预测 EBSN 中的活动参与度：内容，上下文和社会影响

RongDu †,ZhiwenYu †,TaoMei ‡,ZhitaoWang †,ZhuWang †,BinGuo

† Northwestern Polytechnical University, Xi'an 710129, Shaanxi, China ‡ Microsoft Research, Beijing 100080, China zhiwenyu@nwpu.edu.cn, tmei@microsoft.com

摘　要　新兴的基于事件的社交网络（EBSN）连接线上和线下的社交网络，为理解人们社交行为提供了巨大的机会。虽然现有的努力主要集中在调查传统社交网络服务（SNS）中的用户行为，但本文旨在探索 EBSN 中的个人行为，这仍然是一个未解决的问题。特别是，我们的方法通过发现一系列连接物理和网络空间的因素和影响个体参与 EBSN 活动的因素来预测活动参与度。这些因素，包括内容偏好，上下文（空间和时间）和社会影响，都是使用不同的模型和技术来提取的。我们进一步提出了一种新的奇异值分解与多因子邻域（SVD-MFN）算法，通过将发现的异质因素整合到一个框架中来预测活动参与度，这些因子通过邻域集合进行融合。基于来自豆瓣事件的实际数据的实验证明了所提出的 SVDMFN 算法优于现有技术的预测方法。

关 键 词　**活动预测；EBSN；内容偏好；上下文；社会影响**

1　绪论

随着以事件和活动为基础的社交应用的繁荣，人们的社交方式由线上交流向线下活动转移。人们可以通过这些应用组织并参加各种不同的社交活动，这可以进一步促进面对面的社交互动。这种社交媒体被称为基于事件的社交网络（EBSNs）[20]。尽管学者已经对 EBSN 进行了各式各样的研究，但在行为预测方面的研究仍然比较欠缺。理解用户参与事件的动机对于更好理解人们的社交网络并提供推荐和推送广告至关重要。与传统的社交网络服务（SNS）相比，EBSN 中的用户行为主要受到线下活动的驱动，并且受到一系列独特因素（如时空约束和特殊社交关系（主持人和参与者））的高度影响。因此，线下活动的这些性质在 EBSNs 中扮演着推动行为预测的角色，与 SNS 中的在线行为预测有所不同。图 1 显示了在中国的人们在一个典型 EBSN 事件中的属性，如下所述。

1.地点：活动举行的地方。通常，一项活动在一个方便和受欢迎的地方举行。

图 1、一个豆瓣上典型事件的例子：该事件有五个关键属性：地点，事件，组织者，参与者和内容

2.时间：活动开始和结束的时候。事实上，时间的分配高度依赖于活动的内容。

3.参与者：点击“我想要参加”的用户

4.举办者：举办事件的用户

5.内容：事件的详细信息，包括种类，标题和事件描述

基于这些要素，我们可以使用不同的方法预测用户的活动出勤率。但是，如何系统地将异构信息进行全面组合仍然是一个未解决的问题。在本文中，我们将上述要素建模型，主要由三个不同因素组成：1）内容。决定用户是否会参与某种活动的一个关键因素。我们使用内容偏好来表示服务对象的不同活动。 2）上下文。我们方法中的上下文是指空间上下文和时间上下文。具体而言，用户在参加不同的流动活动时可能会有不同的时间和地点偏好。例如，如果用户是学生，即使他/她喜欢该活动，他/她参加在工作日期间举行的活动的可能性也会很低。同样，如果用户距离活动地点很远，则他也极有可能不参加该事件。 3）社会影响。在 EBSN 中，一个人参加某项活动的意愿可能会受到他或她与事件组织者以及其他事件参与者的社会关系的影响。然而，由于时间和地点限制的活动，主持人通常会扮演更加重要的角色。原因在于，组织者可以向其他参与者推荐活动，另一方面，如果用户经常参加由有影响力的组织者举办的活动，则参与由他组织的其他活动的可能性也将很高。

基于这些因素和多因素模型，我们提出了一种新的方法，称为奇异向量分解（SVD）和多因子邻域（SVD-MFN），一个基于 SVD 的方法，能够将不同的特征集成到多因素模型中，充分利用了 SVD 和多因素模型的优点。主要贡献可概括如下：

1.我们研究了如何预测即将到来的事件的参与度，这是据我们所知针对该问题的第一次尝试。

2.我们研究了影响用户行为的三个关键因素：内容偏好，地点-时间上下文和社会影响。

3.我们提出了一个新奇的算法（SVD-MFN），用来整合我们在上文提出的异构因素。该算法胜过了现在的最佳方法。

本文的其余部分安排如下：我们首先回顾相关工作。然后我们正式定义问题并提出系统框架。然后描述特征提取和融合，然后详细描述所提出的 SVD-MFN 算法。我们接下来展示实验结果，总结我们的工作，并在最后部分讨论可能的未来方向。

2 相关工作

我们简要回顾一下相关工作，可以分为三类：

第一类是研究理解 EBSN 在线与社会互动之间的关系。刘等人提出了 EBSN 的概念，并利用在线和社交链接[20]关注社区检测。Han 等人试图根据豆瓣事件[12]获取关于用户行为的信息。通过研究豆瓣中的事件，他们展示了与事件相关的结果，参与事件的用户行为以及事件的社会影响，这有助于我们更好地理解事件中影响用户行为的因素。 Xu 等人进行了定量分析，揭示了在线跟踪行为与真实事件特征之间的关系[27]。我们的工作和这一研究的不同之处在于我们进一步推进，以利用 EBSN 的独特特征来预测用户是否参加活动。

第二类包括对活动或事件预测的研究。对于活动建议有一些方法。例如，在匹兹堡地区，文化活动推荐围绕着信任关系而建立[16]。社交网络的许多事件更多关注社交网络分析（SNA）和协作过滤（CF）[14]。 Cornelis 等人开发了一种混合事件推荐方法 [7]。 Dalyetal 建立了一个有趣的事件管理服务，在制定推荐时考虑事件的位置[9]。他们发现，与会者有时来自附近的地点，并提出了一种基于位置的方法来向目标用户推荐当地活动。 Minkov 等人提出了一些建议，但不包括仅考虑内容信息的应用[22]。Zhuang 等人。探索使用语境推荐[32,25]。然而，现有研究仅考虑了 EBSN 事件的两个方面，没有一个研究者已经开发出了综合系统地研究不同信息和异构信息的方法。

