



基于用户兴趣和推荐信任域的微博推荐

徐雅斌^{1,2}, 刘超¹, 武装¹

(1. 北京信息科技大学计算机学院 北京 100101;

2. 网络文化与数字传播北京市重点实验室 北京 100101)

摘要:向用户推荐其感兴趣的微博,是改善用户体验的重要途径。为使推荐的微博更加符合用户的兴趣和品味,提出的微博推荐方法不仅考虑用户自身的特点,而且还考虑所在社区对微博的评价。在技术实现上,采用支持向量机进行文本分类,以便发现用户的兴趣偏好;通过多维 Newman 算法进行用户社区的发现,并将社区视为推荐信任域。最后采用改进的协同过滤算法综合用户兴趣偏好和推荐信任域进行微博推荐,以此提高微博推荐的质量。实验结果表明,提出的微博推荐方法是切实有效的。

关键词:微博推荐;协同过滤;推荐信任域;用户兴趣发现

doi: 10.11959/j.issn.1000-0801.2015042

Micro-Blog Recommendation Based on User Interests and Recommendation Trust Domain

Xu Yabin^{1,2}, Liu Chao¹, Wu Zhuang¹

(1.School of Computer, Beijing Information Science & Technology University, Beijing 100101, China;

2. Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing 100101, China)

Abstract: Recommending micro-blogs to users whom are interested in is an important way to improve the user's experience. In order to make the recommended micro-blog more match user's interests and tastes, a micro-blog recommendation method was proposed. The method considers not only user's own characteristics, but also the evaluation from the user's community. In technology, vector machine (SVM) was supported for text classification to discover user interests, and through multidimensional Newman algorithm to discover user's community. This community will be regarded as the recommendation trust domain. Finally, the improved collaborative filtering algorithm was used. It integrated user's interests and recommendation trust domain to recommend micro-blog, in order to improve the quality of micro-blog recommendation. The experimental results show that, the micro-blog recommendation method is practical and effective.

Key words: micro-blog recommendation, collaborative filtering, recommendation trust domain, user interest discovery

收稿日期:2014-11-15;修回日期:2015-01-09

基金项目:国家自然科学基金资助项目(No.61370139),网络文化与数字传播北京市重点实验室基金资助项目(No.ICDD201309),北京市属高等学校创新团队建设与教师职业发展计划基金资助项目(No.IDHT20130519)

Foundation Items: The National Natural Science Foundation of China (No.61370139), Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (No.ICDD201309), The Project of Construction of Innovative Teams and Teacher Career Development for Universities and Colleges Under Beijing Municipality(No. IDHT20130519)



1 引言

随着互联网和移动终端设备的飞速发展,社交网络服务(social network service)已经成为人们日常生活中的重要组成部分。在国外,著名的社交网站主要有 Facebook、Twitter、Flickr 等;在国内,著名的社交网站主要有新浪微博、开心网、人人网等。

在微博社交网络中,用户获取信息的方式主要有 3 种:一是通过关注用户的方式来获取其他用户发布的微博,用户通过这种方式只能够看到其关注用户公开发布的所有微博,不能根据自身兴趣偏好有选择地获取自己感兴趣的微博;二是按照一定的条件对微博进行检索,但是这种方式往往产生大量的返回结果,导致用户无法进行选择;三是由微博平台的运营者推荐热门话题相关的微博,但是热门话题只针对所有微博用户,并不考虑每个用户的需求。因此,现有的微博获取方式并不能够实现微博的个性化推荐。而且,由于微博数量巨大,更加使得用户无法及时、有效获取自己感兴趣的微博。因而,如何从海量的微博数据中挖掘出用户感兴趣的微博并精准地推荐给目标用户,已经成为微博推荐相关研究的热点。

2 相关工作

为了实现微博内容的个性化推荐,使推荐结果更加准确。一些学者对微博、Twitter、Facebook 等知名社交网络上的个性化推荐进行了深入研究,提出了一些个性化推荐模型和算法。

Earlybird^[1]是运用在 Twitter 上的一种微博搜索引擎,它能够根据微博的检索结果按照微博发布的时间及其受欢迎的程度进行排序。参考文献[2]中,Chen 等人也对 Twitter 上的微博检索方法进行了研究,研究中利用每条微博成为检索结果的概率对微博进行分组。参考文献[3]中,Duan 等人利用 learning to rank(排序学习)框架实现微博推荐,该框架中选择了微博相似度、微博质量、微博所有者影响力 3 个特征来进行微博推荐结果的排序。这 3 种方法虽然在实际应用中能够高效地向用户推荐微博,但是显然不能够满足用户的个性化需求。参考文献[4]中,Chen 等人将用户的微博主题因子与待推荐微博的内容特征结合,并综合用户的影响力及待推荐微博的受欢迎程度,向用户提供个性化的推荐结果。这种方法考虑了用户的待推荐微博所属的主题与用户之间的联系,但是当用户的历史数据较少时,

