一种基于遗忘规律的微博用户关注度计算方法*

崔瑞飞，于洪涛，张 考
(国家数字交换系统工程技术研究中心，河南 郑州 450002)

摘 要：准确把握微博用户所关心的领域对于目标营销和个性化推荐具有重要意义。利用用户历史关注信息预测用户将来可能关注的领域，是一种解决该问题的有效思路。此过程中，若用户关注某一领域的信息越多说明用户将来关注该领域的可能性越大；关注某信息的时间距离当前时间越近，则该信息的参考价值越大。这两点与人类对知识重复学习强化记忆和逐渐遗忘的过程非常相似，因此以人类遗忘规律为基础，提出了一种微博用户关注度计算方法。真实数据集上的实验表明，该方法能够很好地预测用户关注度分布情况。

关键词：微博；关注度；重复学习；遗忘规律
中图分类号：TP393 **文献标识码**：A **文章编号**：0258-7998(2014)08-0112-04

Micro-blog user attention degree method based on forgetting law

Cui Ruifei, Yu Hongtao, Zhang Kao
(National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450002, China)

Abstract: It is significant for targeted marketing and personalized recommendation to grasp accurately the interest of micro-blog user. It is a valid idea to predict current attention by history information of the user. In this process, the higher attention degree to the field is, the more a user pay attention to a concern field, and the stronger the reference value of the message is, the nearer the distance between user attention time to a message and current time is. These two points can be regarded as the process of human's repeated learning and forgetting gradually. Therefore, this paper proposed a method of calculation micro-blog user's attention based on forgetting law. Experiments show that the method can predict micro-blog user's attention better.

Key words: micro-blog; attention degree; repeated learning; forgetting law

微博用户关注是指用户所关心的领域知识或信息，有时也指用户关注的其他用户。在本文中，如无特殊说明，均指前者。

提到用户关注，人们往往会联想起用户兴趣。的确，两者有一定的联系，但是也有着显著的不同。用户关注是客观的，是用户兴趣叠加了社会影响(如朋友推荐、社会事件)之后的外部呈现，如式(1)中的函数 H；而用户兴趣是主观的、相对稳定的，可以从用户关注中挖掘出用户兴趣，如式(1)中的函数 G。

$$\begin{cases} E=H(I,S) \\ I=G(E) \end{cases} \tag{1}$$

其中，I 为用户兴趣，S 为社会影响，E 为用户关注。

由于微博诞生得较晚，所以有关微博用户关注和用户兴趣的研究较少，主要集中在用户兴趣上。WENG J[1]等人使用微博中的标签来构建用户兴趣，指出用户标签是微博用户兴趣的重要来源。WU W[2]等人从用户的微博内容中抽取若干关键词，作为该用户的个性化特征，来描述用户的兴趣。HONG L[3]等人把每个用户发布的所有微博看作一个大的文档，然后扩展了标准的 LDA 模型来发现用户潜在感兴趣的主题。Xu Zhiheng[4]等人对 LDA 模型作了改进，提出了更适合微博的 Twitter-User 模型。

以上研究的思路都是从用户关注中挖掘用户兴趣，进而指导目标营销和个性化推荐。但本文认为，通过用户历史关注信息直接预测用户当前关注的领域，进而指导目标营销和个性化推荐是一种更好的思路。该思路的主要优点体现在：(1)不具体关心社会影响(S)的作用；(2)用户关注易于获取和评价，而主观上的兴趣并不是很容易把握，且评价较困难。本文将基于此思路来获取用户的当前关注。

* 基金项目：国家"863"计划资助项目(2011AA010603，2011AA010605)

