基于矩阵分解模型的微博好友推荐算法①

余 勇[1,2], 郭躬德[1,2]

[1](福建师范大学 数学与计算机科学学院, 福州 350007)
[2](网络安全与密码技术福建省重点实验室(福建师范大学), 福州 350007)

摘 要: 微博作为一种实时的信息传播和分享的社交网络平台, 对人们日常生活的影响越来越大. 在微博中, 用户可以通过关注关系, 添加自己感兴趣的好友, 扩大自己的交际圈. 但如何推荐高质量的关注好友, 一直是个性化服务的难点之一. 针对此种情况, 提出一种微博好友推荐算法, 旨在为用户推荐高质量的关注用户. 该算法是对基于 Seeker-Source 矩阵分解模型的一种改进算法. 文中分析了微博用户的多种数据源信息, 并给出了相应的特征提出方法, 最后将这些特征引入到 Seeker-Source 矩阵分解模型中, 通过对模型的优化求解, 得到最佳的参数因子矩阵, 从而完成好友推荐. 在真实的微博数据集上的实验表明, 本文所提出的算法取得了良好的效果.

关键词: 矩阵分解; 微博; 推荐算法; 社交网络

Algorithm for Micro-blog User's Followee Recommendation Based on Matrix Factorization

YU Yong[1,2], GUO Gong-De[1,2]

[1](School of Mathematics and Computer Science,Fujian Normal University, Fuzhou 350007, China)
[2](Key Laboratory of Network Security and Cryptography in Fujian Province(Fujian Normal University), Fuzhou 35007, China)

Abstract: Micro-blog is a social network platform that provides us a new communication and information sharing service. It has become more and more important in our daily life. An user can follow his interested friends to expand his social circle throw following relationship. But how to recommend high quality following users is always a difficulty of personalized service. For the issue, a Seeker-Source matrix factorization model based on micro-blog features is proposed in this paper. The algorithm is an improved algorithm which is based on "Seeker-Source". We extracted the characteristics of user's interest from each data source, and then introduced into the matrix factorization model which is suitable for recommending followee friends. Finally, we optimize the model and get the best factor parameter matrix to recommend followee friends. The experimental results carried on real data sets show that the proposed method performs better than the traditional matrix factorization model.

Key words: matrix factorization; micro-blog; recommendation algorithm; social network

1 引言

近年来, 随着 web2.0 技术的兴起, 社交网络服务已经发展成为互联网新兴媒体的一个重要服务. 其中, 微博就是一个典型代表. 微博是基于用户之间的关联关系, 构筑而成的一个信息传播和分享的平台. 不同于其他的社交网络服务, 微博不仅具有强大的交互功能, 而且还是一个承载了巨大信息的社会化媒体平台, 已经成为人们日常生活中获取信息和传播信息的重要途径. 国内外著名的微博平台有 Twitter, 新浪微博, 腾讯微博等.

据新浪微博发布的数据显示, 截至 2014 年 6 月底, 新浪微博日活跃用户量达到 6970 万. 如此庞大的日活跃用户量, 必定会产生出海量的微博数据. 那么, 如何从这些海量数据中快速有效地获取用户感兴趣的信息, 一直是信息检索和数据挖掘领域努力解决的问题.

推荐系统在互联网应用中一直占据着重要的作用,

① 基金项目:国家自然科学基金(61070062)
收稿时间:2015-03-31;收到修改稿时间:2015-06-03

如亚马逊和淘宝等电商中的商品推荐, 豆瓣中的电影推荐, 谷歌的新闻推荐等. 有关数据显示, 亚马逊网站推荐的销售转化率高达 35%. 微博作为一种社交网络, 不仅包含了大量的文本信息, 还具有复杂的社交网络特征, 允许一个用户关注其他用户, 从而建立起一个庞大的社交网络. 由于微博的这些特点, 吸引了众多的科研人员从各个方面对微博进行相关研究, 并试图从中挖掘出有用的信息. 如根据用户的资料和兴趣, 可以为其推荐可能感兴趣的个性化标签; 根据用户历史转发的微博和关注的话题, 可以为其推荐可能感兴趣的热点话题; 根据用户的社交拓扑图和兴趣爱好, 可以为其推荐可能感兴趣的好友等. 本文利用矩阵分解模型, 结合微博数据的各项特征, 提出一种针对微博用户可能感兴趣的潜在好友的推荐方法, 旨在给用户提供更好的个性化服务.

本文第 2 节将回顾微博推荐的相关工作; 第 3 节首先对传统的矩阵分解模型进行了介绍, 并随后介绍了一种适用于微博好友推荐的基于 Seeker-Source 的矩阵分解模型; 第 4 节针对微博数据源的不同特征, 首先给出提取特征的方法, 然后将这些特征引入到 Seeker-Source 矩阵分解模型中, 最后得到一种基于不同特征的联合模型. 第 5 节利用真实的微博数据集验证新模型的有效性; 最后总结全文, 并指出下一步的工作.

2 相关工作

在当前的推荐系统中, 无论是工业界还是在学术界, 推荐系统的技术都已经达到成熟, 应用也十分广泛. 就方法的角度看, 传统的推荐系统主要分为内容过滤类方法和协同过滤方法(Collaborative Filtering, CF)两大类. 内容过滤类方法为每个用户或商品建立能表达其自然特征的属性信息, 通过这些属性信息可以把用户和商品关联起来. 但基于内容的策略需要搜集额外的信息, 实现起来可能不是很容易. 协同过滤的方法是基于用户的历史记录或商品的评分, 不需要显式的建立属性信息. 协同过滤通过分析用户之间的关系和商品间的相互依赖关系, 来发现新的用户商品关联对.