推荐效果并不理想。参考文献[5]中,Chirita 等人提出的微博个性化推荐方法依赖于由用户自行输入的兴趣关键词,这种方法可以使推荐系统准确地捕捉用户的兴趣。参考文献[6]中,Ramage 利用 LDA 算法将微博映射到相应的主题空间,以改善推荐的效果。参考文献[7]中,Naveed 等人使用提取微博文本特征的方式来分析微博内容,从而预测该微博被转发的概率。参考文献[8]中,Kapanipathi 等人通过建立微博语义网络,将待推荐微博与用户的个性化信息进行匹配,预测用户是否对该微博感兴趣。

综上所述,现有的个性化微博推荐方法侧重于对用户兴趣以及微博内容和属性进行分析,而忽视了社交网络用户之间的潜在联系在个性化微博推荐中的作用。为了将这种联系合理地应用到个性化微博推荐中,本文在推荐时综合考虑了用户自身的兴趣偏好与同一网络社区用户对微博的评价,提出了一种基于协同过滤技术的微博推荐方法。该方法首先利用文本分类的方法将用户已经发布的微博按照主题类别进行分类,以此来发现用户的兴趣集合;为了解决数据稀疏性给推荐结果带来的影响,引入推荐信任域的概念,通过多维 Newman 算法发现用户所属的社区,并将其作为推荐信任域;在此基础上,本文结合兴趣偏移量及推荐信任域对传统的协同过滤算法进行改进,以此减小数据稀疏性给推荐结果带来的影响,使同一推荐信任域内的用户对不同类别微博的评分在推荐过程中拥有不同的可信度,从而降低由于用户对不同类别微博的熟悉程度不同而给推荐结果带来的误差;在微博内容推荐时,利用线性回归模型将用户兴趣和同一社区用户对微博的综合评价有效结合起来,以改善推荐效果。

3 基于文本分类的微博用户兴趣发现

3.1 微博文本分类

用户评论、转发及发布的微博中通常蕴含着用户的兴趣,用户微博所属的主题类别往往可以反映该用户的兴趣倾向,所以采用微博分类的方法可以有效地判断微博所属的主题类别,进而发现用户的兴趣倾向。

微博分类是一种属于有监督学习的短文本分类,本文采用多分类支持向量机(SVM)作为文本分类模型。SVM 原是一种二分类模型,但是为了解决多分类的问题,研究人员又在二分类的基础上进一步加以改进,形成了基于 1-a-1 的多分类 SVM。该方法的基本思想是:在每 2 个类别之间构造一个二分类 SVM 分类器,共有 $k(k-1)/2$ 个分

类器。为文本分类时,在这些分类器中出现次数最多的类别即为文本所属的类别。在训练分类模型之前,要首先建立微博主题类别体系,该分类体系既要尽可能涉及全部的主题,各类别之间的区分度又要尽可能明显。本文根据新浪微博及各门户网站提供的主题类别体系设置了 8 个主题类别,具体类别名称见表 1。

表 1 微博主题类别体系

主题类别	
体育	娱乐
科技	财经
军事	时尚
旅游	社会

此外,在使用 SVM 进行微博分类时,需要将每条微博用向量空间模型表示,即用一个以特征项权重作为分量的向量 $\vec{d}$ 表示一条微博。因为文本分类所使用的词汇表中大约有十几万个词语,所以会使特征向量 $\vec{d}$ 成为一个十几万维的高维向量。又因为微博属于短文本,其长度不超过 140 字,所以通常只能提取出很少量的能够表示微博的特征词语。在这种情况下,向量 $\vec{d}$ 中的元素几乎全部为 0,使得用来表示微博的特征向量之间的区分度极低,这非常不利于分类器的训练,并且严重影响分类的准确率。因此,本文采用卡方检验的方法从所有训练文本中提取出区分度最高的 k 个词语替代原有的词汇表,使得 $\vec{d}$ 的维度可以降低到几千甚至几百,从而提高了向量的区分度,进而提高分类的准确率。

3.2 用户兴趣发现

假设含有 m 个类别的微博主题类别集合为 $C=\{c_1, c_2, \dots, c_m\}$, 对于某个用户 u , 收集其发布、评论或转发的微博文本集合 $T=\{t_1, t_2, \dots, t_n\}$, 微博文本数目为 n , 经过分类器的预测后, 得到这些文本所对应的类别列表 $P=\{p_1, p_2, \dots, p_i\}$, 其中 $p_i \in C$ 。用户 u 对于微博主题类别 c_i 的兴趣度 $D(u, c_i)$ 为:

$$D(u, c_i) = \frac{n_c}{n} \quad (1)$$

其中, n 为用户 u 的微博总数, n_c 为用户 u 属于类别 c_i 的微博的数量。然后将用户 u 所对应的每个类别的 $D(u, c_i)$ 值经过降序排序后, 选择最靠前的 k 个类别作为用户 u 的兴趣集合。

