基于矩阵分解的社交网络正则化推荐模型①

林晓勇, 代苓苓, 史晟辉, 李 芳

(北京化工大学 信息科学与技术学院, 北京 100029)

摘 要: 社交网站的快速发展和普及使得实现高效的好友推荐成为了一个热点问题, 而矩阵分解算法是被业界广泛采用的方法. 虽然传统的矩阵分解算法能够带来良好的效果, 但是仍然存在一些问题. 首先, 算法没有充分利用用户之间的社交网络结构化关系; 其次, 算法依赖的用户-物品评分矩阵只有二级评分不能充分表达用户的喜好. 提出了一种基于矩阵分解的社交网络正则化推荐模型, 利用社交网络中用户的近邻关系进行建模, 并将其作为一种辅助信息融合到矩阵分解模型当中, 该模型能够解决传统矩阵分解面临的问题. 通过在腾讯微博数据集上进行实验对比, 验证了本文提出的方法与传统的推荐方法相比能取得更高的推荐平均准确度.

关键词: 社交网络; 矩阵分解; 好友推荐; 近邻关系; 平均准确度

Recommendation Model of Matrix Factorization Based on Social Network Regularization

LIN Xiao-Yong, DAI Ling-Ling, SHI Sheng-Hui, LI Fang

(Beijing University of Chemical Technology, College of Information Science & Technology, Beijing 100029, China)

Abstract: With the rapid development and popularization of social network site, how to achieve efficient friend recommendation has become a hot issue. Currently, Matrix Factorization algorithm is widely used method by industry. Although the traditional Matrix Factorization algorithm could bring a good results, but there are still some problems. First, this model does not take full advantage of structural relationship between users in social network; Secondly, this algorithm is dependent on the user-rating matrix, which only has secondary scoring and cannot fully express the user's preferences. In order to solve these two problems, a Matrix Factorization model with social network regularization was proposed in this paper, modeling use of social network users in the model the relationship between neighbors. And as an auxiliary information fusion to the matrix Decomposition Model. This model can solve the problems that traditional Matrix Factorization model cannot solve. Though the contrast experiments on tencent weibo data set, verify that our proposed method could obtain a higher mean average precision than other traditional methods.

Key words: social network; matrix factorization; friend recommendation; neighborhood relation; mean average precision

微博是一种基于关注机制分享简短实时信息的广播式的社交网络平台. 在这个平台上用户能够随时随地发布自己的思想和最新动态. 目前微博已经成为了一种非常受欢迎的网络服务. 中国微博在发展最好的时期仅每天的信息量能达 2 亿多条, 但是海量信息的出现使得微博用户想要找到感兴趣的信息就变得非常困难. 针对这个问题, 如果网站能够为用户建立一个其感兴趣的朋友圈, 用户在这个小世界里分享信息就不会面临信息过载的问题. 社交网站中好友推荐系统的出现使得这个问题迎刃而解. 如何建立一个高效的好友推荐系统也成为了互联网行业的热点问题.

好友推荐系统的目的是根据用户现在的好友关系、用户行为记录给用户推荐新的朋友, 从而增强用户之间的黏性和用户与网站的黏性. 协同过滤算法是

① 基金项目:国家自然科学基金(61304237)

收稿时间:2015-04-16;收到修改稿时间:2015-05-15

推荐领域应用最广泛的算法, 包括 Model-based 协同过滤[1,2]和 Memory-based 协同过滤[3,4]. Model-based 算法近年被很多研究者研究, 其中最著名的为基于矩阵分解的推荐算法[5,6]. 该算法利用低维的向量表示用户和物品, 通过用户与物品的潜在向量内积计算预测值.

基于矩阵分解的模型在电子商务网站中获得了很好的推荐效果, 但是把矩阵分解模型应用于好友推荐存在一些问题: 首先, 矩阵分解依赖于一个用户-物品评分矩阵, 评分值代表用户对物品的打分, 反映了用户的喜好. 好友推荐系统中的评分范围为 0-1. 1 表示接受推荐对象, 0 表示拒绝推荐对象. 该取值不能充分表达用户喜好的强度; 其次, 传统的方法没有充分利用社交网络中的结构化信息. 社交网络是一个关系网, 每一个关注对象都代表了用户的兴趣爱好. 充分挖掘网络中的关系, 将能够更加准确的预测出用户对推荐好友的接受程度.

因此, 针对以上问题, 本文提出了一种基于矩阵分解的社交网络正则化推荐模型. 模型的核心思想是在基于排序的矩阵分解模型中加入带有社交网络结构特征的正则化项, 利用特定网络结构中用户的最近邻有效改善推荐结果.