1 微博用户关注度分析

1.1 关注度的表示

本文借鉴数据分类的思想，将关注信息分成若干个类别，这些类别互不重叠并且所有类别的并集为整个信息空间；然后计算用户对每个类别的关注度，得到用户的关注度分布向量。

对于特定的类别集合 $\boldsymbol{C}=\{c_1,c_2,\cdots,c_N\}$，用户 i 的关注度分布向量表示为：

$$\boldsymbol{E}_i=\{e_{i1},e_{i2},\cdots,e_{iN}\} \tag{2}$$

其中，e_{ij} 为第 i 个用户对第 j 个类别的关注程度，且 $\sum_{j=1}^{N} e_{ij}=1$。

为了得到集合 $\boldsymbol{C}$，本文参考了新浪、搜狐、腾讯和网易四大门户网站的板块划分情况，最终得到 $\boldsymbol{C}=$｛军事，经济，科技，体育，娱乐，教育，政治，医药，交通，环境｝，共 10 个项集。

1.2 用户关注与遗忘规律

本文的基本思路为根据用户历史信息预测当前关注，这个过程刻划为式(3)：

$$\boldsymbol{E}_t=F(W_{t-1},W_{t-2},\cdots,W_{t-n}) \tag{3}$$

其中，t 为当前时间，$\boldsymbol{E}_t$ 代表用户当前关注，W_{t-1}、W_{t-2}、…、W_{t-n} 为用户历史信息。在这个预测过程中，需要考虑两个关键点：

(1)用户历史信息距离当前时间越近参考价值越大(近因效应)，反之则越小；换句话说，信息的参考价值随时间推移逐渐衰减，这与人类的记忆随时间不断衰减的过程非常类似。

(2)用户对某类内容关注越多，则对该类别关注度越高。这就好比人们重复学习某一知识，重复学习次数越多，则记忆量越大，直到记忆稳定。

通过以上分析，本文基于人类遗忘规律，提出一种微博用户关注度计算方法。

2 基于遗忘规律的关注度计算方法

2.1 遗忘规律量化函数

德国心理学家艾宾浩斯揭示了人类的遗忘规律，指出了记忆时效随时间的变化特征[5]，曾东红[6]等人对该特征进行了数学分析，并采用负指数曲线拟合了此规律，在记忆学中该曲线又称为遗忘曲线，其量化函数为式(4)。

$$p(t,k)=p_0\mathrm{e}^{-kt},\ t\in(0,\infty) \tag{4}$$

其中，p_0 为初始记忆量，k 为遗忘速率，它是反映遗忘曲线衰减差异的主要参数。

2.2 关注度计算方法

遗忘曲线体现了单个时间段内记忆量随时间的变化，本小节将以此为基础提出一种微博用户关注度多阶段量化方法。

对同一类别 c，重复学习指用户多次关注该类别的微博，学习的时刻即为关注时刻。令 t_1、t_2、t_3 表示 3 次相邻重复学习的时刻，用户关注度的变化过程如图 1 所示。

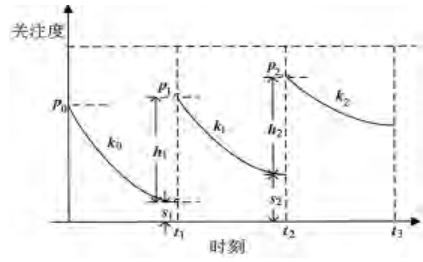

图 1 关注度多阶段量化方法示意图

可以看出，关注时刻把整个过程分成了多个阶段，每个阶段都是一个新的遗忘过程，不同阶段的区别在于阶段初始值与遗忘速率不同。如果能确定任意阶段的初始值 p_n 和遗忘速率 k_n，则可以实现用户关注度的实时度量。下面分别介绍 p_n 和 k_n 的确定策略。