4 推荐信任域及其发现方法

4.1 推荐信任域

在微博社交网络中, 用户数量、微博数量都很庞大, 一

个用户发布的微博通常只是全部微博的极小一部分, 这导致在使用传统协同过滤算法进行微博推荐时, “用户-项目”矩阵极其稀疏, 无法准确计算用户之间相似度, 从而无法准确地找到与目标用户相似度较大的用户, 以至于影响协同过滤算法的效果。解决这一问题的关键在于如何找到一个大小合适的用户集合, 并保证该集合中的用户拥有一定的相似度。本文从微博推荐系统的角度出发, 提出了微博推荐信任域的概念。

定义 1 对于微博用户集合 U , 用户 $u \in U$, 如果能够根据集合 U 中的其他用户信息, 准确预测用户 u 对其未知微博的喜爱程度, 则称集合 U 为用户 u 的推荐信任域。

推荐信任域是一个有共同兴趣且相互之间联系紧密的微博用户集合, 微博社交网络中信息的传播基本上完全依赖于用户之间的关注关系, 即用户所能够看到的微博绝大多数来自于其关注的用户, 所以同属于一个推荐信任域的用户由于关注关系密集而使得这些用户之间能够接收到很多相同的微博, 这就从一定程度上减少了数据稀疏性对推荐结果的影响。

推荐信任域的发现是一个寻找相似用户集合的过程, 参考文献[9]利用社会图谱、个人信息及交互信息对用户进行聚类, 进而产生相似用户集合。参考文献[10]利用现有的用户之间的联系预测潜在的用户关系, 计算用户邻居节点集合的相似度, 生成相似用户集合。Newman 在参考文献[11]中提出了“社区”的概念: “社区”是社交网络中一部分个体由于拥有共同的兴趣偏好或背景而形成的群体。社区内部的用户通常会共同关注某些话题, 联系紧密, 而社区间的用户联系不够紧密且很少有共同的话题。为了准确地发现每个用户所属的推荐信任域, 本文根据社区的概念将社区作为一个推荐信任域, 并以 Newman 快速(NF)算法的思想为基础, 提出了一种多维 NF 算法, 用于发现推荐信任域。

4.2 多维 NF 算法

Newman 快速算法^[12]是一种根据经典 GN 算法^[13]提出的基于贪心思想的凝聚式层次聚类社区发现算法。其主要作用是能够根据网络中各个节点之间联系的紧密程度将整个网络划分成若干个社区。在无权网络中, 这种算法不断沿着模块度增量 ΔQ 最大的方向合并网络中的节点或子社区, 以形成一定的社区结构, 直到所有节点合并为一个社区, 然后找出模块度 Q 最大的一次合并的结果, 并将其作为最终的社区发现结果。模块度 Q 是用来衡量社区发



现质量的标准, Q 值越大, 说明社区发现的效果越理想。

在传统社区发现中, 用户之间的相互关注关系可以看作是用户之间的一种联系, 所以可以将用户作为节点, 用户之间的相互关注作为边, 将整个网络中的用户关系抽象成一个无权网络图。但是在微博社交网络中, 用户的属性、标签及微博也蕴含着大量的信息, 这些信息往往能够体现用户之间兴趣上的相似程度, 对于社区发现有着重要作用。如果只考虑用户之间的关注关系而不考虑其他信息, 则可能导致这种单一维度的 NF 算法效果不佳。推荐信任域是由一些拥有共同兴趣爱好且相互间有一定关注关系的用户组成的, 如果只凭借微博社交网络的拓扑结构信息并不能保证社区内用户兴趣的相似性。因此, 本文提出一种加权网络下的多维 NF 算法来实现推荐信任域的发现。首先, 将微博社交网络抽象成一个加权网络, 网络的边权重为一条边连接的两个用户节点的兴趣相似度, 其计算方法如式(2)所示。

$$p_{ab} = \frac{I_a \cap I_b}{k} \quad (2)$$

其中, $I_a \cap I_b$ 为用户 u_a 和 u_b 兴趣集合中相同主题类别的数量, k 为每个用户兴趣集合中主题类别的数量。传统无权网络社区发现中的模块度增量 ΔQ 及模块度 Q 并不适用于加权网络, 所以式(3)和式(4)定义了推荐信任域发现的 ΔQ 和 Q 。计算 ΔQ 和 Q 需要先定义推荐信任域关系系数矩阵 $E=(e_{ij})$ 。