1 相关工作

在 2006 年, Netflix Prize 开始之后, Simon Funk 公布了潜在因子模型(Latent Factor Model, LFM). 由于其较高的预测准确度, 从而获得了大批研究者的研究. 矩阵分解模型是一种重要的潜在因子模型. 在这个模型中我们使用两个低维的向量 p_u 和 q_i 来描述用户, p_u 表示用户的潜在特征向量, q_i 表示物品的潜在特征向量. 则内积 $p_u{}^T\mathrm{q_i}$ 就可以表示用户 u 对物品 i 的喜好预测值, 预测公式如下:

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

该模型不仅能够带来良好的推荐结果而且易于扩展, 利用社交网络特征扩展该模型的算法被很多研究者提出.

Chen 等人[7]提出了基于社交网络的用户反馈模型, 模型假设曾经关注了某类型的对象的用户更容易接受推荐结果中该类型的其他对象. 这种方法只从用户的角度出发, 考虑了用户与对象间的关注关系, 没有考虑到对象间的关系也是影响用户决策的重要因素.

文献[8]提出在基于排序学习的矩阵分解模型中融合社会关系等网络特征. 好友推荐问题是 TOP-N 问题,这类推荐一般只包含 0、1 两级评分, 对应了排序学习的正例和反例. 基于排序学习的矩阵分解模型能够很好的解决 Top-N 推荐问题, 采用排序学习的思想能够保证推荐结果排序的正确性.

文献[8]提出了一种将推荐对象间的关系作为一种社交网络特征的模型, 模型考虑了用户间的关注关系和推荐对象间的关注关系. 这种方法只考虑了用户间的直接关注关系, 没有考虑社交网络中的具有二度人脉的间接关系, 比如拥有很多共同好友的用户更容易建立好友关系.

文献[10,11]提出将信任传播机制融合到矩阵分解模型. Guo 等人[12]提出了一种信任关系强度敏感的社交网络推荐算法. 算法通过共享的潜在用户特征空间来对信任关系强度和用户兴趣进行建模, 识别出那些与目标用户有着共同爱好的朋友来对求解的过程进行优化. 以上文献用到的信任传播是根据用户间的直接信任关系推导出用户之间的间接信任关系. 但是信任关系忽视了没有路径可达的用户之间也可能相似, 比如两个用户有很多共同的关注对象, 但是他们之间没有间接或直接的关系.

研究者对社交网络进行了很多研究, 均获得了良好的推荐效果, 但是对于社交网络的探索仍有局限. 针对已有的矩阵分解算法改进策略, 我们可以得出社交网络在矩阵分解模型中发挥着重要作用. 下面的章节我们首先介绍基于排序学习的矩阵分解模型然后针对几种社交网络的结构特征进行详细分析, 提出改进的模型.

2 基本排序学习的矩阵分解

在基于矩阵分解的好友推荐系统中, 我们常使用排序学习算法对模型进行优化, 排序学习能够考虑用户对于两个物品偏好的排序关系, 这种关系对好友推荐尤其重要. 近年来一些基于排序的推荐算法被提出, Pessiot 等[13]提出在协同过滤推荐中采用 Learning To Rank 的思想进行排序学习. Rendle 等[14]提出了一种贝叶斯个性化排序算法, 直接优化 AUC. 这些排序算法均带来了良好的推荐效果且适用于好友推荐系统. 本文使用的排序算法认为如果一个用户 u 关注了用户 i, 那么可以认为 u 对 i 的喜好超过了没有被 u 关注的用户. 如果一个用户同时关注了多个用户, 那么该用

户对其关注的多个用户的喜好是相同的, 同样对其没有关注的用户, 喜好也是相同的.

基于排序学习算法我们定义了集合 $K=\{(u,i,j)|u\in U\wedge i\in U\wedge j\in U\wedge u\to i\in E\wedge u\to j\notin E\}$, $(u,i,j)\in K$, 元素(u,i,j)表示用户 u 对 i 的喜爱超过了对 j 的喜爱. 在这个集合上优化矩阵分解的损失函数, 通过迭代训练得到模型中的参数. 损失函数如公式(2)所示.

$$\min_{P,Q} L_1=\sum_{(u,i,j)\in K}-\log\left(\sigma\left(p_u{}^T q_i-p_u{}^T q_j\right)\right)+\frac{\lambda}{2}\|P\|_F^2+\frac{\lambda}{2}\|Q\|_F^2 \quad (2)$$

损失函数的值表示预测值和实际值之间的逼近程度, $P\in R^{d\times|U|}$ 表示关注者的潜在特征矩阵; $Q\in R^{d\times|U|}$ 表示推荐对象的潜在特征矩阵; σ 是逻辑斯蒂函数, $\sigma(x)=\left(1/\left(1+e^{-x}\right)\right)$, λ 是正则化项的非负参数, 可以用来防止过拟合; $\|\bullet\|_F$ 是 Frobenius 范数.