第三类是基于位置的社交网络，它也包含线上和线下社交事件[26,30]。虽然相邻检查表明隐含的社交交往和社交关系[6]，但检查数据通常太稀疏，不足以代表人类行为[23]。 Crandall 等人研究地理特征来推断社会联系[8]。同样，庄等人。形式化地预测了社交网络，并使用因子图模型来预测两个用户未来会遇到的可能性[31]。但是，这些预测方法都不是事件驱动的。此外，虽然上述研究主要使用空间信息进行预测，但本文中考虑更加“细化”的特征，如内容偏好，空间和时间背景以及社会影响。

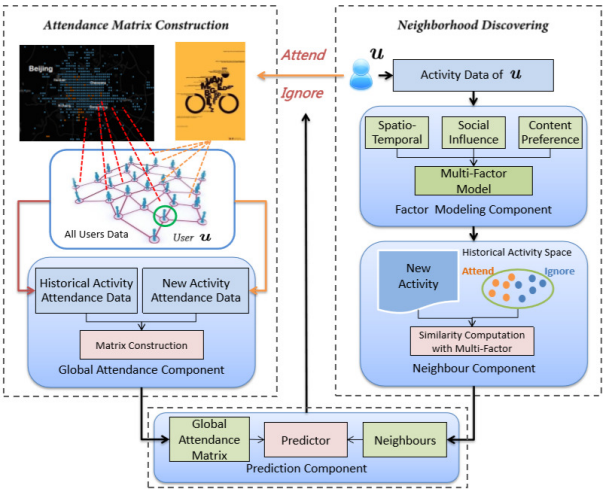


图 2、预测事件参与度的框架

3 问题描述

在 EBSN 中，有一系列活动$\{a_1, a_2, \ldots, a_m\}$和用户列表$\{u_1, u_2, \ldots, u_m\}$。让$a_u$表示您曾经参加过特定用户的所有活动。用户偏好通常是从历史行为中提取的。我们使用$F_u(a)$来表示用户u对活动a的偏好，它由以下五部分组成。$CP_u(a)$是内容偏好，$DP_u(a)$是空间背景下的距离偏好，$WP_u(a)$和$SP_u(a)$是时间背景下的星期几和小时偏好，$SP_u(a)$表示社会影响，这意味着你和事件举办者的关系。这些因素对可以归结为三个宏观方面：内容，背景和社会影响。我们还可以通过结合上述三个方面来获得从A_u中选择的每个用户 - 活动对的活动邻居集合$NS_u(a)$，这意味着从用户的出勤历史中选择的最近的活动。EBSN 中的目标用户的活动出勤预测任务可以被描述为：给用户参与历史A_u和即将到来的活动 a 的参与者U_a集合，我们预测目标用户是否将出席该事件。预测结果表示为r_{ua}，其中可用的值为 1 或 0，1 表示“参加”，0 表示“忽略”。在本文中，我们考虑了$F_u(a)$和$NS_u(a)$来预测有效的活动出勤率。关键问题列举如下：

1）如何衡量不同因素的影响，如何将其组织进一个单一的模型

2）如何提取活动邻居集合$NS_u(a)$。

3）如何在预测r_{ua}中将$F_u(a)$和$NS_u(a)$结合起来。

4 系统概览

图 2 显示了我们框架的概述，它由三部分组成：参与矩阵构建组件（左），邻域发现与多因素组件（右）以及预测组件（底部）

预测矩阵的构建：在一方面，在给出目标用户 u 和目前的参与者 a，我们可以为活动创建一个二维矩阵。在另一方面，我们可以为所有用户分别创建历史参与矩阵。通过矩阵的更新，来获得全局的参与矩阵，并将其作为预测部件的输入。

多因素情况下的邻居发现：第二部分是基于目标用户的历史性考察，这是我们框架中的关键组成部分。我们首先从三个方面为每个活动提取特征：内容偏好，时空背景和社会影响。然后，考虑到这三方面的不同影响，我们提出了一个使用决策树评估其贡献的多因素（MF）模型。基于这个模型，我们可以从目标用户的角度计算事件之间的相似度。此外，我们可以从出席历史中发现目标用户的邻居活动。

预测组件：为了将以前的部分结合到我们的系统中，我们在预测部分中提出了 SVD-MFN 预测器。我们将在 SVD-MFN 算法部分给出细节。

5 特征建模

在本节中，我们首先从三个宏观方面阐述我们的多因素模型中使用的特征：内容偏好，时空背景和社会影响。第一个原则的基础是首先研究影响活动参与的因素，然后，我们展示不同的特征如何融合在一起。

5.1 特征抽取

活动内容在确定用户参与事件的可能性方面起着重要作用[22]。因此，计算用户的历史活动与即将到来的活动之间的内容相似性是关键步骤。适当的相似性度量可以极大地影响我们系统的性能。在豆瓣事件中，有三个元素用于表征活动内容：类别，标题和描述，如图 1 所示。我们将这些元素放在一起作为整个文本，然后将一对活动之间的内容相似度定义为文本相似度。许多研究试图解决文本相似性问题。一种方法是利用搜索引擎扩展和丰富文本中的关键词[4]。另一种方法是使用 WordNet 这样的自编的数据库来挖掘单词之间的关系[17]。虽然词汇数据库中存在大量词汇及其语义关系，但基于词典的相似度计算的应用范围相当有限。隐式狄利克雷分配（LDA）模型，可以解决所有上述问题[3]。LDA 基于文件是话题的混合的想法，其中话题