$$e_{ij} = \frac{\sum_{a \in i} p_{ab}}{\sum_{b \in C} p_{ab}} \quad (3)$$

其中, 矩阵元素 e_{ij} 表示推荐信任域 i 和推荐信任域 j 之间的带权边的权重和占网络中所有边的权重和的比例; a 和 b 为网络中的节点, C 为网络中所有推荐信任域的集合。模块度增量 ΔQ 的定义为:

$$\Delta Q = e_{ij} + e_{ji} - 2a_i a_j \quad (4)$$

其中 $a_i = \frac{\sum_{a \in i} p_{ab}}{\sum_{b \in C} p_{ab}}$, 它表示与推荐信任域 i 相连的边

的权重和占网络中所有边的权重和的比例。模块度 Q 的定义为:

$$Q = \frac{1}{\sum_{a \in C} p_{ab}} \sum_{b \in C} \left[p_{ab} - \frac{\sum_{a \in C} p_{ab} \sum_{b \in C} p_{ab}}{\sum_{b \in C} p_{ab}} \right] \delta(c_a, c_b) \quad (5)$$

其中, p_a 为与节点 a 相连的所有连边权重之和, c_a 为节点 a 所属的推荐信任域, 当 a 和 b 两个节点同属于一个推荐信任域, 则 $\delta(c_a, c_b)=1$; 否则 $\delta(c_a, c_b)=0$ 。

根据上述分析, 基于用户兴趣的微博推荐信任域发现方法具体描述如下。

输入: 经预处理后的加权微博社交网络 G 。

输出: 推荐信任域集合 C 及每个推荐信任域内包含的节点列表 S_c 。

(1) 将网络中的每个节点初始化为一个推荐信任域;

(2) 预先计算每两个推荐信任域合并后的模块度增量 ΔQ 的值;

(3) 沿着 ΔQ 最大的方向进行推荐信任域的合并, 合并后重新计算矩阵 $E=(e_{ij})$, 同时计算当前状态下的模块度 Q ;

(4) 重复步骤(2)和(3)直到所有节点都属于一个推荐信任域, 并选择最大的 Q 值对应的划分状态为最优推荐信任域发现结果。

5 基于改进协同过滤算法的微博推荐

5.1 改进的协同过滤算法

协同过滤是利用相似用户信息实现个性化推荐最为成功的算法。协同过滤算法大致又分为基于用户的协同过滤算法^[14,15]和基于项目的协同过滤算法^[16,17]。由于用户的增长速度比微博增长速度慢, 用户之间的关系比较稳定, 计算用户间的相似度相对容易, 所以在微博推荐过程中使用基于用户的协同过滤更为合适。

为了解决数据的稀疏性, 同时根据推荐信任域的特点, 本文以用户 u 所属的推荐信任域内与 u 的相似度最大的 k 个用户作为协同过滤的邻近用户集合 U 。此外, 信任域内的每个用户对于每个主题类别的熟悉程度可能并不完全相同, 用户对其感兴趣的类别的微博通常比较熟悉, 而对于其他类别的微博可能并不熟悉, 所以对感兴趣的类别的评分可信度比较高。

基于上述分析, 本文对传统的协同过滤方法进行了改进, 使邻近用户对不同类别微博的评分在推荐过程中拥有不同的可信度, 因而降低了由于用户对不同类别的微博熟悉程度不同而给推荐结果带来的误差。式(6)给出了改进后的协同过滤式:

$$R(u_a, I_j) = \overline{R(u_a)} + \frac{\sum_{u_i \in U} (1 + \Delta_{u_i, C}) \times \text{sim}(u_a, u_i) \times (R_{i,j} - \overline{R_i})}{\sum_{u_i \in U} (1 + \Delta_{u_i, C}) \times \text{sim}(u_a, u_i)} \quad (6)$$

其中, $\overline{R(u_a)}$ 表示用户 u_a 对其他微博评分的平均值, $R(u_a, I_j)$ 表示用户 u_a 对微博 I_j 的预测评分值, $\text{sim}(u_a, u_i)$ 是用户 u_a 和 u_i 的相似度, $\overline{R_i}$ 为用户 u_i 对所有微博的平均评分。 $\Delta_{u_i, C}$ 为用户 u_i 对分类 C 的兴趣偏移量。兴趣偏移量可以反映出用户对属于某一主题分类的微博的熟悉程度, 进而反映出用户对该类别微博评价的可信度。兴趣偏移量 $\Delta_{u_i, C}$ 为:

$$\Delta_{u_i, C} = D(u_i, C) - \overline{D_C} \quad (7)$$

其中, $\overline{D_C}$ 为用户集合 U 内的用户对类别 C 的平均兴趣度。

5.2 微博推荐方法

在预测一条微博 t 是否应该被推荐给用户 u 时, 主要考虑如下两个因素:

