基于模糊关联规则的微博用户潜在兴趣发现①

牛朝林, 高茂庭

(上海海事大学 信息工程学院, 上海 201306)

摘 要: 针对微博用户兴趣随时间变化的特征, 提出一种基于模糊关联规则的潜在兴趣发现方法(PIDFAR), 利用 LDA 主题模型表达微博主题分布, 通过时间加权的方式计算出用户现在兴趣的主题分布, 进行模糊关联规则挖掘, 得出关联规则集合以表示和发现用户兴趣随时间发生变化的一般规律, 最后根据关联规则集合中关联规则和用户现在兴趣的主题分布来计算相似度, 取相似度较高的关联规则的后项的集合组成用户的潜在兴趣. 实验表明, PIDFAR 方法能够使得用户潜在兴趣的发现过程脱离用户的好友群体限制, 相比基于协同过滤技术的潜在兴趣发现方法明显提高了发现微博用户潜在兴趣的准确率.

关键词: 潜在兴趣; 关联规则; 主题模型; 加权; 微博

Discovering Micro-Blog User's Potential Interests Based on Fuzzy Association Rules

NIU Chao-Lin, GAO Mao-Ting

(College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)

Abstract: According to the specific characteristic of the micro-blog user's interests changing with time, a potential interest discovery method based on fuzzy association rules (PIDFAR) is proposed. It could express micro-blog theme distribution by using LDA topic model, and then use the time weighted method to calculate user's recent interest subject distribution, for mining fuzzy associate rules. That set of association rules to express and discover the general rule of user's interests change over time. Finally, after calculating the similarity between the the associated rule of association rule in the collection and the topic distribution of user's interest, take the high similarity of the latter collection in association rules set to constitute the user's potential interests. Experiments show that the method of PIDFAR can make the process of discover the user's potential interest break away from the limit of user's friends group, and improve the accuracy of the discovery of potential interests of micro-blog users obviously than the traditional method of discovery potential interest based on collaborative filtering technology.

Key words: potential interest; association rule; topic model; weighting; micro-blog

随着 Web 2.0 技术的进步和移动互联网的兴起, 微博最近几年呈现爆炸性的增长. 与传统媒体相比, 微博平台的信息提供者不受时间和空间的限制, 信息涵盖更加宽泛, 有很快的更新速度和传播速度. 由此带来的是海量微博信息, 如何在这些海量信息中发现用户的潜在兴趣已成为一个重要的研究领域.

目前, 微博平台上用户潜在兴趣主要是根据该用户的好友的兴趣来发现的. 但是, 用户的兴趣随着时间的推移会发生改变, 怎样才能及时发现用户伴随时间变化的潜在兴趣这一问题却考虑得相对较少.

针对微博文本的特点, 本文运用主题模型 LDA(Latent Dirichlet Allocation)[1-4]表示和推断微博的主题分布, 进而得到用户的历史兴趣和现在兴趣, 然后根据所有用户的历史兴趣通过关联规则挖掘方法发现微博平台上的用户兴趣之间的关联规则, 以得到随时间变化的一般规律, 再利用这个一般规律和用户现

① 基金项目:国家自然科学基金(61202022);上海海事大学科研项目(201100051)
收稿时间:2015-05-04;收到修改稿时间: 2015-07-02

在兴趣发现他的潜在兴趣信息. 这项工作面临两大挑战: (1)用户现在兴趣的度量; (2)用户兴趣随时间变化的一般规律的发现. 本文主要的工作可归纳为: (1)对微博主题分布进行时间加权来计算用户的现在兴趣; (2)设计一个基于模糊理论的关联规则挖掘算法, 来发现微博平台上用户兴趣转移的一般规律; (3)利用真实数据集验证本文所提方法的准确性和高效性.

本文内容组织如下: 第 1 部分回顾个性化推荐和关联规则挖掘有关技术; 第 2 部分介绍本文方法的整体流程及其各子流程的细节; 第 3 部分利用真实微博数据集验证本文方法的有效性; 第 4 部分总结全文.

1 相关工作

微博的流行吸引了众多学者对用户潜在兴趣发现技术进行研究, 下面分别介绍 LDA 主题模型和应用广泛的用户潜在兴趣发现技术以及关联规则挖掘技术有关的内容.

1.1 LDA 主题模型

LDA 主题模型主要被用来识别大规模文档集中潜在的主题信息, 是一种非监督自主学习技术[2]. 其基本思想是: 每篇文档可以用主题所构成的一个概率多项分布来表示, 而每个主题又可以用文本单词项构成的一个概率多项分布来表示. 所以 LDA 主题模型是一种生成模型, 其生成过程为: 首先从文本的潜在主题分布中抽取一个主题 z, 然后从主题 z 的单词概率多项分布中抽取一个单词 w_i, 重复上述过程直至遍历文档中的每个单词.