是话语的概率分布。我们从删除停用词（即标点符号）和短语（即“of”和“and”）开始，然后格式化剩余的文本作为 LDA 的输入。之后，通过使用吉布斯采样将格式化的文本映射到主题空间。然后，Jensen-Shannon（JS）距离通常用于计算文本相似度[19]。最后，为了获得 LDA 中最好的主题，我们使用聚类方法[4]。基于 LDA 生成具有 n 个词的文本的过程可以用边缘分布来描述：

$$\mathrm{P(d)} = \int_{\theta}\left(\prod_{i=1}^{n}\sum_{T^{i}} P\left(W^{(i)}|T^{(i)},\beta\right)P\left(T^{(i)}|\beta\right)\right)P(\theta|\alpha)d\theta \quad (1)$$

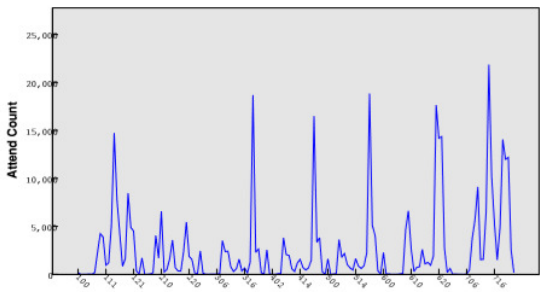
我们使用吉布斯采样来从文本库中抽取主题，然后将采样结果作为文本相似度的输入，为了衡量采样得到的两个分布间距离，我们使用了 KL 散度来计算[19]。

$$\mathrm{D_{KL}}(p,q) = \sum_{j=1}^{T} p_j \log_2 \frac{p_j}{q_i} \quad (2)$$

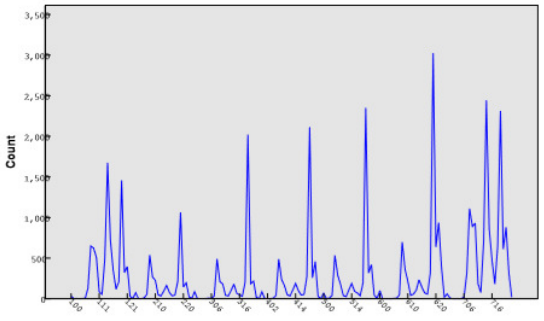
KL 散度是非对称的，但我们可以通过 KL 散度来方便的计算出对称的 JS 散度。如公式（3）。

$$\mathrm{JS(p,q)} = \frac{1}{2}\left[D_{kl}\left(p,\frac{p+q}{2}\right) + D_{kl}\left(q,\frac{p+q}{2}\right)\right] \quad (3)$$

举个例子，考虑两个活动的事件描述：



(a) The user attendance time histogram over hour of one week.



(b) The activity start time histogram over hour of one week

图 3、时间柱状图，横轴为一周的小时数

a1：“ drama British Shakespeare classic Macbeth. ”

a2：“ British Shakespeare King Lears. ”

在应用了 z=2 的 LDA 后，提取的主题词如下：

T1:“ British Shakespeare ”

T2:“ drama classic ”

很明显，第一个活动在两个主题中都有 50％的成员，因为它包含来自两个主题的同等程度的词汇，而活动 a2 分别在 T1 和 T2 中具有 100％的成员。然后，我们可以将每个活动表示为其主题成员的矢量：

$$a_1 = [0.5,0.5]$$
$$a_2 = [1.0,0.0]$$

其中每个向量中的第一个元素对应于它们在主题 T1 中的分布，第二个元素对应于 T2 中的分布，表示为$\theta^{(a1)}$和$\theta^{(a2)}$。因此，可以通过将$\theta^{(a1)}$和$\theta^{(a2)}$计算 a1 和 a2 之间的 JS 散度。等于 0.31。因此，我们可以通过 JS 散度得到 a1 和 a2 之间的内容相似性：在个例子中，内容相似性为 0. 69。

$$\mathrm{Sim(a_1,a_2)} = 1 - \mathrm{JS}\left(\theta^{(a_1)},\theta^{(a_2)}\right) \quad (4)$$

基于两个活动的内容相似度，我们可以计算用户的内容偏好/兴趣。具体而言，我们在工作中采用遗忘机制的兴趣转移[5]，其关键思想如下：

一方面，人们的兴趣随着时间的流逝而逐渐消失，如记忆。例如，用户最近参加的活动应该对未来行为的预测产生的影响比对很久以前发生的事件影响更大。另一方面，随着积累的利益变得更加稳定，遗忘速度减慢。基于这两个原则，我们通过引入遗忘机制，使用短期兴趣模型（STIM）来表示用户最近的兴趣和长期兴趣模型（LTIM）来表示积累的稳定兴趣，从而为不同目的建立了两种兴趣模型。遗忘函数的实现是基于以下等式来模拟用户兴趣的衰减：$\mathrm{I}(a,a^i) = exp\left\{-\frac{ln2\times\left(t_a - t_{a_i}\right)}{hl}\right\}$,其中$\mathrm{I}(a,a^i)$为遗忘系数，表示用户初始兴趣的衰退。$t_a$指用户参与时间，$t_{a_i}$指活动正式举办时间。$hl$是以天为单位的半衰期。针对短期兴趣，$hl$可以是恒定值，针对长期兴趣，用户的兴趣通常会变得更稳定。我们使用$\mathrm{I}(a,a^i) =$

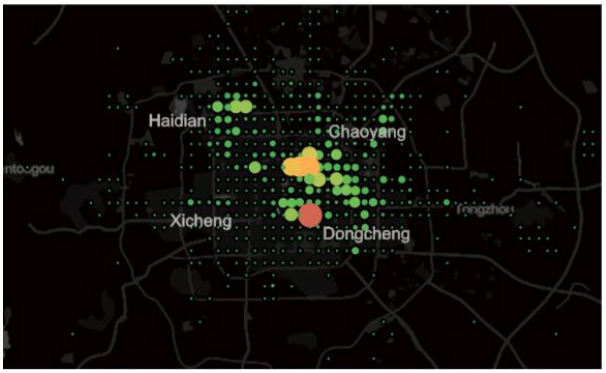


图 4、北京市中活动地点的分布

$exp\left\{-\frac{ln2\times\left(t_a-t_{a_i}\right)}{\hbar l_0+d_{acc}\times s}\right\}$来计算遗忘系数。$\hbar l_0$代表初始半衰期，$d_{acc}$表示初始 LTIM 延续了多少天，常数 s 用来调整d_{acc}带来的影响。总结来说，我们使用公式（5）来表示遗忘系数：

$$\mathrm{I}(a,a_i)=\begin{cases}e^{-\frac{ln2\times\left(t_a-t_{a_i}\right)}{hl_0}}, & \left(t_a-t_{a_i}\right)\le d_{acc}\\ e^{-\frac{ln2\times\left(t_a-t_{a_i}\right)}{\hbar l_0+d_{acc}\times s}}, & \left(t_a-t_{a_i}\right)>d_{acc}\end{cases} \quad (5)$$