相比传统的商品推荐, 微博推荐相关的研究较少, 大部分都是针对 Twitter 的研究, 但都具有一定的借鉴意义. 就微博的推荐方法, 可以分为基于社交拓扑图的方法和基于内容的方法. Armentano[1,2]等利用用户间的关注关系形成的社交拓扑结构来发现用户感兴趣的好友; 基于 PageRank 算法改进的 FolkRank[3]是典型的基于拓扑图的标签推荐系统, 通过对标签在图中的权重排序为用户推荐标签. 基于内容的方法主要是从用户的文本信息中挖掘有用信息. 如 Wu[4]等通过从用户的微博内容中抽取关键词, 作为该用户的个性化标签, 来描述每个用户的兴趣, 此方法还可以通过这些个性化标签词来衡量用户的相似性, 从而为目标用户推荐最相似的用户; 针对微博文本这种蕴含特殊的结构化信息, 张晨逸[5]等提出一种基于 LDA(Latent Dirichlet Allocation)模型的微博生成模型 MB-LDA (MicroBlog-Latent Dirichlet Allocation), 综合考虑了微博的联系人关联关系和文本关联关系, 来辅助进行微博的主题挖掘. 除此之外, 还有人尝试从社交拓扑图和内容相结合的方法着手, 如 Hannon[6]等将两类算法进行线性加权合并.

上述方法中, 无论是基于社交拓扑图的方法, 还是基于内容的方法, 都会遇到一个矩阵稀疏性的问题, 这是解决这类问题的难点之一. 然而, 矩阵分解模型是协同过滤算法中最有效的模型之一, 也是解决矩阵稀疏性问题的有效途径之一. 在 KDDCup2012 track1 (http://www.kddcup2012.org/c/kddcup2012-track1) 的竞赛中, 参赛者将矩阵分解应用于微博好友的推荐中, 在融合了矩阵分解和其他机器学习算法后, 取得了很好的效果[7-9]. Zhou 等[10]将决策树引入到矩阵分解模型中, 用以解决冷启动的问题. 主要思想为利用用户的自然属性(如性别、年龄等)建立决策树, 将决策树得到的分值作为某一维的量. 传统的矩阵分解方法是基于 point-wise 方法的训练和评价, 如 Root-Mean-Square-Error(RMSE). 这类方法倾向于衡量单个预测的准确性, 而忽略了被推荐物品的相对顺序. Rendle[11]等使用 pair-wise 结合 bayesian ranking 的方法, 在标签推荐中得到了优于其他方法的效果. 在微博好友这类推荐中, 由于微博拥有着丰富的数据特征, 且数据形式表现不一, 利用传统的矩阵分解方法, 其推荐的效果不明显. 本文中提出的方法, 考虑了微博好友在社交网络中特殊的角色作用, 并结合微博用户的多种数据特征, 逐步分析了不同特征对推荐结果的影响程度, 得到一个综合的推荐模型.

3 矩阵分解模型

矩阵分解模型是大规模协同过滤问题中最有效的

模型之一. 但传统的矩阵分解模型都是基于用户—项目的推荐, 能够很好的解决矩阵稀疏性问题. 针对微博好友这类特殊的推荐, 由于微博用户不同于商品, 在社交网络中充当不同的角色, 既是资源的提供者, 也是资源的接受者, 因而文献[12]在传统矩阵分解模型的基础上进行改进, 提出了一种适用于微博好友推荐的矩阵分解模型, 即基于 Seeker-Source 的矩阵分解模型. 本节首先对传统的矩阵分解模型进行介绍, 然后介绍了基于 Seeker-Source 的矩阵分解模型. 具体内容如下.

3.1 基于项目推荐的矩阵分解模型

在传统的推荐系统中, 存在一个 $m\times n$ 的用户—项目矩阵 R, 其中 m 代表用户数, n 代表项目数. 矩阵 R 中的每个元素 r_{ui} 为用户 u 对项目 i 的打分, 该值是用户针对历史上消费过的物品进行打分所留下的记录. 如亚马逊和淘宝允许用户对购买过的商品打分, 豆瓣中用户可以对电影进行打分. 显然用户只能对很少的商品进行打分, 矩阵 R 中的每一行都只有很少的元素, 所以矩阵 R 是一个稀疏矩阵. 矩阵分解的基本思想是将用户和项目映射到一个 d 维空间向量中, 每一维可以理解为用户在某一程度上的倾向, 即偏好程度[13].