该模型有两个参数需要推断: 一个是“文档-主题”分布 θ, 另一个是“主题-单词”分布 φ. 通过学习这两个参数, 就可以知道文档所涉及的主题及其分布. 推断方法主要有变分-EM 算法[2]和现在常用的吉布斯抽样算法[5].

1.2 用户潜在兴趣发现

用户潜在兴趣发现[6,7]通过分析用户历史记录及兴趣偏好等来确定用户以后可能的兴趣信息, 以便依此来向用户提供相应的推荐服务. 其中应用最为广泛的是协同过滤技术[8-10].

协同过滤技术最早是由 Goldberg 等学者[10]于 1992 年提出, 该方法通过利用用户或项目之间的相似性, 然后采用邻域的方法来发现用户潜在的兴趣, 根据用户之间相似性或项目间相似性的不同可以分为基于用户的协同过滤和基于项目的协同过滤. 协同过滤技术不需要利用醒目的内容等信息, 通过利用用户对项目的评分就可以发现用户的潜在兴趣, 故该方法对于像电影、音乐这样非结构化的项目也同样具有良好的适用性, 因此应用非常广泛.

尽管协同过滤技术能发现用户潜在的兴趣, 但这种潜在的兴趣依靠的是该用户的好友的兴趣或者项目之间的相似性, 不能反映出微博用户一般意义上的兴趣变化规律.

1.3 关联规则

R. Agrawal 等人[11,12]首先提出了挖掘顾客交易数据库中项集间的关联规则问题, 其核心方法是基于频繁集理论的递推方法. 此后人们对关联规则的挖掘问题进行了大量研究, 以提高算法挖掘的效率.

关联规则挖掘[13-15]问题就是在事务数据库 D 中找出具有用户给定的最小支持度 minsup 和最小置信度 minconf 的关联规则. 关联规则挖掘问题可以分解为以下两个子问题.

1)找出存在于事务数据库中的所有强项集. 若 X 的支持度 support(X)不小于用户给定的最小支持度 minsup, 则称 X 为强项集.

2)利用强项集生成关联规则. 对于每个强项集 A, 若 $B \subset A$, $B \neq \phi$, 且 $\text{support}(A)/\text{support}(B) \geq \text{minconf}$, 则有关联规则 $B \Rightarrow (A-B)$.

第 2 个子问题比较容易, 目前大多数研究集中在第 1 个子问题上.

2 基于模糊关联规则的用户潜在兴趣发现

针对微博信息的文本特点, 为了发现用户的潜在兴趣信息, 需要解决以下问题: (1)根据用户发布的微博信息来表示和发现该用户的兴趣; (2)发现蕴藏在微博平台上用户兴趣随时间变化的一般规律.

首先, 对于问题(1), 本文采用 LDA 主题模型的方法将微博文本转化为可以度量的主题分布信息, 这样, 用户的兴趣通过对其微博主题分布计算得出, 即像微博一样用一个主题分布来表示. 同时随着时间的推移, 微博平台上用户的兴趣也在发生着变化, 不能平等地看待用户不同时间上的微博信息. 本文采用纵向时间加权的方法[16]来进行计算, 以便突出用户近期的微博信息所蕴含的兴趣, 同时抑制用户时间上距离现在较远兴趣的比重, 这样就最大程度上得到了用户现在时

刻的兴趣, 我们称之为时间加权的用户兴趣.

其次, 对于问题(2), 用户兴趣变化的一般规律隐藏在所有的用户兴趣主题分布信息之中, 而关联规则挖掘则可以发现这一规律. 为了更好地发现潜在规律, 不遗漏相关重要信息, 用户的所有微博都要平等看待, 而不能进行纵向时间加权, 我们称之为时间等权的用户兴趣. 同时, 针对微博的特点, 本文设计一个基于模糊理论的关联规则挖掘算法, 以发现用户兴趣变化的一般规律.

这样, 通过用户现在的兴趣和整个微博平台上的用户兴趣变化的一般规律来达到发现用户潜在兴趣的目的了.

综上考虑, 本文提出基于模糊关联规则的潜在兴趣发现方法(Potential Interest Discovery based on Fuzzy Association Rules, 简称为 PIDFAR), 其整体流程可用图 1 表示.