- 微博 t 所对应的主题类别是否符合用户的兴趣;
- 与用户 u 联系较为紧密或兴趣相似的同一网络社区用户对微博 t 的评价。

本文采用结合兴趣偏移量的协同过滤算法来计算邻近用户对微博的总体评分。由于微博平台中并没有用户对微博内容的直接评分, 所以本文首先将用户对微博的转发和评论作为用户对微博内容的隐性评分来确定用户对微博的喜欢程度。微博的转发是信息传播的主要途径之一, 表示用户喜欢该微博, 所以评分应该最高, 其次应该是用户评论微博, 既不转发也不评论表明用户对该微博并不感兴趣。为了方便计算, 将这种隐性评分进行量化: 上述 3 种情况的评分依次定为 5、3、1。

最终的预测结果是由微博主题类别和邻近用户评分两个因素所决定。基于用户兴趣和推荐信任域的微博推荐方法, 简称 RBTD 方法。采用 RBTD 方法计算预测结果的方法如式(8)所示。

$$R_f = R(u_a, I_j) \times w_r + D(u_a, c_i) \times M \times w_d \quad (8)$$

其中, M 为用户对微博评分的上限值, w_r 和 w_d 分别为两部分得分的权重。权重 w_r 和 w_d 的取值直接影响着推荐评分的计算。为了更准确地估计 w_r 和 w_d 的取值, 本文将式(8)作为一个线性回归模型。线性回归是利用数理统计中的回归分析来确定两种及以上变量相互依赖的定量关系的一种统计方法。本文中, 线性回归的因变量为 $y=R_f$, 两个自变量分别为 $x_1=R(u_a, I_j)$ 和 $x_2=D(u_a, c_i) \times M$, 自变量所对应的系数分别为 $\beta_1=w_r$ 和 $\beta_2=w_d$, 假设共有 n 组观测值, 则可由式(8)推导出线性回归表达式为:

$$\begin{cases} y_1 = \beta_1 x_{11} + \beta_2 x_{12} \\ y_2 = \beta_1 x_{21} + \beta_2 x_{22} \\ \vdots \\ y_n = \beta_1 x_{n1} + \beta_2 x_{n2} \end{cases} \quad (9)$$

将各组观测值代入线性回归模型中, 并引入损失函数(式(10)), 求解出损失函数取得最小值时对应的 β_1 和 β_2 值, 即为式(8)中的 w_r 和 w_d 。

$$L(y, f(x_1, x_2)) = (y - f(x_1, x_2))^2 \quad (10)$$

6 实验与结果分析

6.1 实验数据

本文实验数据有两个来源:

- 利用新浪微博爬虫采集 10 万条用户微博内容作为数据集 D ;
- 利用新浪微博爬虫采集的 2 000 个相互邻近的微博用户信息作为数据集 S , 其中包括 2012 年至今的微博(包含评论和转发信息)共 10 万余条以及每个用户的关注用户列表。

数据集 D 的作用是训练和测试 SVM 微博文本分类器, 而数据集 S 则再分为两个部分: S 中的 1 000 个用户节点信息作为一个子数据集 S_1 , 用于计算式(8)中的权重值 w_r 和 w_d ; 其余 1 000 个用户节点信息作为数据集 S_2 , 用于发现微博推荐信任域及推荐结果的测试。

6.2 实验步骤

步骤 1 本实验首先要训练微博文本分类器以实现微博用户兴趣发现。数据集 D 中 80% 的微博内容作为 SVM 微博文本分类器的训练数据, 训练出文本分类器, 而另外 20% 数据则作为文本分类器的测试数据集。实验表明, 通过训练得到的微博文本分类器的准确率为 81.59%, 所以该分类器拥有较好的分类效果, 并可以用于用户的兴趣集合发现。

步骤 2 式(8)是一个多元线性回归模型, 为了估计其中的权重 w_r 和 w_d , 首先将数据集 S_1 中的用户划分为 n 个推荐信任域, 在每个推荐信任域内随机抽取 m 个用户; 之后取出每个用户曾经评论和转发的微博作为训练集; 然后用本文方法计算出每条微博的 $R(u_a, I_j)$ 和 $D(u_a, c_i)$; 最后利用线性回归模型通过迭代得出权重值为 $w_r=0.339\ 5$, $w_d=0.670\ 5$ 。