令 $P^{m\times d}$ 和 $\mathrm{Q}^{d\times \mathrm{n}}$ 分别表示用户和项目的矩阵, 矩阵 P 的每一行代表一个用户的向量, 矩阵 Q 的每一行代表一个项目的向量. 每个用户 u 关联一个未知的特征向量 $p_u\in\Re^d$, 每个项目 i 关联一个未知的特征向量 $q_i\in\Re^d$. 用户 u 对项目 i 的预测评分为[13]:

$$\hat{r}_{ui}=p_u^{T}q_i \tag{1}$$

在最终的优化目标中, 希望预测的评分与真实的评分越接近越好, 但由于输入矩阵 R 是非常稀疏的, 很难处理其中的缺失值. 前期的工作中, 如文献[14]往往采用某种策略将缺失值进行填充, 但会带来繁重的计算量, 并且不适当的策略会使填充的数据不合理, 造成推荐的结果不准确. 文献[15]建议只对 R 中的已有元素进行建模, 故应采取某种策略对 R 中的已有元素进行建模, 可以得到如下优化函数:

$$\min_{P,Q} L=\sum_{r_{ui}\in R}(r_{ui}-\hat{r}_{ui})^2 \tag{2}$$

但在实际应用中为了防止出现过拟合现象, 所以需要添加 l_2 正则项:

$$\min_{P,Q} L=\sum_{r_{ui}\in R}(r_{ui}-\hat{r}_{ui})^2+\lambda\|P\|^2+\lambda\|Q\|^2 \tag{3}$$

其中, λ 为经验参数, 可以通过实验获得. 该优化函数可以用随机梯度下降法[11](Stochastic gradient descent)求解. 随机梯度下降法是最优化理论中最基础的优化算法之一, 它通过求参数的偏导数来找到下降速度最快的方向, 然后通过不断地迭代使目标函数达到极小值. 对于上面定义的优化函数, 首先对其求偏导, 得到如下结果:

$$\frac{\partial L}{\partial p_u}=-2q_ie_{ui}+2\lambda p_u \tag{4}$$

$$\frac{\partial L}{\partial q_i}=-2p_ue_{ui}+2\lambda q_i \tag{5}$$

然后, 根据随机梯度下降法, 将参数沿着下降速度最快的方向推进, 可以得到如下递推公式:

$$p_u\leftarrow p_u+\alpha(q_ie_{ui}-\lambda p_u) \tag{6}$$

$$q_i\leftarrow q_i+\alpha(p_ue_{ui}-\lambda q_i) \tag{7}$$

其中, α 是学习速率(learning rate), 可以通过实验反复测试获得. 但通常为了使算法尽快收敛, 需要在迭代的时候不断减小学习速率. 在实际的实验过程中, 根据递推公式(6)和(7)进行反复迭代, 直到目标函数达到最优值时算法结束, 此时可以得到最佳的因子参数矩阵 P 和 Q, 最后依据最终得到的参数矩阵给用户产生 top k 推荐目标.

3.2 基于 Seeker-Source 推荐的矩阵分解模型

由于矩阵分解很善于解决矩阵稀疏性问题, 所以矩阵分解模型在推荐系统中的应用十分广泛. 尤其是在近年来的 Netflix 和 KDDCup 大数据竞赛中, 矩阵分解模型更是取得了很好地效果[7-9]. 但众多研究者都是基于传统的用户—项目二耦图的推荐方法, 将一部分微博用户视为用户, 另一部分视为项目. 如图 1 所示, 图 1(a)是一个简单的用户间的社交网络拓扑图, 图 1(b)是基于图 1(a)形成的传统的用户—项目二耦图. 但微博好友推荐不同于传统的推荐方法, 微博用户在整个社交网络中不仅充当着资源的提供者(Source), 而且也是资源的接受者(Seeker)[16]. 除此之外, 传统的推荐方法也没有很好的利用社交网络中的一些结构化信息. 在图 1(b)中, 如果将用户 u_2 和 u_5 作为 User, 而其他用户作为 Item, 则就会缺少很多的关注信息, 如: u_2 关

注 u_3, u_3 关注 u_4, u_4 关注 u_2, u_5 关注 u_1 等信息.

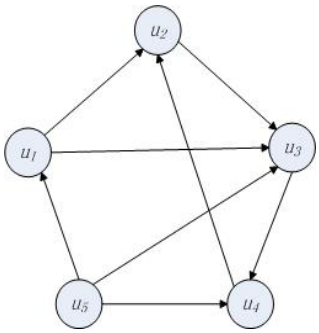

图 1(a)社交网络拓扑图

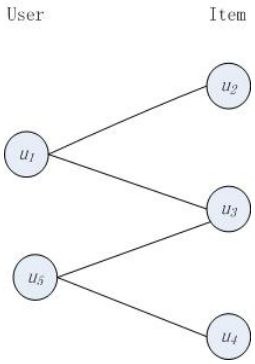

图 1(b)用户-项目二耦图

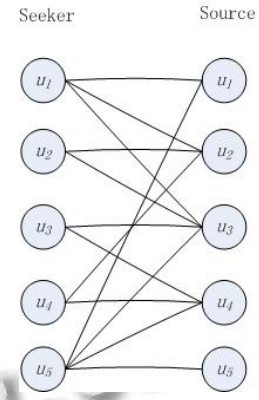

图 1(c)Seeker-Source 二耦图

在微博这种特殊的社交网络中, 用户通过关注可以随意的添加自己感兴趣的用户, 并且可以自动接收已关注用户发布的信息. 通过关注关系, 可以形成一个庞大的社交网络图 $G(U,E)$, 其中, U 表示网络图中的用户集合, E 表示整个社交网络中的关注关系. 如果用户 u 关注用户 v, 则表示为一条有向边 $u \to v \in E$. 用户 u 的关注用户集合可以表示为 $\Gamma_+(u) = \{v \in U \mid u \to v \in E\}$, $|\Gamma_+(u)|$ 表示用户 u 的关注数. 同理, 用户 u 的粉丝集合可以表示为 $\Gamma_-(u) = \{v \in U \mid v \to u \in E\}$, $|\Gamma_-(u)|$ 表示用户 u 的粉丝数.