(1)阶段初始值

从图 1 中可以看出，第 n 个阶段初始值 p_n 是 s_n 与 h_n 叠加的结果，s_n 为上一阶段剩余量，h_n 为重复学习带来的新叠加量。容易确定 s_n 为：

$$s_n=p_{n-1}\mathrm{e}^{-k_{n-1}(t_n-t_{n-1})} \tag{5}$$

下面分析如何确定 h_n。人们重复学习同一知识时，每次获得的新记忆量并不是一致的；随着学习次数的增加，对某知识的记忆总量会不断增加，这个过程不是线性的，而是逐渐趋于平缓并最终收敛于某最大值；所以，重复学习次数越多，每次获得的新记忆量会不断下降。因此，可以用负指数曲线来描述 h_n 与 n 的关系，如式(6)所示。

$$h_n=p_0\mathrm{e}^{-n} \tag{6}$$

综上，第 n 个阶段初始值 p_n 可以表示为：

$$p_n=p_{n-1}\mathrm{e}^{-k_{n-1}(t_n-t_{n-1})}+p_0\mathrm{e}^{-n} \tag{7}$$

(2) 遗忘速率

令 t_{n-1} 和 t_n 为重复学习的任意相邻时刻，k_{n-1} 为遗忘曲线从 t_{n-1} 到 t_n 的遗忘速率，则对于类别 c，该遗忘阶段的量化函数为：

$$p_c(t,k_{n-1})=p_{n-1}\mathrm{e}^{-k_{n-1}(t-t_{n-1})},\ t\in(t_{n-1},t_n) \tag{8}$$

为了分析相邻阶段遗忘速率的关系，将后一阶段遗忘曲线先向 y 轴反向平移，再向 x 轴反向平移，使之与前一阶段遗忘曲线具有共同起点，该操作过程如图 2 所示。

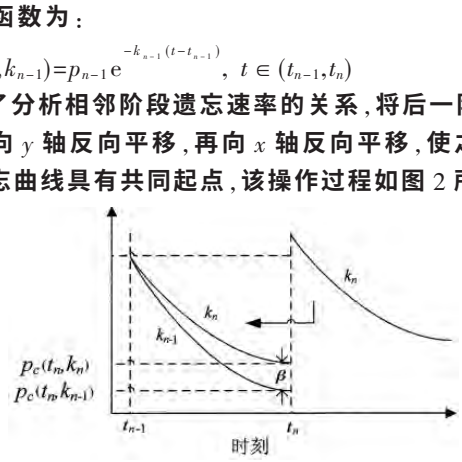

图 2 遗忘速率调整分析图

其中，β 为 t_n-t_{n-1} 时间段内相邻两条遗忘曲线的衰退差异，可表示为：

$$\beta=p_{n-1}(\mathrm{e}^{-k_n(t_n-t_{n-1})}-\mathrm{e}^{-k_{n-1}(t_n-t_{n-1})}) \tag{9}$$

由式(9)可得 k_n 和 k_{n-1} 之间的关系：

$$k_n=\frac{\ln(e^{-k_{n-1}(t_n-t_{n-1})}+\beta/p_{n-1})}{-(t_n-t_{n-1})} \tag{10}$$

将 β 的取值上限 $p_{n-1}(1-p_c(t_n,k_{n-1}))$ 划分为 θ 个线段，则 β 的取值可表示为式(11)。θ 称为惰性因子，它反映了一次重复学习对遗忘速率调整的程度，若两个相邻重复学习时刻的间隔越大，说明用户越懒惰，θ 的取值就越大。

$$\beta=\frac{p_{n-1}(1-p_c(t_n,k_{n-1}))}{\theta} \tag{11}$$

将式(11)代入式(10)，得到 k_n 的计算公式：

$$k_n=\frac{\ln(1+(\theta-1)e^{-k_{n-1}(t_n-t_{n-1})})-\ln\theta}{-(t_n-t_{n-1})} \tag{12}$$

从式(12)中可以看出，如果已知 k_{n-1}、惰性因子 θ 和时间间隔 t_n-t_{n-1}，则可以确定 k_n 的值，进而可计算用户在任意时刻对某类别的关注度。

2.3 用户关注度分布向量

本小节设计了微博用户关注度算法 MUAD (Micro-Blog User Attention Degree Algorithm)，用来计算用户的关注度分布向量。其伪代码如下：