步骤 3 为了验证 RBTD 算法的有效性和准确性, 实验中对基于协同过滤的微博推荐(RBCF)算法、基于传统



协同过滤和用户兴趣的微博推荐(RBICF)算法、基于用户兴趣和推荐信任域的微博推荐(RBITD)方法几种推荐方法在数据集 S_2 上的推荐结果的平均绝对误差(MAE)值、准确率(precision)及召回率(recall)3个指标进行对比。

MAE是预测值与真实值的偏差绝对值和的平均值,其计算方法为:

$$MAE = \frac{\sum_{i=1}^N |p_i - q_i|}{N} \quad (11)$$

其中, N 表示预测评分的个数, p_i 表示微博 i 的预测评分, q_i 表示微博 i 的实际评分。

准确率是推荐结果中有效推荐占有所有推荐的比例,能够反映出推荐算法的有效性,其计算方法如式(12)所示。

$$precision = \frac{N_s}{N_r} \quad (12)$$

其中, N_s 为推荐结果中用户感兴趣的微博(即测试集中评论或转发过的微博)数量, N_r 为推荐的结果数量。

召回率,又称作查全率,是推荐结果中用户感兴趣的微博占其所能接收到的微博(即其好友发布的微博)的比例,其计算方法如式(13)所示。

$$recall = \frac{N_s}{N_r} \quad (13)$$

其中, N_s 为推荐结果中用户感兴趣的微博(即测试集中评论或转发过的微博)数量, N_r 为用户所能接收到的所有微博中其感兴趣的微博数量。

6.3 实验结果及其分析

实验中利用多维 NF 算法对数据集 S_2 进行推荐信任域的划分,共得到 52 个推荐信任域,图 1 为推荐信任域规模统计。由图 1 可知,大多数推荐信任域的规模(即用户数量)都集中在 15~30。这说明多数用户都可以在其所属的推荐信任域内找到 15~30 个用户作为其邻近用户的候选集合。其中,推荐信任域的规模与协同过滤中邻近用户数

量的设定有着直接的关系。

为了对比 3 种算法在不同规模的微博社交网络中的推荐效果,本文使用不同的数据集进行对比实验,此时根据推荐信任域的平均规模将目标用户的邻近用户的数量设置为 20。图 2 为 3 种推荐算法在不同用户数量的数据集上的 MAE 值的比较。从图 2 中可以看出 RBITD 方法的 MAE 值比 RBICF 算法及 RBCF 算法的 MAE 值低,而且从整体上看, RBICF 算法的 MAE 值又比 RBCF 算法的 MAE 值低,这说明 RBICF 算法的推荐效果最好,而 RBICF 算法中引入了对用户兴趣,使得推荐的结果中排除了那些不符合用户兴趣的微博,所以推荐效果也比较好。另外,从图 2 中还可以看出, RBITD 方法的 MAE 值随着微博数量的增加基本没有变化,而另外两种算法的 MAE 值则随着用户数量的增加不断增大。这是因为用户不断增多使数据变得越来越稀疏,对推荐的效果有较大的影响,而 RBITD 方法由于是利用多维 NF 算法找出用户所属的推荐信任域并从中选择出协同过滤中的邻近用户集,即使用户数量增加,推荐信任域的结构也基本稳定,所以 RBITD 方法对于用户数量变化的抗干扰性更强,因此也更适合在较大的数据集上使用。

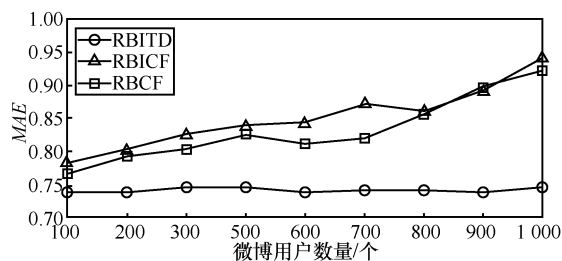


图2 微博推荐算法在不同用户数量数据集上 MAE 对比

为了测试 3 种算法在协同过滤的邻近用户数量不同的情况下的推荐效果,实验中分别设置不同的邻近用户数量并计算相应的 MAE 值,图 3 为 3 种算法 MAE 值的比较(用户数量为 1 000 户)。从图 3 中可以看出, RBITD 方法