由于微博用户不仅充当着资源的提供者, 而且也是资源的接受者. 因而可以用两个特征向量 p_u、q_u 来描述用户 u. 而 $p_u{}^T q_i$ 表示用户 u 对用户 v 的偏好程度. 如图 1(c)所示. 与传统的用户—项目推荐不同, 微博好友推荐就是给目标用户一个 Top K 的用户推荐列表. 因此, 我们可以将其视为是一个基于喜好对的微博用户排序(pairwise-ranking)[12]: 如果用户 u 关注了用户 i, 则我们认为用户 u 对用户 i 的喜好程度要高于没有关注的其他用户. 如图 1(a)中, 用户 u_1 关注了用户 u_2, 而没有关注用户 u_4, 则认为用户 u_1 对用户 u_2 的偏好程度要高于用户 u_4, 将其定义为 $u_2 >_{u_1} u_4$. 我们假设目标用户对自身是感兴趣的, 故对其自身也是可以关注的. 对于训练集

$$D_s = \{(u,i,j) \mid u \in U \wedge i \in U \wedge j \in U \wedge u \to i \in E \wedge u \to j \notin E\},$$

$(u,i,j) \in D_s$ 表示对于用户 i 和 j 而言, 用户 u 更偏好于用户 i, 则模型给出的预测应满足 $\hat{r}_{ui} > \hat{r}_{uj}$. 对于每个用户 u 可以得到优化函数[16]:

$$\min_{P,Q} L = \sum_{(u,i,j) \in D_s} -\log(\sigma(p_u{}^T q_i - p_u{}^T q_j)) + \lambda \|P\|^2 + \lambda \|Q\|^2 \tag{8}$$

其中, σ 是一个逻辑 S 型函数, $\sigma(x) = \dfrac{1}{1+e^{-x}}$; λ 是一个非负的经验参数. 随后, 我们可以用随机梯度下降法来局部最小化目标函数, 并求解其中的特征向量 p_u, q_i 和 q_j:

$$\frac{\partial L}{\partial p_u} = \sum_{(u,i,j) \in D_s} \frac{-(q_i - q_j)}{1 + e^{(p_u{}^T q_i - p_u{}^T q_j)}} + \lambda p_u \tag{9}$$

$$\frac{\partial L}{\partial q_i} = \sum_{(u,i,j) \in D_s} \frac{-p_u}{1 + e^{(p_u{}^T q_i - p_u{}^T q_j)}} + \lambda q_i \tag{10}$$

$$\frac{\partial L}{\partial q_j} = \sum_{(u,i,j) \in D_s} \frac{p_u}{1 + e^{(p_u{}^T q_i - p_u{}^T q_j)}} + \lambda q_j \tag{11}$$

4 基于微博特征的Seeker-Source矩阵分解模型

虽然基于 Seeker-Source 的矩阵分解模型是对传统矩阵分解模型的一种改进, 考虑到了微博用户在社交网络中的不同作用, 其推荐准确率也有所提升. 但由于微博用户拥有多种数据特征, 且每种数据特征都不同程度的反映了用户的兴趣偏好, 所以, 本文在前者研究的基础上, 分别引入了微博用户自然属性特征、微博关键词和标签特征、社交特征、结构化特征, 来分析不同微博特征对推荐好友准确率的贡献度, 并最终得到一个联合模型, 即基于微博特征的 Seeker-Source 矩阵分解模型. 具体方法如下.

4.1 微博用户自然属性特征

微博用户的自然属性特征(profile)包括用户的性别和年龄信息. 不同年龄段和不同性别的用户可能有各自不同的兴趣偏好. 从统计学的角度看, 这些兴趣偏好会随着年龄和性别呈现一定的规律性. 因而, 在建模时应考虑用户的性别和年龄特征. 根据文献[7]提出的方法, 我们将年龄分成 k 个群组, 且每个群组中根据性别再分成两个群组, 共计 $2k$ 个群组. 将用户 u 映射到所属的群组中 $age(u):u\to\{1,2,...,2k\}$, 则用户 u 的年龄、性别特征可以用一个大小为 $2k$ 的向量表示, 用户所属群组的单元值为 1, 其它为 0. 故对每一个群组, 定义一个因子向量 p_l, 则基于用户自然属性特征的偏好预测可以定义为[7]:

$$\hat{r}_{ui}=b_u+b_i+(\sum_{l=1}^{2k}p_l)^T q_i \tag{12}$$

4.2 微博关键词和标签特征

微博关键词(key words)是从用户发布的微博文本中提取出来的一些关键词, 很大程度上反应了用户的兴趣爱好, 常用来作为基于内容的微博推荐. 微博关键词的权重, 常用 TF-IDF(term frequency-inverse document frequency)技术来计算得到.

TF-IDF 是一种用于信息搜索和信息挖掘的常用加权技术. 词频(Term Frequency, TF)是指某个词语在文档中出现的次数. 文档频度(Document Frequency, DF)是指某个词语在文档集合中出现的次数, 每篇文档中只算一次. 其主要思想是: 如果某个词语在一篇文档中出现的频率 TF 较高并且在其他文档中很少出现, 则认为该词语具有很好的类别区分能力.

