}{\partial b_*}=-\sum_i[1-\sigma(y_i f_i^4)]y_i+2\lambda_1\sum_* b_* \tag{8}$$

$$\frac{\partial L_1}{\partial \beta_{r(i),k}}=-\sum_i[1-\sigma(y_i f_i^4)]y_i\phi_{r(i),k}+2\lambda_2\beta_{r(i),k} \tag{9}$$

$$\frac{\partial L_1}{\partial \alpha_{i,k}}=-\sum_i[1-\sigma(y_i f_i^4)]y_i\phi_{i,k}+2\lambda_2\alpha_{i,k} \tag{10}$$

其中，b_* 表示任何偏好项，$\alpha_{i,k}$ 表示状态更新 i 的系数的第 k 个元素，$\beta_{r(i),k}$ 表示 i 的接收者 $r(i)$ 的系数的第 k 个元素，并且 $\sigma(x)=\frac{1}{1+\exp(-x)}$。

3 实验设计与结果分析

3.1 实验环境设置

基于排名的评价指标是信息检索领域采用的常用指标，本文采用基于排名的评价指标来衡量状态更新推荐的有效性。借鉴信息检索中的 MAP(Mean Average Precision)，定义社会媒体推荐中的 $Precision@k=\frac{\#\ \text{of clicks in top positions}}{k}$，其中 k 为排名列表中的第 k 个位置，# of clicks in top positions 为所有用户对第 k 个位置的状态更新的点击次数。给定用户 u，前 m 个位置的平均准确率为 $AP_{u,m,k}=\frac{\sum_{k=1}^{m}Precision@k\times l_k}{\#\ of\ clicks\ for\ ranked\ list\ of\ u}$，其中 l_k 表示位置 k 的状态更新是否被用户点击，m 为被评估的状态更新个数。于是，所有用户对状态更新 k 的平均准确率为

$$MAP_{m,k}=\frac{1}{|U|}\sum_{u\in U}AP_{u,m,k} \tag{11}$$

其中，U 为用户集合，$|\cdot|$ 为集合的势。

实验采用的平台为个人 PC，采用的数据集为 Tencent 数据集(KDDCup2012)。由于该数据集的规模比较大，截取了数据集的前 10%进行算法性能的验证。

3.2 实验结果

在截取的数据集中，随机选取 70%的数据作为训练数据，并用余下的 30%数据进行算法的准确性判断。在实验中，将本文提出的基于线性学习模型的排名算法记为 LM，将其与基于 kNN 的协同过滤算法[15]、学习排名算法 L2R[16] 和 PageRank 算法进行对比。

首先，实验验证了本文所提算法的准确性。在该实验中随机选取了 20 个用户，并对每个用户推荐 20 个状态更新，重复 10 次并取其平均值，实验结果如图 1 所示。从图 1 可以看出，在 4 种算法的对比中，LM 算法的 MAP 最高，L2R 算法次之，kNN 算法最差。其中 LM 和 L2R 算法都是基于学习的排名算法，这两种算法能根据用户及状态更新的变化进行个性化推荐，因此其预测的准确性要高于其它两种算法。

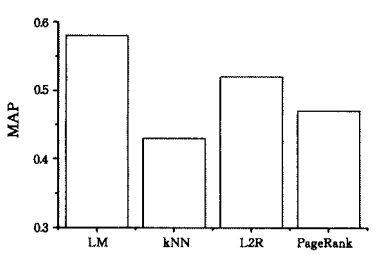

图 1 算法的准确性对比

其次，对 4 种算法的执行效率进行对比，采用的评价标准为算法的执行时间。kNN 算法是一种在线计算方法，当给定用户及状态更新时，在线计算其预测值。其它 3 种方法均为离线计算方法，PageRank 算法根据指定的用户进行个性化排名，LM 和 L2R 算法计算每个用户的预测模型。本实验对比了 3 种离线算法的执行时间，在每种算法中，随机选取 20 个用户，并对状态更新进行个性化排名，每个算法重复 10 次并取其平均值，实验结果如图 2 所示。从图 2 可以看出，LM 算法在进行预测时所需的时间最少，L2R 次之，PageRank 所需的计算时间最长。