利用基于排序学习的矩阵分解模型进行好友推荐是一个线性回归问题, 为了得到模型中的参数, 我们需要局部最小化该损失函数. 随机梯度下降是最小化风险函数、损失函数的一种常用方法, 可以对线性回归问题求解. 该方法包括增量梯度下降和批量梯度下降. 增量梯度下降得到的是一个全局最优解, 但是每迭代一步, 都要用到训练集中所有的数据, 当数据量很大的时候, 时间复杂度会非常高. 随机梯度通过每个样本来迭代更新一次, 可能只需要训练集中很少的样本, 就已经将参数迭代到最优解了. 因此本文我们使用随机梯度下降法更新参数.

3 基于矩阵分解的社交网络正则化推荐

本章首先分析微博平台的社交网络的结构特征, 根据不同的结构特征选择不同的相似度方法计算用户之间的兴趣相似程度, 然后针对不同的网络结构特征分别提出了可以利用网络结构特征进行优化的矩阵分解模型.

我们先对文中的符号进行定义, 把微博中的普通用户和热点用户都统一用 U 表示. E 表示用户之间的关注关系, 社交网络则用有向图 $G(U,E)$表示. 如果用户 u 关注用户 v, 则用有向边 $u\to v\in E$ 表示关注关系. 用户 u 的出度集合记为 $out(u)=\{v\in U|u\to v\in E\}$, $|out(u)|$ 表示出度的个数, $|\bullet|$ 表示集合的大小. 同样的, $in(u)=\{v\in U|v\to u\in E\}$ 表示用户 u 的入度, $|in(u)|$ 表示入度的个数.

由于微博平台上用户的数量很大, 如果对所有好友相互推荐, 则需要一个相当大的评分矩阵, 所以本文只研究对微博的热点用户进行推荐. 热点用户和普通用户的社交网络关系如图 1 所示. 其中, U 表示普通用户, V 表示热点用户. 在微博这个平台上用户可以建立关注关系, 这种关注关系是单向的, 不需要得到目标用户的认可就能建立. 微博中的用户具有双重角色, 即一个用户可以是关注者也可以是被关注者.

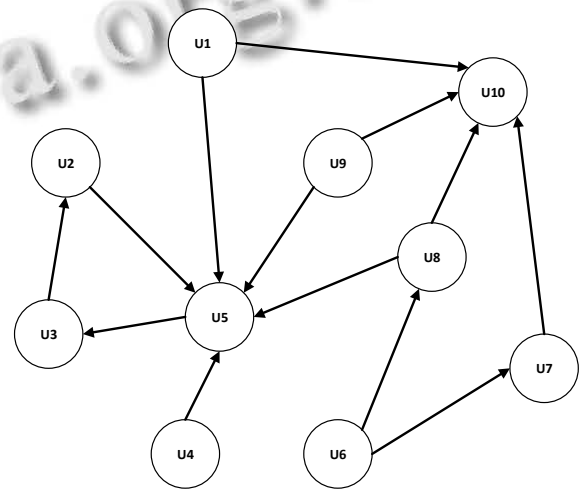
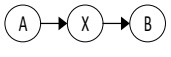
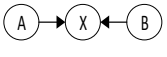
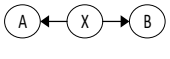


图 1 微博平台社交网络

根据对图 1 所示的社交网络图的分析, 两个没有关注关系的用户可以通过某个用户以下三种方式关联起来: 通过某个用户间接关注、共同关注某个用户和同时被某个用户关注. 根据这三种关系可以抽象出三种典型的网络结构, 如表 1 所示. 其中, Structure1 中 A 和 B 是间接关注关系, Structure2 中 A 和 B 共同关注了 X, Structure3 中 A 和 B 被 X 同时关注. 在这三种结构中, A 和 B 没有直接的关系, 但是它们通过 X 建立了间接的关系, 二者可能彼此互相感兴趣.

表 1 三种网络结构图

Structure 1	Structure 2	Structure 3
		
A → X → B	A → X ← B	A ← X → B
A following B indirectly	A and B following X	A and B followed by X

Structure1、Structure2、Structure3 这三种结构特征是社交网络中较为典型且比较直观的特征, 在本文中主要围绕三个特征进行讨论, 并根据它们建立了一个好友推荐模型.