输入：用户 U 和时刻 t。

输出：用户 U 在时刻 t 的关注度分布向量 $\boldsymbol{E}(t)$。

初始化：k_0，p_0，θ_0，类别列表 $C(N)$ 和时间阈值 T；

算法：

```
AllTweets ← 用户 U 在(t-T, t)内关注的微博集合;
Category [i] ← TextClassify(AllTweets), i ∈ (1,N);
FOR i = 1:N
    Category [i] ← Category [i]按时间升序排列;
    M ← Category [i]中元素的个数;
    t0 ← Category [i]的第一个元素;
    FOR j =1:M-1
        tj= Category [i]的第 j 个元素;
        pj=pj-1 e^(-kj-1(tj-tj-1)) + p0 e^(-j);
        kj=(ln(1+(θ-1)e^(-kj-1(tj-tj-1)))-lnθ)/(-(tj-tj-1));
        θj=(ti-ti-1)×θ0; (ti-ti-1 以天为单位)
    END FOR;
    计算得到 ei(t);
END FOR;
E(t) ← 对{e1(t), e2(t),…,eN(t)}归一化
```

其中，时间阈值 T 的含义是需要参照的历史信息的时间跨度，Category [i]表示属于类别 c_i 的所有微博的关注时间集合。TextClassify()表示文本分类算法，HONG L[3] 等人根据微博文本的特点，提出了一种基于 LDA 模型的分类方法(USER Scheme)，很好地实现了微博文本的分类，其准确率可达到 82.92%，本文使用此方法来实现微博文本的分类。

通过 MUAD 算法，得到了用户在时刻 t 的关注度分布向量 $\boldsymbol{E}(t)$。

3 实验

3.1 实验准备

3.1.1 实验数据

本文实验数据来源于腾讯微博开放平台，原始数据包括 2 000 个用户信息以及这些用户于 2013 年 8 月至 10 月发布、评论、收藏和赞(都称用户关注)的 1 214 980 条微博信息。根据本实验需求，在这些数据中剔除了三个月内关注微博信息数不足阈值 A 的用户及其关注的信息(这类用户非常不活跃，没有研究意义)，最后得到了 1 636 个用户信息和 1 207 431 条微博信息。

3.1.2 实验参数的设定

需设定的参数包括：初始遗忘速率 k_0、首阶段初始值 p_0 和惰性因子基数 θ_0，具体数值如表 1 所示。

表 1 参数设定

参数名称	参数意义	取值
k_0	初始遗忘速率	1.0
p_0	首阶段初始值	1.0
θ_0	惰性因子基数	10

时间阈值 T 的选取会对算法的预测效果产生一定的影响，本文将通过实验选取最优的 T 值。

3.1.3 评价指标

本文使用平均绝对误差 MAE 来评价预测结果的准确性。其计算方法如式(13)所示。

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|Y_{\text{实际 }i}-Y_{\text{预测 }i}| \tag{13}$$

3.2 实验结果与分析

本实验共获取了 3 个月的数据，实验利用前 70 天的数据来预测后 20 天的用户关注情况，然后与后 20 天的实际数据进行对比，得出 MAE 的值。

3.2.1 时间阈值 T 对算法性能的影响

图 3 为算法 MAE 值随 T 的变化趋势。当 T 在 30~45 天这一范围时，MAE 值基本稳定，这是因为偶然性导致的问题逐渐被剔除；当 T>45 时，MAE 值略有上升，这是由于随着参考信息的增加，用户对各个类别的“记忆量”都趋于收敛使得算法区分度下降的缘故。从图 3 中还可