针对目标用户 u，我们使用式$CS_u(a,a_i)=I(a,a_i)\times Sim(a,a_i)$来定义事件间的内容相似度。其中$\mathrm{Sim(a,a_i)}$为用公式（4）计算的内容相似度，为了同时考虑到两事件的内容相似度和用户兴趣的衰减程度，我们使用公式（6）来定义用户u的内容偏好程度$\mathrm{CP_u}(a_i)$:

$$\mathrm{CP_u}(a)=\frac{\sum_{a_i\in A_u}I(a,a_i)\times Sim(a,a_i)}{\sum_{a_i\in A_u}Sim(a,a_i)} \quad (6)$$

5.2 空间和时间上下文

1）时间上下文

时间上下文对活动预测非常重要。一方面，如图 3（a）所示，人类行为表现出强烈的每日和每周的周期性模式。另一方面，活动的开始时间也是周期性的，如图 3（b）所示。

a）周内因素

用户的日常生活呈现按周的周期性，所以我们引入周内因素$\mathrm{WP_u}(a)$作为第一个用户的时间偏好，如公式（7）

$$\mathrm{WS_u}(a,a_i)=\begin{cases}1, & wd(t_a)=wd(t_{a_i})\\ 0, & wd(t_a)\ne wd(t_{a_i})\end{cases} \quad (7)$$

其中$\mathrm{wd_t}(t_a)$表示用户参加事件 a 是在一周的第几天，$\mathrm{wd(t_a)}\in\{1,2,3,4,5,6,7\}$，因此，u 的周内偏好可以被定义为公式（8）

$$\mathrm{WP_u}(a)=\frac{\sum_{a_i\in A_u}WS_u(a,a_i)\times Sim(a,a_i)}{\sum_{a_i\in A_u}Sim(a,a_i)} \quad (8)$$

b) 一天中的时间因素

在图 3（a）中，我们发现一天中用户的活动出勤也是周期性的，可以解释如下。豆瓣上的大部分用户都是学生或白领。如果在学习或工作时间进行活动，那么即使他们对该活动感兴趣，这些用户也不可能参与。我们使用高斯公式来表示时间相似性：$\mathrm{HS_u}(a,a_i)=\exp\left(-\frac{\left(t_a-t_{a_i}\right)^2}{2}\right)$，因此，我们可以通过公式（9）来表达一天中的时间因素：

$$HP_u(a)=\frac{\sum_{a_i\in A_u}HS_u(a,a_i)\times Sim(a,a_i)}{\sum_{a_i\in A_u}Sim(a,a_i)} \quad (9)$$

2) 空间上下文

我们注意到参加活动的可能性在用户的位置和活动的位置增加之间降低了很多，这并不令人惊讶，并且已被众多研究人员证明[28]。而且，用户对地点有个人偏好。例如，如果到一个地方的交通方便，其周围的活动将更受欢迎。图 4 显示了我们数据集中所有活动的位置分布。我们观察到在中国北京的海淀区，朝阳区和东城区有更多的活动，这些区域由图中较大和较亮的圆点代表。此外，类似的活动往往位于同一地区。例如，大多数教育活动都在海淀区举行，那里有许多大学[27]。

为了得出两个活动之间的空间相似性，我们根据用户的出勤历史来计算她的位置偏好，因为用户的真实家庭位置难以获得。与一天中的小时数相似，我们采用高斯公式来计算距离相似度：$\mathrm{DS_u}(a,a_i)=\exp\left\{-\frac{Distance(a,a_i)^2}{2}\right\}$，同样的，公式（10）给出了用户的空间因子的定义。

$$\mathrm{DP_u}(a)=\frac{\sum_{a_i\in A_u}DS_u(a,a_i)\times Sim(a,a_i)}{\sum_{a_i\in A_u}Sim(a,a_i)} \quad (10)$$

5.3 社会影响

社交友谊对于事件预测和推荐是有益的[29]，这是激励用户参与社交活动的关键因素[1]。我们定义用户和事件组织者之间的两种社交关系。第一种类型如下：在豆瓣活动中，活动主持人可以向其追随者发送邀请。因此，如果用户跟随主持人，一旦她被邀请，她可能更愿意参

加由该主持人组织的活动。第二种类型是偏好关系。例如，用户可能参加了由运营健身房的主持人组织的许多体育赛事，因为用户对该主持人感兴趣或者她只是健身房的成员。因此，她更有可能参加今后的体育活动，这些体育活动是由事件组织者来组织的。基于以上描述，我们将即将发生的事件a与过去的事件a_i之间的社会相似度定义为公式（11）

$$SS_u(a,a_i) = S_1\big(u,H(a)\big) * \delta + S_2\big(H(a),H(a_i)\big) * (1-\delta) \quad (11)$$

其中，H(a)为事件 a 的举办者，S_1为用户间的关系，S_2为用户间的偏好关系。如果用户 u 跟随 H(a)，那么$S_1\big(u,H(a)\big)=1$，否则$S_1\big(u,H(a)\big)=0$。上述关系在S_2中也成立：对于由相同事件组织者举办的两个事件a, a_i，如果$H(a)=H(a_i)$，那么$S_2\big(H(a),H(a_i)\big)=1$，否则为 0。我们将参数δ设置为 0.5，即公式（11）的两项有相同的权重。用户的社会影响可定义为式（12）：

$$SP_u(a) = \frac{\sum_{a_i \in A_u} SS_u(a,a_i) \times Sim(a,a_i)}{\sum_{a_i \in A_u} Sim(a,a_i)} \quad (12)$$

