文章编号：2096-1472(2017)-09-15-03

基于张量分解的个性化微博推荐算法研究

秦晓晖
(太原工业学院计算机工程系，山西 太原 030008)

摘　要：随着社交媒体的发展，微博为人们提供的服务正在极大地改变着人们使用互联网的习惯，然而微博上用户发表的大量信息，以及高频率的信息更新，使得用户面临信息过载的问题而无法快速获取他感兴趣的信息。推荐系统是解决此问题的一种很好的方法，它是通过研究用户已有数据来发掘用户兴趣，从而为用户推荐可能感兴趣的对象，如产品、网页、微博等。本文介绍了一种基于张量分解技术的微博推荐算法来预测用户对微博的兴趣度，同时考虑用户与微博、用户与微博发布者影响因素，以及微博与微博发布者的影响因素，提高了已有算法的准确度。

关键词：微博推荐；矩阵分解；张量分解
中图分类号：TP391　　**文献标识码**：A

A Study of the Personalized Micro-Blog Recommendation Algorithm Based on Tensor Factorization

QIN Xiaohui
(*School of Computer Engineer*,*Taiyuan Institute of Technology*,*Taiyuan* 030008,*China*)

Abstract: With the development of social media,the services in micro-blog have significantly changed the way people use the Internet.However,as the large amount of information posted by users and the highly frequent update on micro-blogs,users often face the problem of information overload and miss out the content they are interested in.The recommendation system,which recommends items(such as products,web pages,micro-blogs,etc.)to users based on their interests,is an effective solution to this problem.The paper introduces a micro-blog recommendation algorithm based on the tensor factorization technology to predict the user's interest degree on certain micro-blog.The experimental results on real dataset show that the proposed model achieves desirable performance in characterizing the user's interest and the preprocessing of data on micro-blog.Finally,the paper presents the experimental results which show that the method significantly outperforms the baseline method.

Keywords:micro-blog recommendation;matrix factorization;tensor factorization

1　引言(Introduction)

目前，一些微博推荐算法在发掘用户在社交媒体中的兴趣和行为中表现出了一定的优越性，例如基于内容的推荐算法，但是目前大多数方法都通过内容等显性因素来预测用户的兴趣度而没有考虑一些内在的隐性因素。然而社交网络中的信息是丰富且复杂的，只通过一些显性因素来预测用户兴趣度是不够的。因子分解模型最初被用于推荐系统中来对用户感兴趣的商品进行推荐[1]。为了更好地对用户行为建模，一些研究使用隐因子模型对用户的兴趣度进行预测，而这些无法直接获取的隐性因素是影响用户兴趣度的主要因素。这些方法使用矩阵分解算法分别考虑用户和微博主题，用户和微博发布者之间的社会关系，以及微博发布者与微博主题之间的隐性因素，通过两两之间关系来预测用户对微博的兴趣度。然而，同样内容的微博被不同的发布者发布的话，用户的兴趣度是不同的，因此我们应综合考虑用户与微博，以及微博发布者它们之间的隐性因素共同对微博兴趣度的影响。

张量是对向量和矩阵的扩展[2]，因此它可以表示多元数据，已有的矩阵分解方法丢失了用户与微博，以及微博发布者三者之间在三维空间上对用户兴趣度的影响而张量分解模型很好地解决推荐系统中存在的多元影响因素[3]。而现实生活中的数据一般都具有多元特征，相对复杂，因此张量模型很好地模拟了推荐系统中数据的多元影响关系。

基金项目：太原工业学院科学基金项目(2016LQ14).

2 微博排序优化准则(Optimizing ranking criterion for weibo recommendation)

本文研究用户对微博的喜好度，这里使用协同排序算法对喜好度排序，它是基于隐因子模型的协同过滤方法[4]。我们用$p_u \in R^d$和$q_i \in R^d$表示用户和微博属性空间向量。计算用户u对微博i的喜好度见式(1)：

$$y_{u,i} = {p_u}^T q_i \tag{1}$$

y就是我们预测出来的用户u对微博i的喜好度。对于一条被用户转发了的微博k及一条没被用户转发的微博h来说，我们认为用户对微博k的喜好度将大于用户对微博h的喜好度，也就是$y_{u,k} > y_{u,h}$。因此，为了实现微博的排序我们将用户喜好度排序模型定义为计算求微博k在喜好度排序中比微博h靠前的可能性，详见式(2)：

$$P(r(k) > r(h)|u) = \frac{1}{1 + e^{-(y_{u,k} - y_{u,h})}} \tag{2}$$

$r(k)$和$r(h)$为微博k及微博h在所有微博中的排序，因此这个公司表达了微博k在用户喜好度排序中比微博h前的可能性，公式的右半部分为$y_{u,k} > y_{u,h}$的归一化处理形式。在这里，我们假设用户对他转发过的微博的喜好度将大于他对其他为转发过的微博的喜好度。因此我们需要构建数组D表示用户对转发微博喜好度大于非转发微博喜好度，见式(3)：