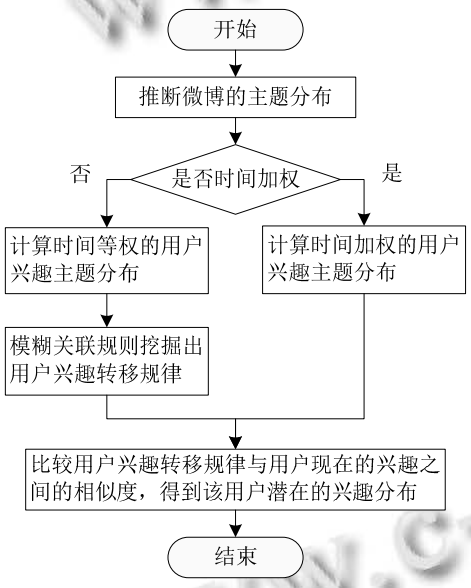

图 1 基于模糊关联规则的微博用户潜在兴趣发现的流程

下面分别详细介绍基于模糊关联规则的微博个性化推荐流程的三个主要部分的处理过程细节. 其中 3.1 节主要介绍 LDA 主题模型推断微博主题分布; 3.2 节主要介绍通过时间加权的方式来计算用户现在兴趣主题分布的内容; 3.3 节主要介绍模糊关联规则挖掘用户兴趣转移规律的内容; 3.4 节主要进行 3.2 节得出的用户现在兴趣主题分布和 3.3 节得出的模糊关联规则的相似度比较, 最终得到用户的潜在兴趣分布.

2.1 推断微博的主题分布

一个用户所发布的微博往往能很好地反映该用户所关心的主题. 因此, 本文采用 LDA 主题模型来推断微博的主题分布, 通过该用户所发布的微博的主题分布来推断出该用户兴趣的主题分布.

2.1.1 微博的主题分布

假设微博主题集合为 $C=\{C_1,C_2,...,C_T\}$, T 表示主题个数, t 为一条微博, 则 $p\{C_i|t\}$ 表示该微博 t 属于主题 C_i 的后验概率, 由这 T 个后验概率组成的向量 $(p\{C_1|t\},p\{C_2|t\},...,p\{C_T|t\})$ 称为微博 t 的主题分布. 则 $p\{C_i|t\}$ 越大, 表明微博 t 属于主题 C_i 的可能性越高.

2.1.2 推断微博的主题分布

对给定的训练数据集进行 LDA 模型学习, 得到每个单词在 T 个主题上的分布 φ, 接下来本文通过 φ 利用吉布斯采样的方法来推断不在训练数据集中的微博的所有单词的主题, 进而推断该微博的主题分布[17].

假设微博 t 由 n 个单词组成, 记为 $\{w_1,w_2,...,w_n\}$, 令随机变量 c_{w_i} 表示单词 w_i 的主题. 在给定其他参数的情形下, 对微博 t 中单词 w_i, $c_{w_i}=j$ (单词 w_i 属于第 j 个主题)的概率计算如下:

$$\begin{aligned} &P(c_{w_i}=j\,|\,C_{t,\neg w_i},t,\varphi,\alpha)\\ &\propto\frac{P(c_{w_i}=j,C_{t,\neg w_i},t\,|\,\varphi,\alpha)}{P(C_{t,\neg w_i},t\,|\,\varphi,\alpha)}\\ &=\frac{n(j,t)+\alpha-1}{n(t)+T\alpha-1}\bullet\varphi_{w_i}^{j} \end{aligned} \tag{1}$$

式(1)中: α 表示主题抽样的先验分布的参数, $\varphi_{w_i}^{j}$ 表示训练得出的单词 w_i 属于第 j 个主题的概率, $C_{t,\neg w_i}$ 表示除了单词 w_i 的主题外在微博 t 中其他所有单词的主题集合, $n(t)$ 表示微博 t 中单词的个数, $n(j,t)$ 表示微博 t 中属于第 j 个主题的单词的个数. 本文的其他公式所涉及的标记意义与(1)式相同.

式(1)表示排除当前词 w_i 的主题, 根据其他词的主题和已知的单词主题概率分布来计算当前词 w_i 的主题的概率. 该式用于吉布斯采样的迭代过程, 直到吉布斯采样收敛.

单词 w_i 的主题分布记为 $V_{w_i}=(v_1,v_2,...,v_T)$, 其中分量 v_j 为标准化上式中的概率, 即:

$$v_j=\frac{P(c_{w_i}=j\,|\,C_{t,\neg i},t,\varphi,\alpha)}{\sum_{j=1}^{T}P(c_{w_i}=j\,|\,C_{t,\neg i},t,\varphi,\alpha)} \tag{2}$$

单词 w_i 的主题从分布 $V_{w_i}=(v_1,v_2,...,v_T)$ 中利用吉布

斯采样得到. 综上, 微博 t 属于第 j 个主题的概率 $\theta_{t,j}$ 估计为:

$$\hat{\theta}_{t,j}=\frac{n(j,t)+\alpha}{n(t)+T\alpha} \tag{3}$$

最后, 微博 t 的主题分布表示为 $\hat{\theta}_t=(\hat{\theta}_{t,1},\hat{\theta}_{t,2},...,\hat{\theta}_{t,T})$.