3.1 基于社交网络结构 structure1 的模型

在 Structure1 中, 用户 A 和用户 B 是间接关注关系, B 是 A 的二度人脉. 这种关注关系可以用图 2 表示.

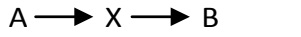


图 2 structure1 的关注关系

在这种关注关系中, 用户 A 很可能关注用户 B. 这是因为用户 A 的兴趣受 X 的影响, 用户 X 的兴趣受 B 的影响, 用户 B 间接影响 A, A 和 B 会有一些共同的特点. 在这种场景中, 求相似度时我们需要考虑用户 A 关注的对象中有多少关注了用户 B. 在这里定义了一种有向的相似度, 如公式(3)所示. 基于 Structure1 这种网络结构, 在矩阵分解模型中用户 A 的潜在特征向量 p_A 应该接近于用户 B 的潜在特征向量 p_B.

$$\mathrm{sim1}(A,B)=\frac{|out(A)\cap in(B)|}{\sqrt{|out(A)||in(B)|}} \tag{3}$$

我们定义矩阵分解模型的正则化项公式如(4)所示. 其中 u 和 f_1 的关系与图三中 A 和 B 的关注关系相同. 使用相似度函数 $sim1(u,f_1)$ 表示用户 u 和用户 $f1$ 的相似度, 用该正则化项来最小化 p_u 和 p_{f1} 之间的距离.

$$\frac{\alpha}{2}\sum_{u\in U}\sum_{f_1\in F_1(u)} sim1(u,f_1)\left\|p_u-p_{f_1}\right\|_F^2 \tag{4}$$

在这里 α 是一个基于 Structure1 的非负系数. $F_1(u)$ 表示与用户 u 的关注关系是间接的且相似度较高的用户集合. 相似度 $sim1(u,f_1)$ 的值越大意味着用户 u 和 f_1 越相似, 同时也意味着潜在的用户特征向量之间的距离越小. 把正则化项添加到损失函数中, 可以得到公式(5)所示的损失函数. 模型中的参数同样可以通过随机梯度下降算法训练得到.

$$\min_{P,Q} L_2=\sum_{(u,i,j)\in K}-\log\left(\sigma\left(p_u{}^T q_i-p_u{}^T q_j\right)\right)+\frac{\alpha}{2}\sum_{u\in U}\sum_{f_1\in F_1(u)} sim1(u,f_1)\left\|p_u-p_{f_1}\right\|_F^2+\frac{\lambda}{2}\|P\|_F^2+\frac{\lambda}{2}\|Q\|_F^2 \tag{5}$$

3.2 基于社交网络结构 structure2 的模型

在 Structure2 中用户 A 和用户 B 都关注了用户 X. 即不同的用户关注了同一个用户, 如图 3 所示.

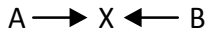


图 3 structure2 的关注关系

X 是微博中的一种信息源, 如果 A 关注了 X 说明 A 对该信息源感兴趣, 其他用户也关注了 X 就说明其他用户也有相同兴趣. 对于具有相同兴趣的用户, 他们之间的相似度也会相对更高. 在 Structure1 的情况下, 对于关注了同一个用户的用户 A 和用户 B, 我们可以用共同好友比例计算他们的相似度. $out(A)$是在社交网络图中用户 A 指向的其他好友的集合, 则计算 A 和 B 相似度的公式可以表示如下:

$$\mathrm{sim2}(A,B)=\frac{|out(A)\cap out(B)|}{\sqrt{|out(A)||out(B)|}} \tag{6}$$

我们定义矩阵分解模型的正则化项公式如(7)所示. 其中 u 和 f_2 的关系与图四中 A 和 B 的关注关系相同. 使用相似度函数 $sim2(u,f_2)$ 表示用户 u 和用户 f_2 的相似度, 用该正则化项来最小化 p_u 和 p_{f_2} 之间的距离.

$$\frac{\beta}{2}\sum_{u\in U}\sum_{f_2\in F_2(u)} sim2(u,f_2)\left\|p_u-p_{f_2}\right\|_F^2 \tag{7}$$