到目前为止，我们已经从三个层面来提取相似特征，接下来我们需要将其有机结合在一起，以方便接下来的活动参与度预测。

5.4 特征融合

由于不同的特征对用户偏好有不同的影响，所以如何合理的对特征进行评估，然后将它们融合在一起是一个挑战。在 EBSN 中，用户对活动的反馈是二元的（即 1 或 0），这意味着她参加或忽略了一个事件。因此，我们可以将用户是否参与该活动的预测转化为分类问题。然后用分类算法来解决。在分类结果的基础上，我们首先得到不同特征的贡献权重，然后将这些特征按贡献权重线性组合。我们通过式（13）来定义即将到来的事件a和已知某事件a_i对某用户u的总体相似度：

$$Sim_u(a,a_i) = \alpha * CS_u(a,a_i) + \beta * WS_u(a,a_i) + \gamma * HS_u(a,a_i) + \delta * DS_u(a,a_i) + \varepsilon * SS_u(a,a_i) \quad (13)$$

其中，$\alpha,\beta,\gamma,\delta,\varepsilon$代表着不同项的权重，分别的，α代表整体内容偏好权重，β,γ代表对星期几的偏好和对一天中时间因素的权重，δ代表空间上下文的权重，ε代表时间上下文的权重。

6 SVD-MFN 算法

为有效利用不同的特征进行活动参与预测，我们提出了一种基于 SVD 的新的多因子邻域奇异值分解算法（SVD-MFN）。本节我们将首先简单介绍 SVD，接着对本文提出的算法进行详细说明。

6.1 矩阵分解：SVD

矩阵的奇异值分解（SVD）是一个著名的矩阵因式分解的技术，它能处理稀疏的，高维的数据[15]。在一些领域例如广告投放和电子商务中获得了广泛的应用[18][21]。SVD 的基本思想是将矩阵由一组正交的基向量和对角矩阵来表示。在本文中，SVD 被用来处理事件参与矩阵。我们使用向量p_u来表示用户 u，q_a来表示事件 a，其中$p_u, q_a \in R^d$。相应的，用户 u 对事件 a 的参与度被表示为$r_{ua} = p_u^T \cdot q_a$。我们使用一个回归方程将参与度归一化到(0,1)区间中，然后设置 0.5 作为划分线：即如果参与度大于 0.5，该用户就会参加该事件，反之则会忽略该事件。我们通过最小化目标函数（14）来设置参数。

$$\min_{p^*,q^*} \sum_{(u,a)\in R} (r_{ua} - p_u^T \cdot q_a)^2 + \lambda\left(\sum_u |p_u|^2 + \sum_a |q_a|^2\right) \quad (14)$$

其中，λ为预先设置的正则化参数，在实验中我们将其设置为 0.01，我们使用随机梯度下降的算法来训练模型。

6.2 SVD-MFN:将多因子领域方法与SVD结合

传统的邻域方法着重于计算活动者或用户之间的关系。尽管它们在检测空间相邻性以及在邻居数量足够多的情况下表现不错，但当其观察的样本比较稀疏，周围没有或只有有限个样本时，它们可能无法工作。相比之下，经 SVD 分解后获得的潜在信息能够有效的捕捉到全局范围内的信息，因而有着更好的效果。我们通过公式（15）将基于邻域的预测方法和矩阵因式分解结合起来：

$$\widetilde{r_{ua}} = p_u^T \cdot q_a + |N(u,a;k)|^{-\frac{1}{2}} \sum_{i\in N(u,a;k)} w_{ai}(r_{ui} - \bar{r_u}) \quad (15)$$

其中$\bar{r_u}$指用户 u 的平均参与度，$N(u,a;k)$为基于 knn 选出的用户 u 参加的事件 a 的 k 个邻居。特别的，w_{ai}为矩阵因式分解获得的对角阵。在计算中，我们只需要存储和更新每个事件最接近的 k 个邻居事件，我们使用$Sim_u(a,a_i)$来

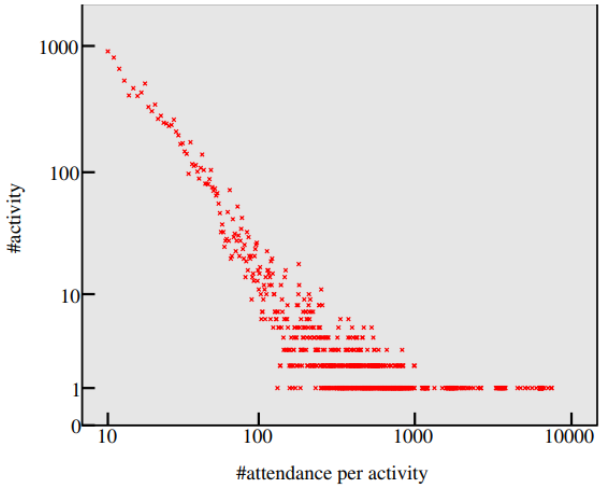


图 6、事件参与人数的分布

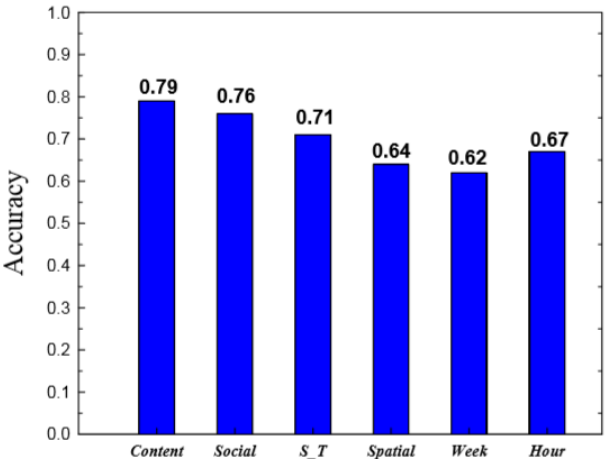


图 5、使用决策树预测参与度中各个属性的表现

确定$N(u,a;k)$中的 k 个近邻事件，通过这种方式我们整合了所有事件特征

7 实验结果

在本节中，我们将根据真实的 EBSNs 数据集评估所提出的框架和方法。我们首先介绍 EBSNs 数据集，然后介绍框架中不同过程的参数学习实验。之后，根据决策树对多因素模型中不同特征的权重进行评估。最后，我们将 SVD-MFN 方法与另外三种现有方法进行比较，结果表明我们的方法性能更好。