对于用户发布的微博、转发的微博、评论的微博(评论的微博和内容连接在一起看成是一条微博)都可以依照上述方法分别来推断它们的主题分布, 以形成可以用来计算用户兴趣主题分布的数据集.

在式(1)中, 只有 $\varphi_{w_i}^j$ 是独立于微博 t, 该变量用于衡量用户在微博平台上的词和主题之间的阶段性对应关系. 所以随着时间的变化, 要定期的构建 LDA 训练数据集来更新 $\varphi_{w_i}^j$, 以便及时准确的推断微博的主题.

2.2 用户兴趣主题分布与时间加权的用户兴趣主题分布

假设用户的微博集合为 $\{t_1,t_2,...,t_d\}$, d 表示其中的微博总个数, 则该微博集合所代表的用户的兴趣主题分布可用向量 $(v_1,v_2,...,v_T)$ 表示, T 表示主题个数, 其中:

$$v_i=\frac{1}{d}\sum_{j=1}^{d}p\{C_i\,|\,t_j\} \tag{4}$$

由于随着时间的推移, 用户兴趣会发生不同的转变, 因此有必要对用户兴趣主题分布进行时间加权, 使得用户以前关注而现在不关注的主题权重降低而以前不关注现在关注的主题权重提高, 以更加真实的反映出用户现在的兴趣. 加上时间权重之后可表示成:

$$v_i=\frac{1}{d}\sum_{j=1}^{d}\left(\frac{p\{C_i\,|\,t_j\}}{\text{currTime}-\text{weiboTime}+1}\right) \tag{5}$$

式(5)中, currTime 为现在的时间, weiboTime 为该微博 t_j 发布的时间.

2.3 对用户兴趣主题分布进行模糊关联规则挖掘

在进行关联规则挖掘中用到的用户兴趣表示该用户过去和现在感兴趣的所有主题的分布, 是前述的时间等权的用户兴趣主题分布. 即用户的兴趣表示为 $(v_1,v_2,...,v_T)$, 其中 v_i 如式(4)表示.

针对用户兴趣主题分布的数值型特点, 本文引入模糊理论来解决关联规则挖掘中的数值型属性问题. 设 $I=\{I_1,I_2,...,I_T\}$ 为微博所有用户时间等权的兴趣主题分布构成的数据库 D 的主题属性集. 对每一个主题属性来说, 利用隶属函数将其划分为高、中、低三个模糊属性. 则划分后的数据库 D 转化成模糊数据库 D_f, D_f 的主题属性集为:

$$I_f=\{I_1^{\text{high}},I_1^{\text{middle}},I_1^{\text{low}},...,I_T^{\text{high}},I_T^{\text{middle}},I_T^{\text{low}}\}$$

模糊关联规则是形如 $X\Rightarrow Y$ 的规则, 其中 $X\subseteq I_f$, $Y\subseteq I_f$, $X\cap Y=\phi$, 并且 $X\cup Y$ 不能同时包含同一个属性扩展而来的任何两个项目(例如, 不能同时包含 I_i^{high} 和 I_i^{middle} 等).

对于项目集 $X=\{X_1,X_2,...,X_p\}\subset I_f$, 在模糊数据库 D_f 的第 i 条数据中, X 的支持度为: $Sup_i(X)=X_{1_i}\times X_{2_i}\times...\times X_{p_i}$, 其中, X_{j_i} 是 X_j 在第 i 条数据中的模糊值, 且 $X_{j_i}\in[0,1]$.

X 在整个模糊数据库 D_f 的支持度为:

$$Sup(X)=\sum_{i=1}^{n}Sup_i(X)\Big/\left|D_f\right| \tag{6}$$

模糊关联规则 $X\Rightarrow Y$ 的支持度为:

$$Sup(X\Rightarrow Y)=\frac{\sum_{i=1}^{n}Sup_i(X\cup Y)}{\left|D_f\right|} \tag{7}$$

模糊关联规则 $X\Rightarrow Y$ 的置信度为:

$$Conf(X\Rightarrow Y)=\frac{\sum_{i=1}^{n}Sup_i(X\cup Y)}{Sup(X)} \tag{8}$$

上述两式中 $\left|D_f\right|=n$, 表示模糊数据库 D_f 中记录的总个数.

同时满足给定的最小支持度和最小置信度的模糊关联规则被称为是强模糊关联规则.