$$D = \{< u, k, h > | k \in Re(u), h \notin Re(u)\} \tag{3}$$

$Re(u)$为用户u转发过的所有微博的集合。这里我们的假设，用户对他转发过的微博会比其他所有他没有转发过的微博的喜好度都高，所以用户u可见的所有微博，Re集合里的每一个k元素都会与不在Re集合中的h元素组成一个D中的元素<u,k,h>。这里我们定义k为正例，h为负例。对公式(3)取对数进行最大似然估计更加便于计算，因此可以最终转化为求解目标函数(4)：

$$\text{Min} \sum_{<u,k,h>\in D} \ln\left(1 + \mathrm{e}^{-(y_{u,k} - y_{u,h})}\right) + r \tag{4}$$

其中，r为正则化参数。

3 基于张量的分解模型(Tensor factorization model)

本文需要同时考虑用户、微博、微博发布者这三个因素来预测用户对微博的兴趣度，即将二维矩阵拓展为三维张量来表示影响兴趣度的隐性因素，也就是分解用户—微博—发布者张量来预测用户对微博的喜好度。

为了和大多数的基于矩阵分解的推荐系统中的方法对比，我们可以将三维张量理解为在传统二维矩阵的基础上增加一个维度，即一种典型的张量分解方法Tucker分解，该分解模型产生的类似于SVD的左右奇异矩阵子结构方便与已有算法SVD进行实验结果对比[5,6]。Tucker分解把原张量分解为一个核心张量与一系列矩阵的乘积。这里我们以对三维张量$X \in R^{Ni \times Nj \times Nk}$的分解为例说明Tucker的具体分解过程，详见公式(5)：

$$X = S \times U \times V \times W = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} U_{ip} V_{jq} W_{kr} S_{pqr} \tag{5}$$

其中，S_{pqr}核心张量表示各维度上的相互作用，U、V、W分别是各维度上的因子矩阵，本文预测用户u对发布者p发布的微博t的喜好度，这里假设有N_u个用户，N_p个发布者，N_t条微博，那么预测一个用户对发布者发布的微博的兴趣度用一个三维张量y来表示，详见式(6)：

$$y = [y_{ijk}] N_u \times N_t \times N_p \tag{6}$$

其中，y_{ijk}表示用户u_i对于发布者p_k所发布的微博t_j的喜好度。三维张量可以分解为一个中心张量和三个矩阵，中心张量表示为$S \in R^{K \times K \times K}$，$K$为模型中隐因子的维度。潜在特征矩阵表示为$u \in R^{K \times Nu}$，$t \in R^{K \times Nt}$，$p \in R^{K \times Np}$，其中列向量$u_i$表示用户$u_i$的$K$维隐特征向量，列向量$t_j$表示微博$t_j$的$K$维隐特征向量，列向量$p_k$表示发布者$p_k$的$K$维隐特征向量。因此在矩阵分解模型的基础上，我们可以定义影响用户兴趣度隐因子的预测式(7)：

$$y_{u,i} = S \cdot {p_u}^T \left(\frac{1}{Z} \sum_{w \in T_i} q_w\right) d_{p(i)} \tag{7}$$

我们可以通过梯度下降的方法对这个协同排序模型进行学习。梯度下降的中的梯度指的就是数学中的求导方法，根据对公式中的每个维数公式求偏导得到梯度。因为函数将会沿着负梯度方向的下降最快，所以我们可以通过梯度下降的方法快速找到该函数的最优解。这里我们将数据分为多个数据组D，每组D包含一个用户的一个正例及一个负例。当我们使用梯度下降的方法时，我们要对每一组数据D都计算一次梯度，并对矩阵中的值进行更新直到循环终止得到最优解。其中，梯度更新系数详见式(8)—式(14)：

$$-\frac{\partial L}{\partial p_u} = \hat{e}\left(S \frac{1}{Zs^+} \sum_{s \in k} q_s - S \frac{1}{Zs^-} \sum_{s \in h} q_s + (d_{p(k)} - d_{p(h)})\right) - \sigma_1 p_u \tag{8}$$

$$-\frac{\partial L}{\partial {q_w}^+} = \hat{e} \cdot S\left(\frac{1}{Zj^+} p_u + d_{p(k)}\right) + \sigma_2 {q_w}^+ \tag{9}$$

$$-\frac{\partial L}{\partial {q_w}^-} = \hat{e} \cdot S\left(\frac{1}{Zj^-} p_u + d_{p(k)}\right) - \sigma_2 {q_w}^- \tag{10}$$

$$-\frac{\partial L}{\partial d_{p(k)}} = \hat{e} \cdot S\left(p_u + \frac{1}{Zj^+} {q_w}^+\right) - \sigma_3 d_{p(k)} \tag{11}$$

$$-\frac{\partial L}{\partial d_{p(h)}} = \hat{e} \cdot S\left(p_u + \frac{1}{Zj^-} {q_w}^-\right) - \sigma_3 d_{p(h)} \tag{12}$$

$$-\frac{\partial L}{\partial S}=\hat{e}\left(p_u\cdot\frac{1}{Zs^+}\sum\nolimits_{s\in k}q_s\cdot d_{p(k)}-p_u\cdot\frac{1}{Zs^-}\sum\nolimits_{s\in h}q_s\cdot d_{p(h)}\right)-\sigma_4 \quad (13)$$

$$-\frac{\partial L}{\partial b_j}=\hat{e}\left(\gamma_j{}^+ + \gamma_j{}^-\right)-\sigma_5 b_j \quad (14)$$