对于微博而言, 我们将微博数据集中所有用户发布的微博作为一个文档集, 而每个用户发布的微博作为一个文档. 为了表示 TF-IDF, 假设微博数据集中共有 N 个用户, 即 N 篇文档, 每篇文档中有 n 个关键词 $K(u)=\{k_1,k_2,...,k_n\}$, 其每个关键词的 TF 向量为 $(f_1,f_2,...,f_n)$, 其中 f_i 表示关键词 k_i 的 TF 值. 每个关键词的 DF 向量为 $(f_1^{'},f_2^{'},...,f_n^{'})$, 其中 $f_i^{'}$ 表示关键词 k_i 的 DF 值. 故用户 u 的中关键词 k_i 在文档向量中的 TF-IDF 权重计算公式为:

$$w_{u,i}=TF(i)\times IDF(i)=f_i\times\log(\frac{N}{f_i^{'}}) \tag{13}$$

对每个用户的关键词进行归一化处理, 得到如下公式:

$$w_{u,i}^{'}=\frac{w_{u,i}}{\|w_u\|_2} \tag{14}$$

其中 w_u 表示用户 u 的所有关键词权重向量, 即 $w_u=(w_{u1},w_{u2},...,w_{un})$.

在矩阵分解模型中引入关键词特征, 则用户 u 的潜在因子向量可以定义为[17]:

$$p_u^{'}=p_u+\frac{1}{\|w_u\|_2}\sum_{k\in K(u)}w_{u,k}y_k \tag{15}$$

其中 $K(u)$表示用户 u 的关键词集合, y_k 表示关键词 k 的潜在因子, $w_{u,k}$ 表示用户 u 中的关键词 k 的权重, w_u 表示用户 u 的所有关键词权重向量.

对于微博的另一内容类特征—标签(tags), 是微博用户自行添加的个性化信息, 一般由词语或短语组成, 用以表征用户对某些领域的兴趣. 定义用户 u 的标签集合为 $Tag(u)=\{t_1,t_2,...,t_n\}$, 故引入用户的标签特征, 则用户 u 的潜在因子向量可以定义为[17]:

$$p_u^{'}=p_u+\frac{1}{\sqrt{|Tag(u)|}}\sum_{t\in Tag(u)}p_t \tag{16}$$

其中, Tag(u)是用户 u 的标签集合, $1/\sqrt{|Tag(u)|}$ 是对标签的归一化处理, p_t 是用户 u 对标签 t 的潜在因子.

故综合考虑微博关键词和标签特征, 相应的模型为:

$$\hat{r}_{ui}=b_u+b_i+(\frac{1}{\|W_u\|_2}\sum_{k\in K(u)}W_{u,k}y_k+\frac{1}{\sqrt{|Tag(u)|}}\sum_{t\in Tag(u)}p_t)^T q_i \tag{17}$$

4.3 社交特征

社交功能是微博的重要功能之一, 微博用户可以在微博平台中随意添加自己感兴趣的用户作为好友, 实时关注他的动态. 假设用户 u 的关注用户集合为 $F(u)$, 则 $F(u)$ 中每个用户对用户 u 都会产生一定的影响. 故在矩阵分解模型中引入用户的关注好友信息, 用户 u 的潜在因子向量可以定义为[7]:

$$p_u^{'}=p_u+\frac{1}{\sqrt{|F(u)|}}\sum_{k\in F(u)}p_k \tag{18}$$

其中, $F(u)$ 表示用户 u 的关注好友集合, $1/\sqrt{|F(u)|}$ 是对关注好友的归一化处理, p_k 表示用户 u 对关注好友 k 的潜在因子向量.

社交信息中, 除了用户之间的关注信息外, 还有用户之间的交互行为信息. 用户可以对感兴趣的微博

进行转发、评论, 还可以在微博文本中@(提及)其他用户. 假设用户 u 的交互用户集合为 $A(u)$, 对于某一用户 $x(x \in A(u))$, 用 $a_{u,x}$ 表示用户 u 对用户 x 的交互次数, 即用户 u 对用户 x 转发、评论、提及的总次数. 则用户 u 的交互向量可以表示为 $a(u)$. 故在矩阵分解模型中引入用户的交互行为信息后, 用户 u 的潜在因子向量可以定义为[7]:

$$p_u^{'} = p_u + \frac{1}{\| a(u) \|_2} \sum_{x \in A(u)} a_{u,x} p_x^{'} \quad (19)$$

其中, $1/\| a(u) \|_2$ 是对交互行为的归一化处理.

综上所述综合考虑用户的关注信息和交互行为信息, 可以得到如下模型:

$$\hat{r}_{ui} = b_u + b_i + (\frac{1}{\sqrt{|F(u)|}} \sum_{k \in F(u)} p_k + \frac{1}{\| a(u) \|_2} \sum_{x \in A(u)} a_{u,x} p_x^{'})^T q_i \quad (20)$$

其中, $F(u)$ 表示用户 u 关注的用户集合, $A(u)$ 表示与用户 u 发生交互行为的交互用户集合, $a_{u,x}$ 表示用户 u 对用户 x 的交互行为的次数, p_k 表示用户 u 关注的用户 k 的潜在因子, $p_x^{'}$ 表示与用户 u 发生交互行为的用户 x 的潜在因子.