模糊关联规则挖掘过程可以分为两个步骤:

1)通过生成的隶属度函数将原始数据库转化为模糊数据库, 利用上述公式来计算各项目集的支持度, 给定最小支持度进而生成满足最小支持度的频繁模糊项目集;

2)由频繁模糊项目集中生成规则, 通过与给定的最小置信度比较得到满足需要的强模糊关联规则.

具体到本文问题的模糊关联规则挖掘算法描述如下:

输入: 主题分布模糊数据库 D_f, 最小支持度 minsup, 最小置信度 minconf

输出: 模糊关联规则集合

符号约定:

f_{ij}: 第 i 个主题在第 j 个用户兴趣主题分布上的隶属度

L_{k-1}: 频繁(k-1)-项集

C_k: 候选 k-项集

C_{k_m}: C_k 中第 m 个成员

S_{k-1}: C_{k_m} 的(*k*-1)-项子集
FOR(*i*=1; *i*<=3T; *i*++) BEGIN
计算主题 *i* 的模糊支持度 $\sup_i = \sum_{j=1}^{n} f_{ij}$
IF $\sup_i \geq$minsup THEN
主题 *i* 加入到频繁 1-项集 L_1 中
END
FOR(*k*=2; $L_{k-1} \neq \Phi$;*k*++) BEGIN
连接生成候选 *k*-项集 C_k
FOR EACH $C_{k_m} \in C_k$ BEGIN
FOR EACH $S_{k-1} \subset C_{k_m}$ BEGIN
IF $S_{k-1} \notin L_{k-1}$ THEN
删除 C_{k_m}
ELSE BEGIN
计算 C_{k_m} 的模糊支持度 $\sup_{k_m} = \sum_{j=1}^{n} \prod_{i=1}^{k} f_{ij}$
IF $\sup_{k_m}$ < minsup THEN
删除 C_{k_m}
END
END
END
频繁 *k*-项集 $L_k = C_k$
END
FOR EACH $l \in L_k$ BEGIN
FOR 主题集 $X \subseteq l$ AND $Y \subseteq l$ BEGIN
IF 规则 $X \Rightarrow Y$ 的置信度 $Conf(X \Rightarrow Y) \geq$minconf THEN
RETURN $X \Rightarrow Y$ AND $Conf(X \Rightarrow Y)$;
END
END

在得到的强模糊关联规则集合中, 去掉冗余规则、不符合事实的规则以及后项是感兴趣程度低的规则, 得到最终的模糊关联规则数据集并对其按照置信度高低进行排序, 这也就形成整个微博平台的用户兴趣变化规律.

2.4 推断用户潜在主题分布

将用户现在兴趣主题分布和关联规则数据集中的每条关联规则的前项进行相似度比较.

首先, 将带有时间加权的用户兴趣主题分布通过隶属函数模糊化, 模糊化后的用户兴趣表示为:

$$V = (v_1^{\text{high}}, v_1^{\text{middle}}, v_1^{\text{low}}, \ldots, v_T^{\text{high}}, v_T^{\text{middle}}, v_T^{\text{low}})$$

其次, 对模糊关联规则中的前项中存在的属性的值设为 1, 加入主题集中该关联规则的前项的属性集的补集, 使其成为 3*T* 维的向量, 并将补集中的属性的值设为 0. 转换之后的模糊关联规则的前项表示为:

$$L^{\text{front}} = (I_1^{\text{high}}, I_1^{\text{middle}}, I_1^{\text{low}}, \ldots, I_T^{\text{high}}, I_T^{\text{middle}}, I_T^{\text{low}})$$

最后, 利用余弦定理来计算向量 V 和向量 L^{front} 的相似度, 即:

$$\text{Sim}(V, L^{\text{front}}) = \frac{\sum_{i=1}^{3T} (V_i \times L^{\text{front}}{}_i)}{\sqrt{\sum_{i=1}^{3T} (V_i)^2} \times \sqrt{\sum_{i=1}^{3T} (L^{\text{front}}{}_i)^2}} \quad (9)$$

式(9)中, V_i 和 $L^{\text{front}}{}_i$ 分别表示向量 V 和向量 L^{front} 中的第 *i* 个分量.

对相似度设定一个阈值 $\min_{\text{sim}}$, 去除那些 $\text{Sim}(V, L^{\text{front}}) < \min_{\text{sim}}$ 的关联规则, 对剩下的关联规则的后项进行统计并标准化. 形成的潜在兴趣主题分布向量表示为:

$$\text{Rmd} = (C_1', C_2' \ldots, C_p')$$

其中, p 表示用户潜在感兴趣的主题的个数, $p = T$, 这里的 $C_i'(i \in [0, T])$ 表示为主题属性.