在这里β是一个基于 Structure2 结构的正则化项的非负系数, F_2(u) 表示和用户 u 有共同关注对象且相似度较高的用户集合. $sim2(u,f_2)$ 表示用户 u 和用户 f_2 的相似度. 把以上的正则化项添加到损失函数中, 可以得到下面的损失函数:

$$\min_{P,Q} L_3=\sum_{(u,i,j)\in K}-\log\left(\sigma\left(p_u{}^T q_i-p_u{}^T q_j\right)\right)+\frac{\beta}{2}\sum_{u\in U}\sum_{f_2\in F_2(u)} sim2(u,f_2)\left\|p_u-p_{f_2}\right\|_F^2+\frac{\lambda}{2}\|P\|_F^2+\frac{\lambda}{2}\|Q\|_F^2 \tag{8}$$

3.3 基于社交网络结构 structure3 的模型

在 Structure3 中用户 A 和用户 B 都关注了用户 X. 即同一个用户关注了两个不同的对象, 如图 4 所示.

图 4 Structure3 的关注关系

在微博中如果很多人在关注了用户 A 的时候也关注了用户 B, 那么用户 A 和用户 B 一定会有一些共同的特点吸引了这些关注他们的人. 所以我们可以根据共同关注了 A 和 B 的用户的数量来表示他们之间的相似度. $in(A)$是在社交网络图中指向用户 A 的用户的集

合. *in(B)*是在网络图中指向用户 *B* 的用户的集合. 基于 Structure3 我们使用公式(9)计算 *A* 和 *B* 的相似度.

$$\mathrm{sim3}(A,B)=\frac{\left|in(A)\cap in(B)\right|}{\sqrt{\left|in(A)\right|\left|in(B)\right|}} \tag{9}$$

我们定义矩阵分解模型的正则化项公式如(10)所示. 其中 u 和 f_3 的关系与图五中 *A* 和 *B* 的关注关系相同. F_3(i) 表示用户 u 关注的对象中, 除了 i 之外的对象集合且集合中的对象与 i 的相似度较高. 使用相似度函数 $sim3(i,f_3)$ 表示用户 u 和用户 f_3 的相似度, 用该正则化项来最小化 p_u 和 p_{f_3} 之间的距离.

$$\frac{\gamma}{2}\sum_{i\in U}\sum_{f_3\in F_3(i)} sim3(i,f_3)\left\|q_i-q_{f_3}\right\|_F^2 \tag{10}$$

在这里γ是基于 Structure3 结构的正则化项的一个非负系数, $sim3(i,f_3)$是相似度函数. 把以上的结构正则化项加入到损失函数中, 得到一个新的损失函数:

$$\begin{aligned}\min_{P,Q} L_4 &= \sum_{(u,i,j)\in K} -\log\left(\sigma\left(p_u{}^T q_i - p_u{}^T q_j\right)\right)\\ &+\frac{\gamma}{2}\sum_{i\in U}\sum_{f_3\in F_3(i)} sim3(i,f_3)\left\|q_i-q_{f_3}\right\|_F^2\\ &+\frac{\lambda}{2}\|P\|_F^2+\frac{\lambda}{2}\|Q\|_F^2\end{aligned} \tag{11}$$

3.4 基于社交网络结构特征的模型

在一个复杂网络中会包含多种网络特征, 前面我们已经详细讨论了三种简单特征对矩阵分解模型中潜在向量的影响, 下面我们将这些有利于改善推荐效果的多个特征融合到矩阵分解模型中, 从而形成一个综合的模型, 这个模型的损失函数如下:

$$\begin{aligned}\min_{P,Q} L_5 &= \sum_{(u,i,j)\in K} -\log\left(\sigma\left(p_u{}^T q_i - p_u{}^T q_j\right)\right)\\ &+\frac{\alpha}{2}\sum_{u\in U}\sum_{f_1\in F_1(u)} sim1(u,f_1)\left\|p_u-p_{f_1}\right\|_F^2\\ &+\frac{\beta}{2}\sum_{u\in U}\sum_{f_2\in F_2(u)} sim2(u,f_2)\left\|p_u-p_{f_2}\right\|_F^2\\ &+\frac{\gamma}{2}\sum_{i\in U}\sum_{f_3\in F_3(i)} sim3(i,f_3)\left\|q_i-q_{f_3}\right\|_F^2+\frac{\lambda}{2}\|P\|_F^2+\frac{\lambda}{2}\|Q\|_F^2\end{aligned} \tag{12}$$

我们仍然使用随机梯度下降算法求得潜在特征向量p_u, q_i和q_j. 对特征向量p_u, q_i和q_j求偏导的结果如下:

$$\begin{aligned}\frac{\partial L_5}{\partial p_u} &= \sum_{(u,i,j)\in K}\frac{-\left(q_i-q_j\right)}{1+e^{\left(p_u{}^T q_i - p_u{}^T q_j\right)}}\\ &+\alpha\sum_{f_1\in F_1(u)} sim1(u,f_1)\left(p_u-p_{f_1}\right)\\ &+\beta\sum_{f_2\in F_2(u)} sim2(u,f_2)\left(p_u-p_{f_2}\right)+\lambda p_u ?,\end{aligned}$$