微博中, 用户间的关系无外乎有四种关系: ① $u_1 \rightarrow u_2$; ② $u_1 \leftarrow u_2$; ③ $u_1 \leftrightarrow u_2$; ④u_1 和 u_2 没有直接关注关系. 用户间通过这些关系直接或间接的分享其他微博用户发布的信息. 另外, 社交网络中的结构化信息(structure)对微博好友的推荐有非常重要的作用[18,19]. 如图 2(a)所示是用户间的传递性: 如果 $u_1 \rightarrow u_2$, 并且 $u_2 \rightarrow u_3$, 则可以认为 u_1 是 u_3 的潜在关注用户, 即 $u_1 \rightarrow u_3$. 故可以用用户 u 关注的用户 f 的偏好 p_f 来代替用户 u 的偏好 p_u. 图 2(b)所示的是共同关注型的情形: 如果 $u_1 \rightarrow u_2$, 且 $u_3 \rightarrow u_2$, 则认为用户 u_1 和 u_3 是相似的 Seeker. 此种情形可以用与用户 u 相似的 Seeker 用户 g 的偏好 p_g 来代替用户 u 的偏好 p_u. 图 2(c)所示的是共同被关注型的情形: 如果 $u_2 \rightarrow u_1$, 且 $u_2 \rightarrow u_3$, 则认为用户 u_1 和 u_2 是相似的 Source. 同理, 可以用与用户 u 相似的 Source 用户 h 的偏好 q_h 来代替用户 u 的偏好 q_u. 在这些结构化信息中, 利用一种改进的 Jaccard 系数来衡量两个用户的相似度, 公式如下:

$$Sim(u,v) = \frac{\sum_{w \in \Gamma_+(u) \cap \Gamma_+(v)} \log |\Gamma_-(w)|}{\sum_{w \in \Gamma_+(u) \cup \Gamma_+(v)} \log |\Gamma_-(w)|} \quad (21)$$

4.4 结构化信息

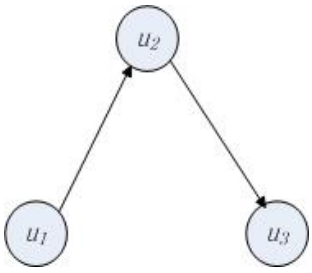

图 2(a)传递性

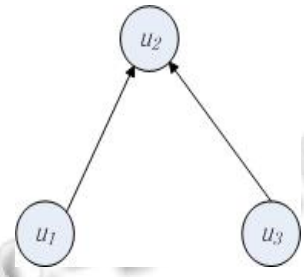

图 2(b)共同关注型

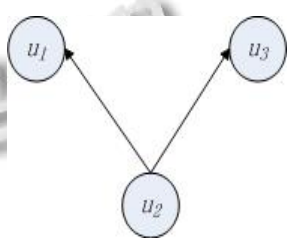

图 2(c)共同被关注

将这三种结构化信息作为基于 Seeker-Source 矩阵分解模型的正则项[16]:

$$\frac{\alpha}{2} \sum_{u \in U} \sum_{f \in \Gamma_+(u)} Sim(u,f) \| p_u - p_f \|^2 + \frac{\beta}{2} \sum_{u \in U} \sum_{g \in G(u)} Sim(u,g) \| p_u - p_g \|^2 + \frac{\gamma}{2} \sum_{u \in U} \sum_{h \in H(u)} Sim(u,h) \| q_u - q_h \|^2 \quad (22)$$

其中, $\Gamma_+(u)$ 表示用户 u 关注的用户集合, α 表述关于用户传递性的非负参数, $G(u)$ 表示与用户 u 相似的 Seeker 用户集合, β 表示关于相似 Seeker 的非负参数, $H(u)$ 表示与用户 u 相似的 Source 用户集合, γ 表示关于相似 Source 的非负参数, $Sim()$ 表示两个用户的相似度.

4.5 联合模型

以上对微博的各项特征信息进行了分析, 并都巧妙的引入了矩阵分解模型中. 每种数据特征都能够反映用户潜在的兴趣, 综合这些特征将对用户的潜在兴趣挖掘, 有着极大地帮助. 首先, 我们综合考虑微博用户的自然属性信息、微博关键词信息、标签信息、

社交信息, 得到一个综合的模型, 如下:

$$\hat{r}_{ui}=b_u+b_i+(\sum_{l=1}^{2k}p_l+\frac{1}{\|w_u\|_2}\sum_{k\in K(u)}w_{u,k}y_k+\frac{1}{\sqrt{|Tag(u)|}}\sum_{t\in Tag(u)}p_t+\frac{1}{\sqrt{|F(u)|}}\sum_{k\in F(u)}p_k+\frac{1}{\|a(u)\|_2}\sum_{x\in A(u)}a_{u,x}p_x')^Tq_i \tag{23}$$

而对于结构化信息, 我们将其作为上述模型的正则项, 故模型的优化函数为:

$$\begin{aligned}\min_{P,Q}L=&\sum_{(u,i,j)\in D_s}-\log(\sigma(p_u{}^Tq_i-p_u{}^Tq_j))\\&+\frac{\alpha}{2}\sum_{u\in U}\sum_{f\in\Gamma_+(u)}Sim(u,f)\|p_u-p_f\|^2\\&+\frac{\beta}{2}\sum_{u\in U}\sum_{g\in G(u)}Sim(u,g)\|p_u-p_g\|^2\\&+\frac{\gamma}{2}\sum_{u\in U}\sum_{h\in H(u)}Sim(u,h)\|q_u-q_h\|^2\\&+\lambda\|P\|^2+\lambda\|Q\|^2\end{aligned} \tag{24}$$