3 实验分析

3.1 推断用户潜在主题分布

实验在装有 Windows 7 操作系统、2G 内存的 PC 机上进行, 程序代码使用 Java 语言实现. 实验数据来自于浙江大学电子服务研究中心和一些个人抓取的数据, 包括新浪微博上的 5000 名微博认证用户从 2011 年 11 月到 2012 年 4 月以来的微博信息[18-19], 包括发表的具体时间, 共有微博消息 7834690 条, 利用 MySQL 数据库系统管理这些数据. 采用 NLPIR/ICTCLAS2014 分词系统对中文微博信息进行分词处理.

本文采用经验假定的方法, 即不断枚举主题的数目来观察主题模型训练结果的好坏, 比如观察高概率的主题词汇、语义是否一致等, 并最终确定了数据集中的主题个数为 13, 表 1 为实验中各参数的设置.

表 1 实验中的各参数及其默认值

参数	默认值
主题个数 T	13
主题抽样的先验分布参数 α	0.5
对主题贡献最大的词的个数 towords	50
LDA 训练迭代次数 iteration	100
隶属度函数的参数	0.05, 0.07, 0.1
最小支持度 minsup	0.3
最小置信度 minconf	0.8
用户兴趣主题分布与规则的相似度阈值	0.6

minsimrule	
推荐微博的相似度阈值 minsumweibo	0.8

为了验证本文中提出的用户潜在兴趣发现方法的准确性和效率，我们分析部分用户在 2012 年 4 月以后的微博信息，同样利用 LDA 主题模型来推断这些用户后来的微博信息表现出的新的兴趣主题，与本文方法得出的用户的潜在兴趣主题进行比较来考察本文提出的方法，并且采用了一个基于协同过滤的用户潜在兴趣发现方法(CF)[9]进行比较.

对比方法 CF 用于发现用户潜在兴趣时，也采用主题分布表示用户的兴趣. 对于目标用户来说，首先利用上述 2.1 节和 2.2 节介绍的方法计算其好友的时间加权的兴趣主题分布；然后综合考虑所有好友的兴趣主题分布，按照式(10)推算出目标用户的潜在兴趣主题分布 $(C'_1, C'_2, ..., C'_T)$.

$$C'_i = \frac{1}{n}\sum_{k=1}^{n} v_{ik} \tag{10}$$

式(10)中 $k \in [1, n]$ 表示目标用户的好友总个数，$C'_i (i \in [0, T])$ 表示为主题属性.

3.2 评价标准

本文通过以下度量来评价算法的性能:

1)单个目标用户潜在兴趣的准确率，衡量发现的目标用户潜在兴趣主题的结果占该用户之后表现的兴趣主题的比例;

2)单个目标用户潜在兴趣重合率，衡量发现的目标用户潜在兴趣主题占该用户已经感兴趣的主题的比例;

3)平均准确率，所有用户潜在兴趣主题发现的准确率的平均值;

4)平均重合率，所有用户的潜在兴趣主题发现的重合率的平均值.

3.3 实验结果与分析

对上述微博数据集进行实验，首先任选一个用户为目标用户，分别通过变动该用户关注的好友的微博数量和该用户关注的好友的数量，以潜在兴趣准确率、潜在兴趣重合率两个指标来衡量本文方法 PIDFAR 和对比方法 CF 的有效性和稳定性；然后针对所有用户，通过变动用户的数量，以平均准确率和平均重合率来衡量上述两个方法的性能，结果如图 2-图 7 所示.

由图 2 可以看出，随着目标用户 A 关注的好友微博数量的变化，CF 方法的潜在兴趣准确率有很大的变动，而 PIDFAR 方法的潜在兴趣准确率却一直很稳定. 这主要是因为目标用户 A 关注的好友微博数量的变化导致了他的潜在兴趣信息来源的变化，这直接就影响了他的潜在兴趣发现结果；而 PIDFAR 方法中目标用户 A 的潜在兴趣信息来源是整个平台上的用户兴趣随时间变化的一般规则，少量用户的微博数量变化可能会改变这些用户的兴趣主题分布，却不会导致整体规则的变化，因而也就很稳定.