表格 1、豆瓣事件的详细数据

时间	2012-01-01 to 2013-10-01
用户数量	15,050
活动数量	45,561
组织者	6,570
关注人数	313,479
用户-活动对数量	481,325

7.1 数据集

我们的实验是在从豆瓣事件收集的数据集上进行的。我们使用豆瓣提供的 API 在指定的时间间隔内抓取所有有效的活动。我们从 2013 年开始参加三项以上活动的北京用户中选择了用户。之后，我们从 2012 年 2 月到 2013 年 10 月对这些用户的活动历史进行了检索。总共得到了三种数据：

1）用户活动数据，即用户活动历史

2）活动属性数据，对每一个事件，我们爬取了其开始时间，地址，地理座标，种类，标题和事件描述。

3）社会数据，为了计算社会影响，我们也收集了事件参与者和事件举办者的关系。

详细的数据集信息如表 1 所示，为了研究人们参与事件的特性，我们首先对事件参与人数进行了简单的数据分析。如图 5 所示，结果呈现长尾分布，这说明了大多数事件只有少部分人参加。而少部分事件有大量人参加。在我们爬取的数据集中，每个事件的参与人数的平均数为 31.98。

在获得的活动属性数据基础上，我们构建了特征模型并实现了我们的预测方法。具体来说，在实验中，我们将数据集分成两部分。第一部分（2012 年 2 月至 2013 年 5 月）用于训练模型，第二部分（2013 年 6 月至 2013 年 10 月）用于测试。同时，收集的用户活动对被视为正样本。但是，为了无差错地训练二元分类器，应该有相同数量的正负样本。由于我们的数据集共有 481,325 个正样本，因此我们从活动用户中随机选择了 481,325 个用户活动对，并且将其作为负样本参加相应的活动集合。

7.2 特征衡量

在多因素模型中，我们从三个不同方面考虑用户偏好：内容，时间背景和社会影响。为了评估他们对活动参与预测的贡献，我们采用了决策树来分别使用每个类别的特征来预测参与度。我们运行 10 折交叉验证来执行预测并采用精度作为衡量标准。所有的测试都基于 WEKA [11]，结果如图 6 所示

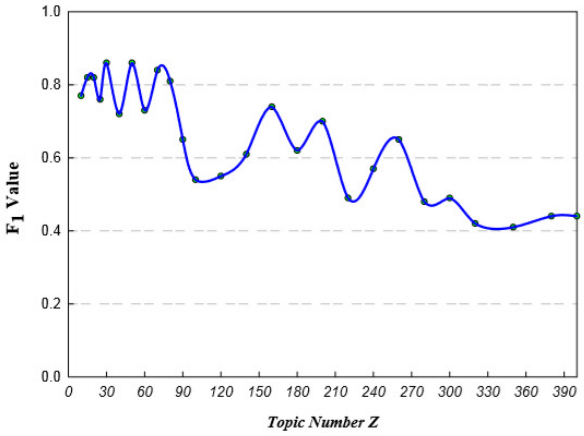


图 8、不同话题数下的F_1 值

根据结果，我们发现不同的特征在预测中表现有差异。具体而言，内容偏好以 0.79 的准确度达到最佳性能。社会影响排在第二位，准确度为 0.76。空间和时间上下文（S_T）也对预测作出了贡献，但其表现并不像其他两个因素那样好。

7.3 参数学习

LDA 中的参数学习

在提取内容偏好时，我们需要确定 LDA 中参数的最优值，即 α， β 和主题 Z 的数量. α 和 β 的最优值取决于 Z 和文档中词汇的大小收集，它们通常设定为 α = 50 / Z 和 β = 0.01 [19]。 Z 可以影响提取的主题的可解释性，从而影响文本相似度的计算结果。如果 Z 的值很小，那么获得的主题将会过于笼统，如果 Z 过大，则会导致主题词的适用范围过于狭窄，导致所有文本都不被认为相似。

为了选择 Z 的最佳值，我们使用了文本分类方法。豆瓣活动中的所有活动分为 10 类：音乐，电影，沙龙，体育，英联邦，派对，旅游，展览，戏剧等。基于 JS 距离，我们首先计算每对活动之间的内容相似度。之后，我们使用 k-means ++将所有活动分为 10 组。然后采用 F1 评估方法进行绩效评估[24]。最好的主题编号 Z 应该对应于最高的 F1 值。结果如图 7 所示。根据图，我们将主题数 Z 设置为 50，其中 F1 达到最佳性能，等于 0.86。

SVD 中的参数学习

在设置 SVD 的参数时，维数 d 是我们最密切考虑的。我们用精度来评估它对 SVD 性能的影响。

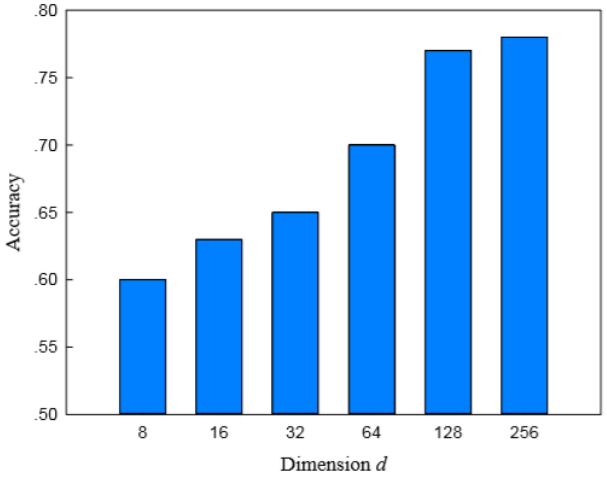


图 7、不同维度 d 下的表现

如图 8 所示，维数 d 的多少对预测的准确性有很大的影响。换句话说，维数对维数高的表示有效。我们进行了 d 值范围从 8 到 256 的各种不同的实验。根据图 8，虽然维数较低，但是其性能持续提高。然而，当 d 大于 128 时，耗时急剧增加。为了平衡准确性和时间复杂度，在下面的实验中，d 被固定为 128，对应的准确度为 0.77。