其中,

$$\begin{aligned}p_u=&\sum_{l=1}^{2k}p_l+\frac{1}{\|W_u\|_2}\sum_{k\in K(u)}W_{u,k}y_k+\\&\frac{1}{\sqrt{|Tag(u)|}}\sum_{t\in Tag(u)}p_t+\frac{1}{\sqrt{|F(u)|}}\sum_{k\in F(u)}p_k\\&+\frac{1}{\|A(u)\|_2}\sum_{t\in A(u)}a_{u,t}p_t'\end{aligned}$$

5 实验与分析

为了测试本文所提出的算法, 进行了相关的对比实验: (1)基于传统的 user-item 推荐的矩阵分解模型; (2)基于 Seeker-Source 推荐的矩阵分解模型; (3)基于各种微博特征的矩阵分解模型; (4)联合矩阵分解模型.

5.1 实验数据集与实验环境

为了保证实验结果的可信性, 本文采用 KDDCup 2012 track1(http://www.kddcup2012.org/c/kddcup 2012-track1)中所公开使用的数据集, 该数据集来源于腾讯微博推荐系统的后台数据. 该数据集包含了丰富的微博数据信息, 如: 用户的个人信息, 用户的个性化标签信息, 用户的微博关键词信息, 用户间的社交关注信息, 以及用户的交互行为信息等等.

实验中根据用户的关注时间, 将数据集划分成两部分: 训练集和测试集. 训练集包含 2011 年 11 月之前关注的用户信息, 测试集则包含 2011 年 11 月之后关注的信息.

实验中所采用的实验环境为: Intel(R) Core(TM) i5-3470 CPU @3.20GHz 3.20GHz 处理器, 4GB 内存, Windows 7 操作系统, Eclipse 开发平台.

5.2 实验结果与分析

实验中根据文献[19]的建议, 将潜在特征的维数 λ 设置为 50, 矩阵分解模型的学习率设置为 0.0005, 参数设置为 0.004. 实验中采用 MAP@k 来评估微博好友的推荐准确度(k 分别取 1、3、5、10), 该值越大, 说明推荐效果越准确.

首先对本文中提出的各模型进行独立的实验, 如表 1 所示. 前两种分别是基于传统的 User-Item 推荐方法和基于 Seeker-Source 的方法, 后四种是在基于 Seeker-Source 方法的基础上引入不同的微博特征的方法. 从实验结果可以看出, 基于 Seeker-Source 的矩阵分解模型要明显好于传统的 User-Item 矩阵分解模型, 这与文献[19]所得出的实验结论是一致的. 另外, 在融合不同微博特征后的 Seeker-Source 矩阵分解模型, 也都表现出很好的推荐效果. 不同微博特征对好友推荐的准确率方面都有所提升, 尤其是微博的社交特征和结构化信息特征, 要明显强于其他特征.

表 1 几种算法的对比实验结果

Algorithm	MAP@1	MAP@3	MAP@5	MAP@10
User_Item	0.4180	0.2475	0.1842	0.1182
Seeker_Source	0.4236	0.2510	0.1857	0.1195
Seeker_Source_profile	0.4495	0.2717	0.2024	0.1326
Seeker_Source_keys_tags	0.4309	0.2583	0.1916	0.1200
Seeker_Source_sns	0.4516	0.2810	0.2056	0.1367
Seeker_Source_structure	0.4583	0.2899	0.2103	0.1422

表 2 所展示的是在 Seeker-Source 方法的基础上, 逐步考虑微博用户特征(自然属性特征、关键词和标签特征、社交特征、结构化特征)的对比实验结果. 从表 2 中可以看出, 每加入一种微博特征后, 其推荐准确度

都有所提升. 但针对不同的微博特征, 提升的推荐准确度是不同的, 这说明不同博数据源的特征对用户的贡献度是不一样的. 图 3 是 MAP@k(k 取 1、3、5、10)时的总平均准确度. 从图中可以看出, 社交网络特征对推荐准确度的提升作用最大, 而关键词和标签特征对推荐准确度的提升作用不大. 究其原因, 首先关键词是从用户发布的微博文本中提取出来的, 而微博文本一般都很简短, 含有大量多的网络用语、表情符号和链接信息等, 并且用户发布微博都很随意, 大部分都是表达自己的心情感受类的信息, 因而用这些关键词很难完整表达用户的兴趣; 其次标签是用户自身添加用来表征自己兴趣爱好的关键词, 但绝大多数的用户都没有添加标签; 无论是关键词矩阵还是标签矩阵, 都非常稀疏. 综合这些因素, 微博用户的关键词和标签特征对好友推荐的作用不是很大. 在融合各种微博特征信息之后的联合模型中, 好友推荐准确度大幅度提升. 这是因为, 随着多种数据特征的加入, 用户的兴趣偏好才会更加准确地刻画出来, 这样一来, 在推荐好友时, 才会更加容易的找到兴趣偏好相似的潜在好友, 且推荐质量也会更高.