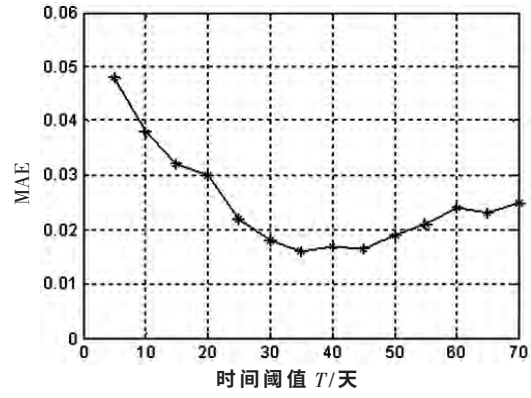

图 3 时间阈值 T 对算法性能的影响

以看出，T 取 35 时 MAE 的值最优(下文 T 取 35)。

3.2.2 算法对比实验

本实验选取标准 LDA 模型[3]和 Twitter-User[4](简称 T-U)模型作为本文的对比算法。此外，当前在预测用户关注时，一些场合也采用简单统计的方法(称为 General 方法)，该方法统计过去一段时间用户关注分布情况当作用户当前关注，本文也将此方法作对比。

图 4 为 3 种对比算法和本文算法的预测误差平均值对比图。从图中可以看出，MUAD 算法的预测误差都要低于其他 3 种方法。

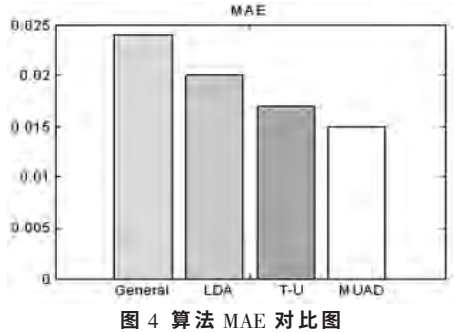

图 4 算法 MAE 对比图

MUAD 算法优于 LDA 和 T-U，是因为 MUAD 直接通过用户历史关注信息预测未来关注的领域，而 LDA 和 T-U 更侧重于用户主观上的兴趣。

MUAD 算法优于 General 方法，是因为相比于 General 方法，MUAD 考虑的因素更全面更合理，主要体现在：(1)考虑了近因效应，距离当前时间越近，影响效果越明显；(2)人类的记忆是会遗忘的，而不是一经学习永久存储的；(3)随着对同一知识记忆量的增加，每次重复学习产生的新记忆量不断下降，总的记忆量逐渐收敛。

本文首先对微博用户关注信息分成了 10 个类别；然后利用用户历史关注信息，借鉴人类遗忘规律的相关知识，提出了一种用户关注度计算方法；最后设计了 MUAD 算法，得到了用户的关注度分布向量。实验表明，该方法能够准确地发现微博用户的关注分布情况，具有较强的实用性。

参考文献

[1] WENG J, LIM E P, HE Q, et al. What do people want in microblogs Measuring interestingness of hashtags in twitter[C]. Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010:1121-1126.

[2] WU W, ZHANG B, OSTENDORF M. Automatic generation of personalized annotation tags for twitter users[C]. 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.

[3] HONG L, DAVISON B D. Empirical study of topic modeling in twitter[C]. Proceedings of the First Workshop on Social Media Analytics. ACM, 2010:80-88.

[4] Xu Zhiheng, Long Ru, Liang Xiang, et al. Discovering user interest on twitter with a modified author-topic model[C]. International Conferences on Web Intelligence and Intelligent Agent Technology. IEEE/WIC/ACM, 2011:422-429.

[5] HERMANN E. Memory:a contribution to experimental psychology[EB/OL].(2011-12-09)[2014-01-16].http://psy.ed.asu.edu/~classics/Ebbinghaus/index.ht-m.

[6] Zeng Donghong, Wang Tao, Yan Shuifa, et al. A collaborative filtering recommendation algorithm based on exponential forgetting function[J]. Science Mosaic,2013(7):10-15.

(收稿日期：2014-03-16)

作者简介：

崔瑞飞，男，1989 年生，硕士研究生，主要研究方向：通信与信息系统。

于洪涛，男，1970 年生，教授，主要研究方向：通信与信息系统。

张考，男，1989 年生，硕士研究生，主要研究方向：通信与信息系统。