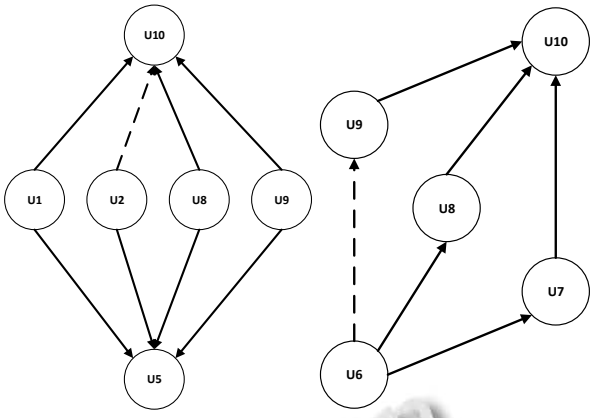
$$\begin{aligned}\frac{\partial L_5}{\partial q_i} &= \sum_{(u,i,j)\in K}\frac{-p_u}{1+e^{\left(p_u{}^T q_i - p_u{}^T q_j\right)}}\\ &+\gamma\sum_{f_3\in F_3(i)} sim3(i,f_3)\left(q_i-q_{f_3}\right)+\lambda q_i ?\end{aligned} \tag{13}$$

$$\frac{\partial L_5}{\partial q_j} = \sum_{(u,i,j)\in K}\frac{p_u}{1+e^{\left(p_u{}^T q_i - p_u{}^T q_j\right)}}+\lambda q_j$$

这个统一的模型中融合了三种社交网络结构化的信息, 这些结构化信息有助于构造更好的用户潜在模型, 从而产生更加准确的推荐结果.

3.5 融合更多社交网络结构特征

在实际的社交网络中, 存在着各种各样的结构特征, 也往往会遇到多种网络结构特征并存的问题. 例如, 可以从图 1 所示的社交网络图中抽取出更复杂的结构特征, 如图 5 所示.



(a) 复杂结构 1 (b) 复杂结构 2

图 5 更复杂的社交网络结构特征

在图5(a)中, 用户 *U1*, *U8*, *U9* 同时关注了用户 *U5* 和用户 *U10*, 而 *U2* 也关注了 *U5*, 所以 *U2* 可能也会对 *U10* 感兴趣; 图5(b)中, 用户 *U6* 关注的 *U7*, *U8* 都关注了用户 *U10*, 而 *U9* 也关注了 *U10*, 因此 *U6* 可能也会关注 *U9*. 在这两种结构中我们用 u 表示目标用户(*U2* 和 *U6*), v 表示目标用户可能兴趣用户(*U9* 和 *U10*). 那么可以通过这两种复杂的结构特征, 首先找到用户 u 的可能感兴趣的用户 v 的集合, 然后选择合适的相似度计算方法求得 u 和 v 之间的相似度, 最后再定义新的正则化项来最小化它们的潜在向量之间的距离. 从而可以将该正则化项加入到最终的损失函数中, 会进一步提升推荐结果的准确度. 该损失函数的形式和求解方式和 3.4 节中的模型类似, 在这里就不多加讨论了.

4 实验及结果分析

4.1 数据集

本文使用腾讯微博的数据集, 该数据集在 KDD Cup 2012 Track1 中被使用. 数据集包含七个文件. 本实验使用了其中的四个文件. 数据规模如表 2 所示.

由于训练集中正负样本的比例为 1:12.9, 需要进行预处理. 本文抽取出数目相同的正样本和负样本, 由正负样本形成的二元组适用于基于排序学习的矩阵分解模型.

表 2 腾讯微博数据集规模

文件名称	文件大小
rec_log_train.txt	1.99Gb
rec_log_test.txt	943Mb
item.txt	1.18Mb
user_sns.txt	740Mb

在实验中为了减少计算时间, 我们随机从 user_profile.txt 文件中选取了其中的 100000 个用户, 然后从 user_sns.txt 文件中抽取了这 100000 个用户的 2,037,284 条历史关注记录, 我们主要根据用户的关注记录计算用户之间的相似度. 然后我们再把与这些用户相关的数据从 rec_log_train.txt 中抽取出来作为实验的训练集, 从 rec_log_test.txt 中抽取出和这些用户相关的数据作为实验数据集.