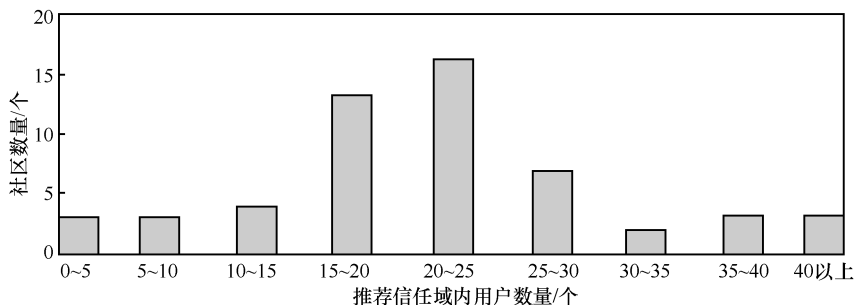


图1 推荐信任域规模统计

的推荐效果要明显好于另外两种算法。但是通过具体分析可以发现,邻近用户数量从 5 个逐渐增加到 20 个,MAE 值逐渐减小,推荐结果越来越好。邻近用户数量从 25 个增加到 35 个,RBICF 及 RBCF 算法的 MAE 值虽然继续减小,但却趋于平稳,因为这时邻近用户集合中存在着一些与目标用户相似度较小的邻近用户,所以这些用户的历史评分对推荐结果的影响较小,所以推荐的结果并不会有很明显的变化。而邻近用户数量在 25~35 个时,RBITD 方法的 MAE 值虽然趋于平稳,但是却略有增大,因为此时的邻近用户数量已经超过一些推荐信任域的用户数量,在协同过滤的过程中不得不从信任域之外的用户中寻找相似度较大的用户,由此导致推荐效果略有下降。因此在设置邻近用户数量时,应该充分考虑推荐信任域的大小,以达到最佳的推荐效果。

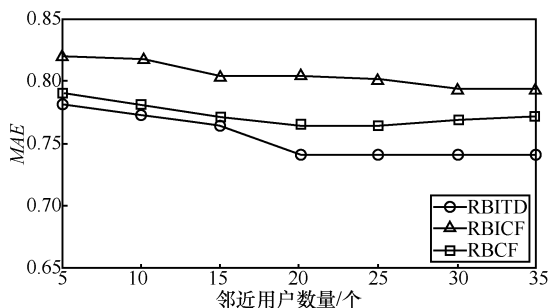


图3 微博推荐算法在不同数量邻近用户情况 MAE 对比

通过上述实验可以得出两个结论:一是当数据量较大时,RBITD 方法计算得出的微博评分与真实评分的平均绝对误差值(MAE)比其他两种方法更小,结果更加准确;二是在设置不同邻近用户数量的情况下,RBITD 方法得出结果与真实值的 MAE 值较另两种方法小,且在邻近用户数量为 20 个时,误差最小。

平均绝对误差 (MAE) 能够反映出微博评分的准确性,但由于微博的数量过大,在实际推荐时需要将微博按照预测评分由高到低排序,选取评分较高的 k 个作为推荐结果。因此,微博推荐的准确率及召回率与 k 值有着直接的关系。又因为每个用户能够接收到的微博数量不同,甚至会有很大差异,所以不能仅仅将 k 设为一个定值,所以在实验中需要按照每个用户能够接收到的微博的一定比例 R 来设置推荐列表长度 k 。以下为 R 在不同取值的情况下 3 种微博推荐方法的准确率与召回率的对比与分析。

图 4 和图 5 分别为 3 种微博推荐方法准确率和召回率

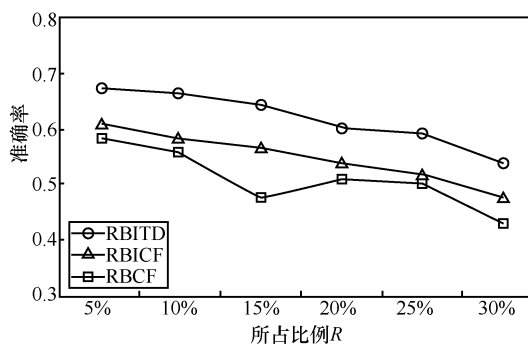


图4 准确率对比

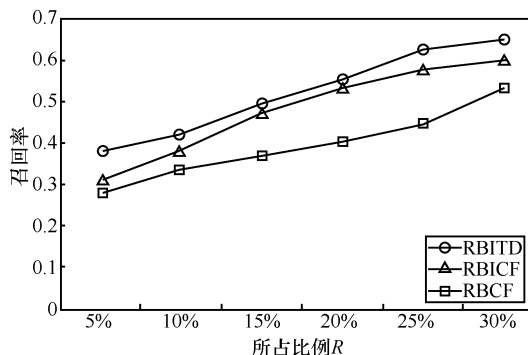


图5 召回率对比

的对比。图 4 表明,随着推荐列表长度的增加,准确率不断地下降。这是因为微博按照评分由高到低排序,当 R 值较大时,就会有一些用户不太感兴趣的微博加入到推荐列表中,造成准确率的降低。图 5 表明,随着推荐列表长度的增加,召回率不断上升。这是因为当 R 值较大时,推荐列表中加入了更多用户感兴趣的微博,导致召回率的上升。从图 4 和图 5 可以看出,虽然 3 种方法准确率和召回率随着 R 值增加的变化趋势大致相同,但 RBITD 方法在准确率和召回率上明显优于其他两种算法,这说明 RBITD 方法拥有更好性能。