表 2 基于不同特征的 Seeker-Source 算法实验结果

ID	Algorithm	MAP@1	MAP@3	MAP@5	MAP@10
1	Seeker_Source	0.4236	0.2510	0.1857	0.1195
2	1+ profile	0.4495	0.2717	0.2024	0.1326
3	2+ keys_tags	0.4529	0.2806	0.2144	0.1503
4	3+ sns	0.4837	0.3175	0.2689	0.1778
5	4+structure	0.5122	0.3581	0.3002	0.1909

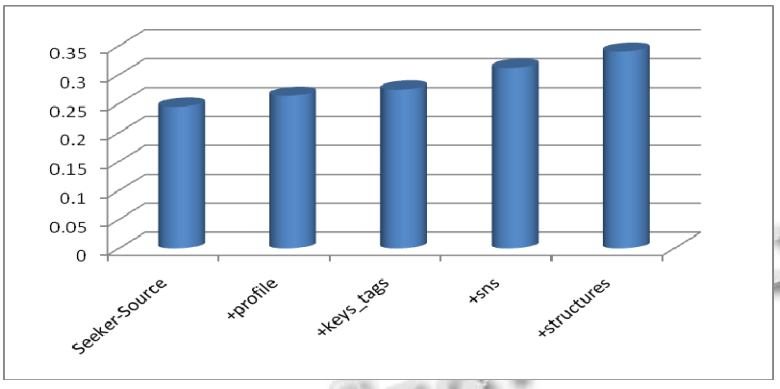

图 3 Seeker-Source 利用几种微博特征数据的总平均推荐准确度

6 结语

本章针对社交网络中用户可能感兴趣的潜在好友, 提出了一种基于矩阵分解模型的微博好友推荐算法. 该算法充分考虑了微博数据源的多样性, 针对每种数据源进行特征提取, 并将其引入到适合微博好友推荐的 Seeker-Source 矩阵分解模型中, 通过对模型的优化求解, 找到最佳的参数因子矩阵, 完成好友推荐. 最后通过在真实微博数据集上的实验验证了本文所提出算法的有效性. 在未来的工作中, 我们将继续围绕矩阵分解模型, 更加深入的考虑微博数据特征内在的关联性, 以便提高推荐质量.

参考文献

1 Armentano MG, Godoy D, Amandi A. Towards a followee recommender system for information seeking users in twitter. Proc. of the Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop Proceedings. 2011, 730: 27–38.

2 Armentano MG, Godoy D, Amandi A. Topology-based recommendation of users in micro-blogging communities. Journal of Computer Science and Technology, 2012, 27(3): 624–634.

3 Jäschke R, Marinho L, Hotho A, et al. Tag recommendations

in folksonomies. Knowledge Discovery in Databases: PKDD 2007. Springer Berlin Heidelberg, 2007: 506–514.

4 Wu W, Zhang B, Ostendorf M. Automatic generation of personalized annotation tags for twitter users. Human language technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 689–692.

5 张晨逸,孙建伶,丁轶群.基于 MB-LDA 模型的微博主题挖掘.计算机研究与发展,2011,48(10):1795–1802.

6 Hannon J, Bennett M, Smyth B. Recommending twitter users to follow using content and collaborative filtering approaches. Proc. of the Fourth ACM Conference on Recommender Systems. ACM, 2010: 199–206.

7 Chen T, Tang L, Liu Q, et al. Combining factorization model and additive forest for collaborative followee recommend-dation. KDD CUP, 2012.

8 Chen Y, Liu Z, Ji D, et al. Context-aware ensemble of multifaceted factorization models for recommendation prediction in social networks. KDD-Cup Workshop. 2012.

9 Ma T, Yang Y, Wang L, et al. Recommending people to follow using asymmetric factor models with social graphs. Soft Computing in Industrial Applications. Springer International Publishing, 2014: 265–276.

10 Zhou K, Yang SH, Zha H. Functional matrix factorizations for cold-start recommendation. Proc. of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2011: 315–324.

11 Rendle S, Freudenthaler C, Gantner Z, et al. BPR: Bayesian personalized ranking from implicit feedback. Proc. of the 25th Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2009: 452–461.

12 Yin D, Hong L, Davison BD. Structural link analysis and prediction in microblogs. Proc. of the 20th ACM International Conference on Information and Knowledge Management. ACM, 2011: 1163–1168.

13 Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. Computer, 2009 (8): 30–37.

14 Sarwar B, Karypis G, Konstan J, et al. Application of dimensionality reduction in recommender system-a case study. Minnesota Univ Minneapolis Dept of Computer Science, 2000.

15 Paterek A. Improving regularized singular value decomposition for collaborative filtering. Proceedings of KDD cup and workshop. 2007, 2007: 5–8.

16 Java A, Song X, Finin T, et al. Why we twitter: understanding microblogging usage and communities. Proc. of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web Mining and Social Network Analysis. ACM, 2007: 56–65.

17 Zhao X. Scorecard with latent factor models for user follow prediction problem. KDD-Cup Workshop. 2012.

18 Golder SA, Yardi S. Structural predictors of tie formation in twitter: Transitivity and mutuality. Social Computing (SocialCom), 2010 IEEE Second International Conference on. IEEE, 2010: 88–95.

19 Yu Y, Qiu RG. Followee Recommendation in Microblog Using Matrix Factorization Model with Structural Regularization. The Scientific World Journal, 2014, 2014.