4.2 测量指标

推荐系统的测量标准是衡量推荐结果效果的标尺, Zhu 等人[15]总结了推荐系统评价指标的最新研究进展. 其中 MAP 得到了很多推荐系统研究者的广泛采用[16]. 本文所使用的评价指标是 MAP(Mean Average Precision)和 AP@n. AP@n 和 MAP 能够更全面的对 TOP-N 推荐结果进行评价. AP@n 和 MAP@n 的定义如下:

$$\mathrm{AP@n}=\sum_{k=1}^{n}P(k)/m \tag{14}$$

$$\mathrm{MAP@n}=\sum_{i=1}^{N}AP@n_i/N \tag{15}$$

其中 m 为用户实际接受的推荐结果总数, AP@n 是含有 n 个对象的单个推荐列表中准确率的平均值. MAP@n 是所有推荐列表的的平均准确率的平均值. 被接受的推荐对象在推荐列表中越靠前, MAP 就可能越高.

4.3 结果及分析

我们进行了三组实验, 从不同方面验证算法的有效性. 试验中 α,β,γ 的值均设置为 1, 矩阵分解模型的学习速率设置为 0.0005, 正则化参数 λ 设置为 0.002, 用户的潜在因子数设置为 150. 虽然潜在因子个数越多, 预测越准确但是计算量也会随之变大, 150 是我们选择的较为理想的值.

4.3.1 各类算法的比较

本节实验对各类算法进行比较. 参与比较的算法有基于用户的协同过滤(User Based Collaborative Filtering)、基于物品的协同过滤(Item Based Collaborative Filtering)、基于排序学习的矩阵分解模型(Ranking Model)、带有 Structure1 特征的模型、带有 Structure2 特征的模型、带有 Structure3 特征的模型和最终改进模型(Unified Model). 实验结果如图 6 所示.

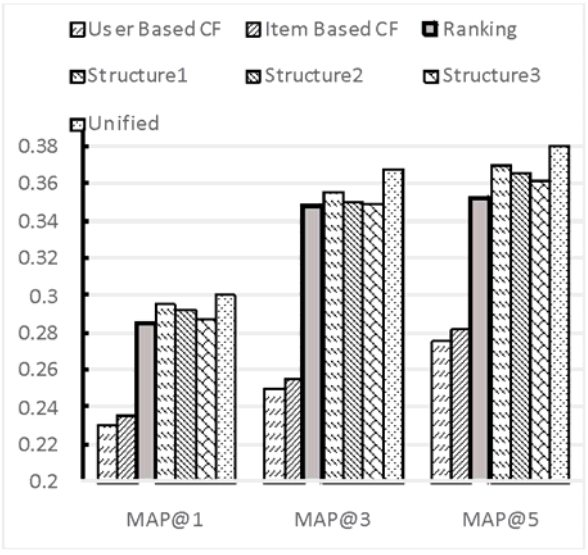


图 6 不同推荐模型推荐准确度比较

由实验结果显示了加入了结构特征的模型和基于评分的协同过滤的实验结果对比, 前者明显优于后者. 加入了三种特征的模型比加入了一种特征的模型效果要好. 基于评分的协同过滤依赖, 存在评分数据稀疏的问题, 影响了结果准确率. 改进的模型利用了评分和社交网络特征得到了更好的效果, 由此可以验证网络结构特征能够帮助我们提高推荐效果.

4.3.2 不同邻居个数对结果的影响

本组实验对比不同的邻居个数对最终改进模型的推荐结果的影响, 实验结果如图 7 所示, 由实验结果可以看出, 当邻居个数为 150 的时候, 结果的 MAP 最好, 当邻居个数增多的时候 MAP 呈下降趋势.

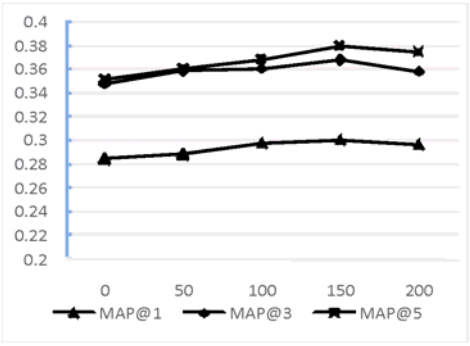


图 7 不同的最近邻个数对推荐结果的影响

实验结果表明, 最近邻的个数会影响推荐的准确度. 这是因为当最近邻个数太少时, 相应的正则化项对目标用户的潜在因子向量影响小, 从而准确度的提升也较小. 当最近邻个数太多时, 相似度不高的用户会对目标用户的潜在因子向量产生干扰, 也会影响准确度的提升.