SVD-MFN 中的参数学习

在 SVD-MFN 中，我们需要重点关注的参数是选取相似的邻居数 k。为了检验预测结果对参数 k 大小的敏感性，我们在各个邻居的数量从 1 到 25 的范围内进行了实验，然后计算出相应的预测精度。图 9 显示了实验结果。

从图 9 可以看出，k 的大小确实会影响预测精度。具体而言，预测精度随着邻居数量增加而增加，然后达到极限后开始下降。这可能是由于太多邻居导致太噪音的事实。我们选择 k=7 作为

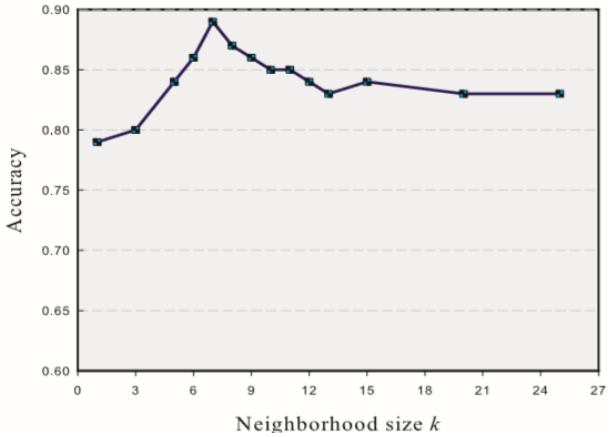


图 9、不同 k 对实验结果的影响

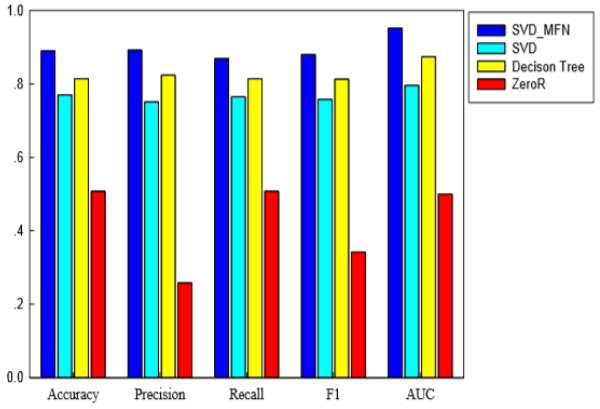


图 10、SVD-MFN 与其他方法的比较

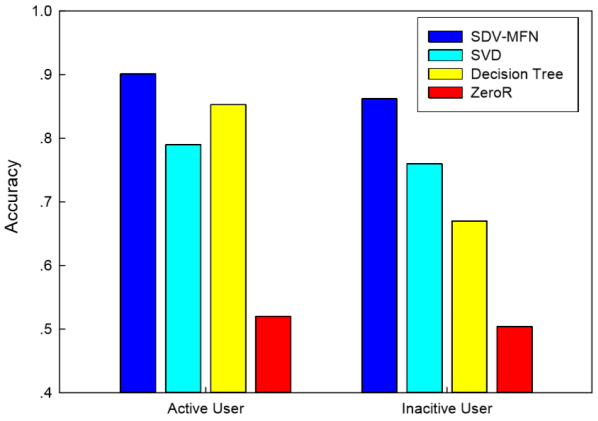


图 11、不同种类用户下的表现

正式实验中使用的参数。

和其他方法的比较

为了检验 SVD-MFN 的性能，我们将比较三种方法，即决策树，SVD 和 ZeroR。对于决策树，我们使用我们的因子模型中的所有特征。对于 SVD，我们将维数设置为 128，并保持其他参数与 SVD-MFN 中的相同。对于 ZeroR 来说，它是最简单的分类方法，常常被用作其他分类方法的基准。

根据图 10 所示的结果，我们比较了 SVD，决策树和 ZeroR 的各种测量方法的性能，结果表明我们的方法取得了较好的性能，其中包括准确度，精度，召回率，F 值和 AUC 等。具体而言，在准确性方面，SVD-MFN 达到 0.89，比决策树高 0.08，比 SVD 高 0.12，比 ZeroR 高 0.39。同样，SVD-MFN 的 F 值也最高（0.88），比 SVD 高 0.122，决定树高 0.067，比 ZeroR 高 0.538。在 AUC 的情况下，SVD-MFN 的性能高达 0.952。由于 SVD 仅使用用户活动矩阵，因此 SVD 的性能较差。决策树的性能优于 SVD，因为它也考虑了活动特征。为了更深入地了解我们算法的性能，我们还计算了 Cohen 的 Kappa [2]值。Kappa 值是作为衡量心理行为观察者之间的一致性。后来才发现，它也可以用来衡量分类器准确度，通过评估分类器与现实之间的一致程度。 SVD-MFN 中的 Cohen Kappa 为 0.758，这是一个相当高的值，证明了我们算法的合理性。

SVD-MFN 在不同类型用户中的表现测试

在 EBSN 中，用户在活动级别上的行为往往不同。一些用户比其他用户更频繁地参加活动。由于这种现象，每个用户可以被定义为活跃或不活跃。在我们的数据集中，每位用户的平均出勤人数为 31.98。为了获得不同类型用户下模型的表现，我们将测试数据集分为两个子集：活跃用户（出席数≥32）和非活跃用户（出席数<32）。结果如图 11 所示。我们观察到，与其他方法相比，SVD-MFN 为活跃和非活跃用户提供了高准确度，证明了其一致性。

8 结论和展望

在本文中，我们专注于预测 EBSN 中活动参与度的问题。我们利用内容，时间，空间上下文和社会影响力等特征对 EBSN 进行建模，其中包含了不同特征的权重。基于这个模型，我们获得了当前用户-活动对的所有相似事件。具体而言，我们提出了将 SVD 与相似事件相结合的 SVD-MFN 用于解决预测参与度问题。我们的实验结果表明我们的模型优于其他模型。我们的工作对于更好地了解 EBSN，支持 EBSN 服务的个性化推荐和定位广告，从而提高客户对 EBSN 服务的满意度有着重要意义。