其中，“+”表示属于正例里的值，“−”表示属于负例中的值。$\hat{e}$表示真实值与预测值之间偏差的概率，详见式(15)。

$$\hat{e}=1-P\left(r(k)>r\left(\hat{k}\right)\middle|u\right)=1-\frac{1}{1+e^{-(y_{u,k}-y_{u,h})}} \quad (15)$$

在算法的循环中，我们将对每一组D中的值都进行一次梯度计算，并对矩阵进行更新。不断循环的过程中模型中的权重值将会沿着梯度下降的方向变化，直到获得模型的最优解。

4　实验(Experiment)

4.1　数据来源

本文数据来源于新浪微博，使用爬虫系统根据本文需求爬取相关数据[7]。网络爬虫作为一种自动提取网页信息的计算机程序或者自动化脚本[8]，它是搜索引擎的核心技术。本文先随机选取一个微博用户以发射状不断爬取该用户的关注者的数据，以及关注者的关注者的数据，然后从这些数据中选出1024个微博用户的主页信息，但这些用户的关注者人数需超过15。

4.2　评价标准

本文通过平均准确率评估预测结果的准确度。本文推荐模型的结果是微博的排序，同时微博的排序位置还关联了准确度使得推荐模型能得到更准确的评估，即微博成功推荐，如果它的排序越靠前那么平均准确率就越高。如果系成功推荐的微博个数为0那么准确率为0。评估公式见式(16)：

$$AP=\frac{\sum_{n=1}^{N}P@n\times retweet(n)}{|R|} \quad (16)$$

N是实验数据中测试集的微博数，R是用户转发的微博数，$retweet(n)$是布朗函数，如果第n条推荐微博用户转发过，即成功推荐，那么$retweet(n)$的值为1，如果第n条微博用户没有转发，那么$retweet(n)$的值为0。$P@n$是计算取排序结果中前n条微博时的准确度。MAP的值是取所有用户的AP值的平均值，见式(17)，N是用户总数。

$$MAP=\frac{AP}{N} \quad (17)$$

4.3　实验结果

为了验证算法的有效性，本文增加其他几种方法来对比实验结果，包括按照时间排序的方法、按相似度排序的方法、矩阵分解模型算法SVD[9]。张量分解算法(TF)综合考虑用户、微博和微博发布者三者之间的关系，较SVD更加准确地评估对用户兴趣度的影响。张量分解算法使用随机梯度算法来估计实验参数，矩阵分解过程中K值取30准确率最高。

表1 所有方法的MAP值

Tab.1 MAP results of all methods

方法	时间	相似度	SVD	TF
MAP	0.158	0.102	0.304	0.339

5　结论(Conclusion)

时间排序的推荐方法由于依赖用户的登录时间而对登录时间前后的微博转发的概率大，因此预测的准确度很低。相似度排序的算法只通过关键词计算微博表面相似度来预测而忽略了内在的语义。SVD只考虑用户、微博与微博发布者两两之间的关系，忽略三者之间的共同作用没有反映数据的真实信息而准确度低于TF方法。

参考文献(References)

[1] Lu J,et al.Recommender system application developments:a survey[J].Decision Support Systems,2015,74:12−32.

[2]Jain P,Oh S.Provable tensor factorization with missing data[C].Advances in Neural Information Processing Systems,2014:1431−1439.

[3] Ding G,Guo Y,Zhou J.Collective matrix factorization hashing for multimodal data[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2014:2075−2082.

[4] 冷亚军,陆青,梁昌勇.协同过滤推荐技术综述[J].模式识别与人工智能,2014,27(8):720−734.

[5] Rendle S.Factorization machines[A].The IEEE International Conference on Data Mining.Sydney:2010:995−1000.

[6] Cao Y.,et al.Adapting ranking SVM to document retrieval[C].The 29th Annual International SIGIR Conference.Seattle,WA:2006:186−193.

[7] 孙立伟,何国辉,吴礼发.网络爬虫技术的研究[J].电脑知识与技术,2010,6(15):4112−4115.

[8] 高建煌.个性化推荐系统技术与成用[D].中国科学技术大学,2010.

[9] 秦晓晖.基于协同过滤的个性化微博推荐算法研究[J].软件工程,2017,20(3):14−17.

作者简介：

秦晓晖(1987−)，女，硕士，助教.研究领域：中文信息处理，人工智能.