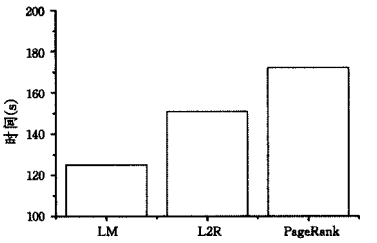

图 2 算法执行时间对比

最后，对 LM 算法中参数 ζ 的取值进行评估。ζ 的取值为从 0.01 到 1000，每次扩大 10 倍，共取值 6 次，观察不同 ζ 下 LM 算法的预测准确性 MAP。从图 3 可以看出，随着 ζ 取值的不断增大，LM 算法的 MAP 先迅速增大，当增大到一定值后缓慢降低。这表明在进行用户的个性化排名时，ζ 的取值很重要，通常在一定的范围内(10 左右)。

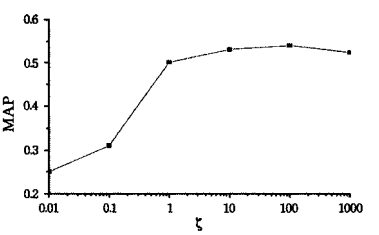

图 3 LM 算法中参数 ζ 的评估

结束语 社会媒体已经成为人们日常获取信息的主要来源。面对海量的社会媒体信息，用户同时面临着信息过载和信息匮乏问题。为了提高用户获取信息的质量和效率，本文基于线性学习模型，提出了一种社会媒体中的状态更新排名模型。该模型结合了潜在变量模型和特征模型，并引入了状态更新的时间效应。实验表明，本文提出的算法与其它相关算法相比，预测准确性更高，执行效率更快。

参 考 文 献

[1] Kaplan A M, Haenlein M. Users of the world, unite! The challenges and opportunities of Social Media[J]. Business horizons, 2010, 53(1): 59-68

[2] 徐恪，张赛，陈昊，等. 在线社会网络的测量与分析[J]. 计算机学报，2014，37(1)：165-188
Xu Ke, Zhang Sai, Chen Hao, et al. Measurement and Analysis of Online Social Networks[J]. Chinese Journal of Computers, 2014, 37(1): 165-188

[3] 陈晓江，房鼎益，刘炜，等. 基于 CORBA 的媒体流构件模型[J]. 西北大学学报(自然科学版)，2005，35(2)：151-154

(下转第 291 页)

安. 2012
Geng Guo-hua. The research and Application of technology for Cultural heritage digitalization and virtual restoration [R]. Xi'an. 2012
[2] 朱新懿,耿国华. 一种结合局部对称的三维模型对齐方法[J]. 计算机科学,2015,42(2):277-279
Zhu Xin-yi, Geng Guo-hua. 3D Model's Alignment Approach Combining Partial Symmetry [J]. Computer Science, 2015, 42(2):277-279
[3] Yemez Y, Schmitt F. 3D reconstruction of real objects with high resolution shape and texture[J]. Image and Vision Computing, 2004(22):1137-1153
[4] 刘钢,彭群生,鲍虎军. 基于多幅实拍照片为真实景物模型添加纹理[J]. 软件学报,2005,16(11):2014-2019
Liu Gang, Peng Qun-sheng, Bao Hu-jun. Texture Mapping on Real World Models from Multiple Photographic Images [J]. Journal of Software, 2005, 16(11):2014-2019
[5] 崔桂涣,张之江,董志华. 自由多视角恢复表面纹理的三维重建[J]. 微计算机应用,2009,30(1):1-5
Cui Gui-huan, Zhang Zhi-jiang, Dong Zhi-hua. The 3D Reconstruction of Recovering Surface Texture from Free Multiple Views [J]. Microcomputer Application, 2009, 30(1):1-5
[6] Gomes L, et al. 3D reconstruction methods for digital preservation of cultural heritage[J]. Pattern Recognition Lett, 2014, 50:3-14
[7] Zha H, Wang P. Realistic face modeling by registration of a 3D mesh model and multi-view color images[C]// Proc. of the 8th Int'l Conf on CAD/Graphics. Macao: Welfare Printing limited, 2003:217-222
[8] Liu L, Stamos I. Multiview geometry for texture mapping 2D images onto 3D range data[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2006:2293-2300
[9] Stamos I, Allen P K. 3D model construction using range and image data[C]// Proceedings of Conference on Computer Vision and Pattern Recognition. 2000:531-536
[10] Iwashita Y, Kurazume R, Hasegawa T, et al. Fast alignment of 3D geometrical models and 2D color images using 2D distance maps[C]// Proceedings of the Conference on 3D Digital Imaging and Modeling. 2005:164-171
[11] 崔桂涣. 自由多视角恢复表面纹理的三维重建的研究[D]. 上海:上海大学,2009
Cui Gui-huan. Research on 3D reconstruction with free multi view to surface texture restoration [D]. Shanghai: Shanghai University, 2009
[12] Hui Wang, Yue Zhao. A New Planar Circle-based Approach for Camera Self-calibration[J]. Journal of Computational Information Systems, 2010, 6(9):2877-2883
[13] Wang Xiao-gang. Intelligent multi-camera video surveillance: A review[J]. Pattern Recognition Letters, 2013, 34(1):13-19
[14] Diao Chang-yu, Lu Dong-ming. Interactive high resolution texture mapping for the 3D models of cultural heritages [C]// VSMM 2007. Brisbane, Australia, 2007
[15] Fu Yan, Sun Jin. Graphic Texture Mapping Models for Animation Modeling[J]. Journal of Computational Information Systems, 2014, 10(16):6957-6964