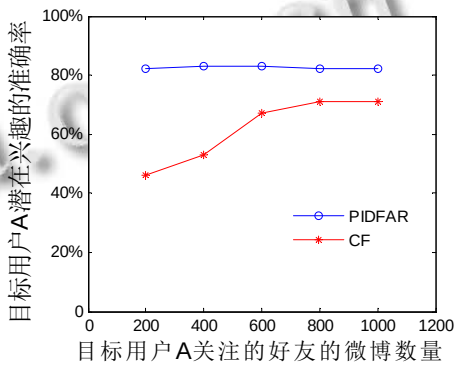

图 2 目标用户相对于好友微博数量的潜在兴趣准确率

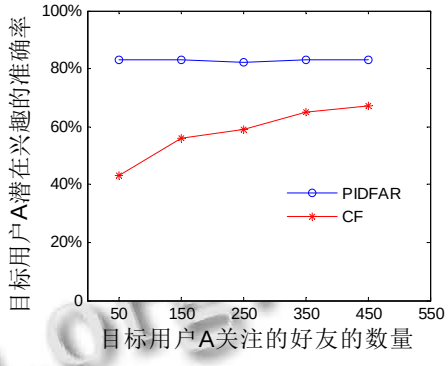

图 3 目标用户相对于好友数量的潜在兴趣准确率

从图 3 可以看出，随着目标用户 A 关注的好友数量的增加, CF 方法的潜在兴趣准确率也随之出现增加，原因在于关注的好友数量的增加能够提供给目标用户 A 更多的新的兴趣发现，从而发现更多符合目标用户 A 的潜在兴趣；而 PIDFAR 方法中，相对于所有用户数量来讲，目标用户 A 关注的好友数量很小，少量用户的加入并不能直接改变规则，因而 PIDFAR 方法表现很稳定.

从图 4 和图 5 不难发现，随着目标用户 A 关注的好友的微博数量或者好友数量的增长，采用 PIDFAR 方法得出的用户的潜兴趣重合率一直维持在较低的位置，比较稳定，但是采用 CF 方法得出的潜在兴趣重合

率则一直在较高的位置变动. 这说明本文提出的算法能够发现更多的用户未真正接触的新的兴趣信息, 并且具有较好的稳定性.

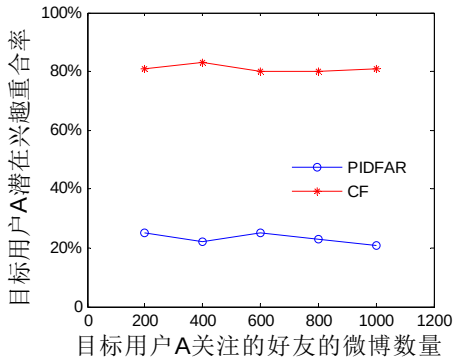

图 4 目标用户相对于好友微博数量的潜在兴趣重合率

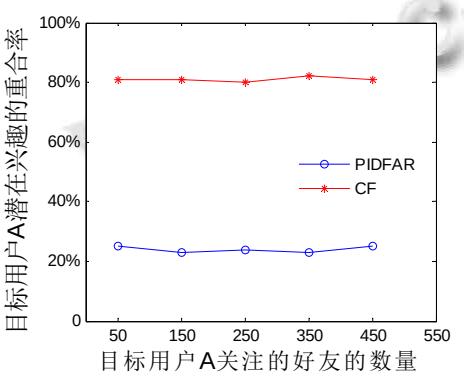

图 5 目标用户相对于好友数量的潜在兴趣重合率

图 6 通过变动用户数量来展示 PIDFAR 方法和 CF 方法的平均准确率. 不难看出, 当用户数量很少的时候, 本文提出的 PIDFAR 方法平均准确率没有 CF 方法高, 这主要是因为用户数量少导致挖掘出的关联规则不能很好地反映出兴趣变化的普遍规律, 而 CF 方法依照好友之间的兴趣来发现目标用户的潜在兴趣则可以得到相对较高的准确率. 不过随着用户数量的增加, 相应的微博数量也就大量增加, 这样挖掘出的关联规则就会更加接近真实规律, 发现的用户潜在兴趣的准确率就会很快增加, 虽然 CF 方法也会随着好友的增加, 发现目标用户的潜在兴趣准确率增加, 但是效果明显不及 PIDFAR 方法. 说明本文提出的 PIDFAR 方法能够进一步提高发现用户潜在兴趣的准确率.

图 7 通过变动用户数量来展示 PIDFAR 方法和 CF 的平均重合率. 由图 7 容易看出, PIDFAR 方法推荐的潜在兴趣与目标用户现在的兴趣重合率要小很多, 说明本文提出的方法能够更容易的发现用户的大量潜在兴趣.