尽管实验结果表明我们提出的方法在预测活动参与度中是有效的，但结果仅基于特定的社交网络服务，即豆瓣事件。因此，将来有必要将 SVD-MFN 算法扩展到其他社交媒体。此外，我们还希望通过结合影响力最大化将我们的工作扩展到广告。例如，作为事件组织者，他/她邀请谁参加该活动以获得最佳的广告效果。

9 致谢

这项工作得到了中国国家基础研究计划（2012CB316400），中国国家自然科学基金（61222209,61373119,61332005），高等教育博士点专项科研基金（20126102110043）和微软的部

分支持。

10 参考文献

1. L. Backstrom, D. P. Huttenlocher, J. M. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In Knowledge Discovery and Data Mining, pages 44–54, 2006.
2. A. Ben-David. Comparison of classification accuracy using cohens weighted kappa. Expert Systems with Applications, 34(2):825–832, 2008.
3. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Advances in neural information processing systems, 1:601–608, 2002.
4. D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using web search engines. In World Wide Web Conference Series, pages 757–766, 2007.
5. Y. Cheng, G. Qiu, J. Bu, K. Liu, Y. Han, C. Wang, and C. Chen. Model bloggers' interests based on forgetting mechanism. In World Wide Web Conference Series, pages 1129–1130, 2008.
6. E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In Knowledge Discovery and Data Mining, pages 1082–1090. ACM, 2011.
7. C. Cornelis, X. Guo, J. Lu, and G. Zhang. A Fuzzy Relational Approach to Event Recommendation. In Indian International Conference on Artificial Intelligence, pages 2231–2242, 2005.
8. D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. Proceedings of the National Academy of Sciences, 107(52):22436–22441, 2010.
9. E. M. Daly and W. Geyer. Effective event discovery: using location and social information for scoping event recommendations. In Proceedings of ACM Conference on Recommender Systems, pages 277–280. ACM, 2011.
10. T. L. Griffiths. Finding scientific topics. Proceedings of The National Academy of Sciences, 101:5228–5235, 2004.
11. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. Sigkdd Explorations, 11:10–18, 2009.
12. J. Han, J. Niu, A. Chin, W. Wang, C. Tong, and X. Wang. How online social network affects offline events: A case study on douban. In Ubiquitous Intelligence & Computing and International Conference on Autonomic & Trusted Computing, pages 752–757, 2012.
13. J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In Research and Development in Information Retrieval, pages 230–237, 1999.
14. R. Klamma, P. M. Cuong, and Y. Cao. You never walk alone: Recommending academic events based on social network analysis. In Complex Sciences, pages 657–670. 2009.
15. Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. Computer, 42(8):30–37, 2009.
16. D. H. Lee. PITTCULT: trust-based cultural event recommender. In Conference on Recommender Systems, pages 311–314, 2008.
17. Y. Li, D. Mclean, Z. A. Bandar, J. D. O'Shea, and K. A. Crockett. Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Transactions on Knowledge and Data Engineering, 18:1138–1150, 2006.
18. Y.-M. Li, C.-T. Wu, and C.-Y. Lai. A social recommender mechanism for e-commerce: Combining similarity, trust, and relationship. Decision Support Systems, 55(3):740–752, 2013.
19. J. Lin. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory, 37:145–151, 1991.
20. X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han. Event-based social networks: linking the online and offline social worlds. In Knowledge Discovery and Data Mining, pages 1032–1040, 2012.
21. A. K. Menon, K.-P. Chitrapura, S. Garg, D. Agarwal, and N. Kota. Response prediction using collaborative filtering with hierarchies and side-information. In Knowledge Discovery and Data Mining, pages 1032–1040, 2012.
22. E. Minkov, B. Charrow, J. Ledlie, S. J. Teller, and T. Jaakkola. Collaborative future event recommendation. In International Conference on Information and Knowledge Management, pages 819–828, 2010.
23. A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. International Conference on Weblogs and Social Media, pages 70–573, 2011.
24. T. Peng, W. Zuo, and F. He. Svm based adaptive learning method for text classification from positive and unlabeled documents. Knowledge and Information Systems, 16(3):281–301, 2008.
25. J. Sang, T. Mei, J.-T. Sun, C. Xu, and S. Li. Probabilistic sequential pois recommendation via check-in data. In Proceedings of International Conference on Advances in Geographic Information Systems, pages 402–405, 2012.

26. Z. Wang, D. Zhang, X. Zhou, D. Yang, Z. Yu, and Z. Yu. Discovering and profiling overlapping communities in location-based social networks. Systems, Man, and Cybernetics: Systems, IEEE Transactions on, 44(4):499–509, April 2014.
27. B. Xu, A. Chin, and D. Cosley. On how event size and interactivity affect social networks. In CHI Extended Abstracts on Human Factors in Computing Systems, pages 865–870, 2013.
28. D. Yang, D. Zhang, Z. Yu, and Z. Yu. Fine-grained preference-aware location search leveraging crowdsourced digital footprints from lbsns. In Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, pages 479–488, 2013.
29. M. Ye, X. Liu, and W.-C. Lee. Exploring social influence for recommendation: a generative model approach. In Proceedings of ACM International Conference on Research and Development in Information Retrieval, pages 671–680, 2012.
30. Z. Yu, Y. Yang, X. Zhou, Y. Zheng, and X. Xing. Investigating how user's activities in both virtual and physical world impact each other leveraging lbsn data. International Journal of Distributed Sensor Networks, 2014.
31. H. Zhuang, A. Chin, S. Wu, W. Wang, X. Wang, and J. Tang. Inferring geographic coincidence in ephemeral social networks. In Machine Learning and Knowledge Discovery in Databases, pages 613–628. 2012.
32. J. Zhuang, T. Mei, S. C. Hoi, Y.-Q. Xu, and S. Li. When recommendation meets mobile: contextual and personalized recommendation on the go. In Proceedings of the ACM International Conference on Ubiquitous Computing, pages 153–162, 2011.