4.3.3 不同的相似度方法对推荐结果的影响

在社交网络中常采用基于共同好友的相似度计算方法, 下面分别介绍两种常用的相似度方法:

1) Jaccard Coefficient:

$$sim(\mathrm{A},\mathrm{B})=\frac{|out(\mathrm{A})\cap \mathrm{out}(\mathrm{B})|}{|out(\mathrm{A})\cup \mathrm{out}(\mathrm{B})|} \tag{16}$$

2) Cosine 相似度

$$sim(\mathrm{A},\mathrm{B})=\frac{|out(\mathrm{A})\cap \mathrm{out}(\mathrm{B})|}{\sqrt{|out(\mathrm{A})||out(\mathrm{B})|}} \tag{17}$$

以上两种算法均根据用户的共同好友求相似度. 本文我们对于不同的网络结构使用了不同的计算方法方法. 本组实验把公式(16)(17)分别代入我们的改进模型进行结果验证. 实验结果如图 7 所示.

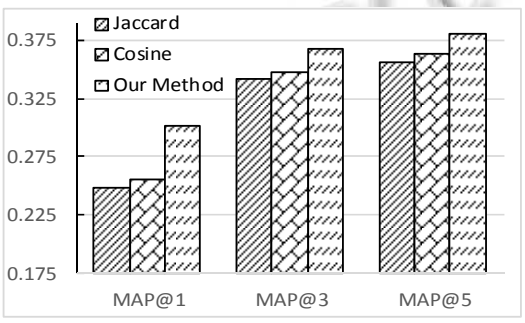


图 8 不同的相似度方法对推荐结果的影响

5 结语

本文为了帮助微博平台上的用户扩展好友圈, 提出了基于矩阵分解的社交网正则化推荐模型. 新型能够利用社交网络中用户的近邻进行建模, 从而得到了更加准确的推荐结果. 尽管本文中的推荐模型能够带来良好的效果但是该模型仍然有需要继续改进的地方. 社交网络中存在着很多种结构特征, 本文只讨论了其中三种比较直观典型的结构特征, 由于篇幅有限, 没有能够涉及其余的一些结构特征. 因此, 如何将更多的社交网络结构特征合理的融合到一起, 从而进一步提高推荐结果的平均准确度, 还需要更加深入的研究.

参考文献

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

2 Li Q, Yao M, Yang J, et al. Genetic algorithm and graph theory based matrix factorization method for online friend recommendation. The Scientific World Journal, 2014, Article ID 162148.

3 Shi Y, Larson M, Hanjalic A. Exploiting user similarity based on rated-item pools for improved user-based collaborative filtering. Proc. of the 3rd ACM conference on Recommender systems. ACM. 2009. 125–132.

4 Ma H, King I, Lyu MR. Effective missing data prediction for collaborative filtering. Proc. of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM. 2007. 39–46.

5 Yuan Q, Chen L, Zhao S. Factorization vs. regularization: fusing heterogeneous social relationships in top-n recommendation. Proc. of the Fifth ACM Conference on Recommender Systems. ACM. 2011. 245–252.

6 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.

7 Chen T, Tang L, Liu Q, et al. Combining factorization model and additive forest for collaborative followee recommendation. KDD-Cup Workshop. 2012.

8 陈渊.基于社交网络的 Top–N 推荐问题研究[硕士学位论文].哈尔滨:哈尔滨工业大学,2013.

9 郭磊,马军,陈竹敏.一种结合推荐对象间关联关系的社会化推荐算法.计算机学报,2014,37(1):219–228.

10 Jamali M, Ester M. A matrix factorization technique with

trust propagation for recommendation in social networks. Proc. of the Fourth ACM Conference on Recommender Systems. ACM. 2010. 135–142.

11 于洪涛,周静,张付志.融合信任传播和矩阵分解的协同推荐算法.燕山大学学报,2013,37(5):424–429.

12 郭磊,马军,陈竹敏.一种信任关系强度敏感的社会化推荐算法.计算机研究与发展,2013,50(9):1805–1813.

13 Pessiot JF, Truong V, Usunier N, Gallinari P. Learning to rank for collaborative filtering. ACM Trans. on Information Systems. 2007.

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

15 朱郁筱,吕琳媛.推荐系统评价指标综述.电子科技大学学报,2012,41(2):164–175.

16 孙建凯.面向排序的个性化推荐算法研究与实现[硕士学位论文].济南:山东大学,2014.