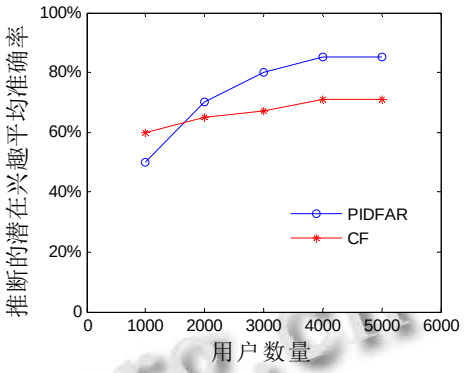

图 6 潜在兴趣的平均准确率

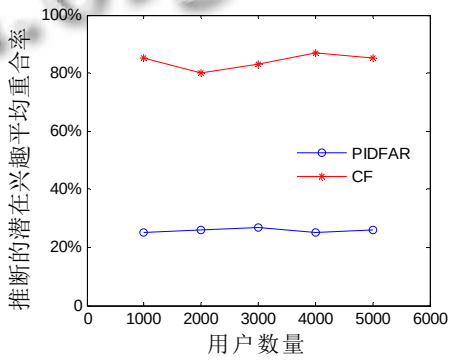

图 7 潜在兴趣的平均重合率

4 结语

发现用户的潜在兴趣, 以向其推荐符合该潜在兴趣的微博信息, 对于微博平台上的个性化推荐来说非常重要. 传统的协同过滤技术只能根据用户的好友的兴趣来判断该用户的潜在兴趣, 具有很大的局限性, 不能很好地利用兴趣随时间转变的一般规律. 本文的方法旨在发现微博平台上用户潜在兴趣随时间转移的一般规律, 进而根据这一规律和用户现在的兴趣来判断用户的潜在兴趣.

针对微博的特点, 本文采用 LDA 主题模型来表示和推断用户的兴趣和微博的内容, 利用模糊化的关联规则进行兴趣随时间转变的一般规律, 并且采用时间加权的方式使所发现的用户现在兴趣主题分布更加接近实际情况, 在很大程度上提高了发现潜在兴趣的准确率, 同时又降低了发现潜在兴趣的重合率.

参考文献

1 Bei DM, Lafferty J. Text Mining: Theory and Applications. Chapter Topic Models, Taylor and Francis, London, 2009.

2 Bei DM, Ng AY, Jadan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3(4/5): 993–1022.

3 Steyvers M, Griffiths T. Probabilistic topic models. Latent Semantic Analysis: A Road to Meaning, Laurence Erlbaum, 2005.

4 Weng JS, Lim EP, Jing J, el al. TwitterRank: finding topic-sensitive influential twitters. Proc. of the 3th ACM International Conference on Web Search and Data Mining. New York City, NY, USA. 2010. 261–270.

5 Wang XR, McCallum A. Topics over time: a non-markov continuous-time model of topical trends. Proc. of the 12th ACM SIGKDD International Coference on Knowledge Discovery and Data Mining. Philadelphia, USA. 2006. 424–433.

6 Varian R. Recommender systems. Communication of the ACM, 1997, 40(3): 56–58.

7 Lawrence RD, Almasi GS, Kotl YV, et al. Personalization of supermarket product recopmmendations. IBM Research Report, 2000.

8 Resnick P, Iacovou N, Suchak M, et al. Grouplens: an open architecture for collaborative filter of netnews. Proc. of the Conference on Computer Supported Cooperative Work. Chapel Hill, NC. 1994. 175–186.

9 Koren Y, Bell R. Advances in collaborative filtering. Recommender Systems Handbook. Springer US, 2011: 145–186.

10 GoldBerg D, Nichols D, Oki BM, et al. Using collaborative filtering to weave an information apestry. Communications of the ACM, 1992, 35(12): 61–70.

11 Agraqal R, Srikant R. Mining association rules between sets of items in large database. Probabilistic ACM SIGMOD International Conference Management of Data. Washington DC. 1993. 207–216.

12 Agraqal R, Srikant R. Fast algorithms for mining association rules. Probabilistic 20th International Conference Very Large Database. Santiago, Chile. 1994. 487–499.

13 Park JS, Chen MS, Yu PS. An effective hash-based algorithm for mining association rules. Probabilistic ACM SIGMOD International Conference Management of data. San Jose. 1995. 175–186.

14 Savasere A, Omiecinski E, Navathe S. An effective algorithm for mining association rules in large database. Probabilistic 21th International Conference Management of Data. San Jose. 1995. 175–186.

15 Cai CH, Fu WC, Cheng CH, el al. Mining association rules with eighted items. IEEE International Database Engineering and Applications Symposium. Cardiff. 1998.

16 陆建江.加权关联规则挖掘算法的研究.计算机研究与发展,2002,39(10):1281–1286.

17 高明,金澈清,钱卫宁,王晓玲,周傲英.面向微博系统的实时个性化推荐.计算机学报,2014,37(4):963–975.

18 浙江大学电子服务研究中心.微博消息数据集. http://www.datatang.com/data/444482/.

19 微博用户信息及微博数据集.http://www.datatang.com/data/44220/.