(上接第 274 页)

Chen Xiao-jiang, Fang Ding-yi, Liu Wei, et al. Media stream component model based on CORBA[J]. Journal of Northwest University(Natural Science), 2005, 35(2):151-154
[4] Paris C, Wan S. Listening to the community: social media monitoring tasks for improving government services[C]// Proceedings of the International Conference on Human Factors in Computing Systems(CHI 2011). Extended Abstracts Volume, Vancouver, BC, 2011:2095-2100
[5] Braunstein S L, Pirandola S, Zyczkowski K. Better late than never: information retrieval from black holes[J]. Physical review letters, 2013, 110(10):101-108
[6] Fercoq O, Akian M, Bouhtou M, et al. Ergodic control and polyhedral approaches to PageRank optimization[J]. IEEE Transactions on Automatic Control, 2013, 58(1):134-148
[7] Venkatraman V, Ritchie D W. Flexible protein docking refinement using pose-dependent normal mode analysis[J]. Proteins: Structure, Function, and Bioinformatics, 2012, 80(9):2262-2274
[8] Sakakura Y, Yamaguchi Y, Amagasa T, et al. A Local Method for ObjectRank Estimation[C]// Proceedings of International Conference on Information Integration and Web-based Applications & Services. ACM, 2013:92-98
[9] Cao L, Cho B, Kim H D, et al. Delta-SimRank Computing on MapReduce[C]// BigMine'12, 2012. New York, NY, USA, 2012:28-35
[10] 李栋,徐志明,李生,等. 在线社会网络中信息扩散[J]. 计算机学报,2014,37(1):189-206
Li Dong, Xu Zhi-ming, Li Sheng, et al. A survey on Information Diffusion in Online Social Networks [J]. Chinese Journal of Computers, 2014, 37(1):189-206
[11] Zanardi V, Capra L. Uncovering Relevant Content Using Tag-based Recommender Systems [C]// RecSys'08, 2008. New York, NY, USA, 2008:51-58
[12] Sharma S K, Suman U. A Trust-based Architectural Framework for Collaborative Filtering Recommender System[J]. Int. J. Bus. Inf. Syst., 2014, 16(2):134-153
[13] Carrer-Neto W, María L, Valencia-García R, et al. Social Knowledge-based Recommender System[J]. Application to the Movies Domain. Expert Syst. Appl., 2012, 39(12):10990-11000
[14] Di Noia T, Mirizzi R, Ostuni V, et al. Exploiting the Web of Data in Model-based Recommender Systems[C]// RecSys'12, 2012, New York, NY, USA, 2012:253-256
[15] Wang B, Liao Q, Zhang C. Weight Based KNN Recommender System[C]// IHMSC'13, 2013. Washington DC, USA, 2013:449-452
[16] Verberne S, Halteren H, Theijssen D, et al. Learning to Rank for Why-question Answering[J]. Inf. Retr., 2011, 14(2):107-132
[17] 卞先华,陈亮,郑倩冰. 基于文本内容和社会结构的可信度[J]. 重庆理工大学学报(自然科学版),2013,27(1):57-61
Bian Xian-hua, Chen Liang, Zheng Qian-bing. Reliability Research Based on Text Context and Community Structure [J]. Journal of Chongqing University of Technology (Natural Science), 2013, 27(1):57-61