7 结束语

微博的个性化推荐能够帮助用户更容易从海量的微博中找到符合自己需求的微博。为了准确预测用户对微博内容是否感兴趣,实现微博内容的个性化推荐,本文综合考虑了用户自身兴趣偏好与同一社区用户对微博的评价,提出了一种基于改进协同过滤方法的微博推荐方法。首先利用文本分类的方法发现用户的兴趣集合;同时提出了推荐信任域的概念,并利用多维 NF 算法发现用户所属的推荐信任域,以解决微博平台中数据的极度稀疏性对协同过滤算法造成的影响;之后,结合用户的兴趣偏移量及



推荐信任域对传统协同过滤算法进行改进;最后利用线性回归模型获取最终的预测结果。实验结果表明,本文提出的微博推荐算法有着更好的推荐效果,在一定程度上提高了推荐的质量。

参考文献

- 1 Busch M, Gade K, Larson B, *et al.* Earlybird: real-time search at Twitter. Proceedings of IEEE the 28th International Conference on Data Engineering, Washington, USA, 2012: 185~193
- 2 Chen C, Li F, Ooi B C, *et al.* TI: an efficient indexing mechanism for real-time search on tweets. Proceedings of the ACM International Conference on Management of Data, Athens, Greece, 2011: 649~660
- 3 Duan Y, Jiang L, Qin T, *et al.* An empirical study on learning to rank of tweets. Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, 2010: 295~303
- 4 Chen K, Chen T, Zheng G, *et al.* Collaborative personalized tweet recommendation. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, USA, 2012: 661~670
- 5 Chirita P, Nejdl W, Paiu R, *et al.* Using ODP metadata to personalize search. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, 2005: 178~185
- 6 Ramage D, Dumais S, *et al.* Characterizing microblogs with topic models. Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, USA, 2010: 130~137
- 7 Naveed N, Gottron T, Kunegis J, *et al.* Bad news travel fast: a content-based analysis of interestingness on twitter. Proceedings of the ACM WebSci'11, Koblenz, Germany, 2011: 1~7
- 8 Kapanipathi P, Orlandi F, Sheth A, *et al.* Personalized filtering of the twitter stream. Proceedings of Semantic Personalized Information Management Workshop, Bonn, Germany, 2011: 6~13
- 9 Chen L, Nayak R, Xu Y. A recommendation method for online dating networks based on social relations and demographic information. Proceedings of International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Kaohsiung, Taiwan, China, 2011: 407~411
- 10 Nayak R, Zhang M, Chen L. A social matching system for an online dating network: a preliminary study. 2010 Proceedings of IEEE International Conference on Data Mining Workshops, Sydney, Australia, 2010: 352~357
- 11 Newman M E J. Communities, modules and large-scale structure in networks. Nature Physics, 2012, 8(1): 25~31
- 12 Newman M E J. Fast algorithm for detecting Community structure in networks. Physical Review E, 2004, 69(6): 66~133
- 13 Girvan M, Newman M E J. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821~7826
- 14 Zhao Z D, Shang M S. User-based collaborative-filtering recommendation algorithms on Hadoop. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (WKDD'10), Portland, USA, 2010: 478~481
- 15 陈可寒, 韩盼盼, 吴健. 基于用户聚类的异构社交网络推荐算法. 计算机学报, 2013, 36(2): 349~359
Chen K H, Han P P, Wu J. User clustering based social network recommendation. Chinese Journal of Computers, 2013, 36(2): 349~359
- 16 罗奇, 余英, 赵呈领等. 自适应推荐算法在电子超市个性化服务系统中的应用研究. 通信学报, 2006, 27(11): 183~186
Luo Q, Yu Y, Zhao C L, *et al.* Research on personalized service system in E-supermarket by using adaptive recommendation algorithm. Journal on Communications, 2006, 27(11): 183~186
- 17 邓爱林, 朱扬勇, 施伯乐. 基于项目评分预测的协同过滤推荐算法. 软件学报, 2003, 14(9): 1621~1628
Deng A L, Zhu Y Y, Shi B L. A collaborative filtering recommendation algorithm based on item rating prediction. Journal of Software, 2003, 14(9): 1621~1628

[作者简介]



徐雅斌,男,北京信息科技大学教授、系主任,主要研究方向为社交网络、未来网络。



刘超,男,北京信息科技大学硕士生,主要研究方向为社交网络。



武装,男,博士,北京信息科技大学教授,主要研究方向为社交网络、未来网络。