一种基于用户动态兴趣和社交网络的微博推荐方法

陈 杰[1],刘学军[1],李 斌[1],章 玮[2]
(1. 南京工业大学计算机科学与技术学院,江苏南京 211816;2. 中国人民解放军 73677 部队,江苏南京 210016)

摘 要: 针对为微博用户推荐符合其兴趣取向的个性化微博信息的问题,结合 LDA 主题模型,提出了一种基于用户动态兴趣和社交网络(DISN)的微博推荐方法. DISN 方法首先引入时间函数,推断出用户的兴趣向量,通过对新发布的微博数据内容进行聚类分组,以用户兴趣向量筛选与用户最匹配的分组,随后以网格索引的形式对选定的分组中微博进行查询,计算微博发布者被目标用户关注的可能性并进行排序,最终形成推荐列表. 实验验证了 DISN 方法较之传统方法更具有效性和高效性.

关键词: 动态兴趣;社交网络;LDA;网格查询;个性化推荐;微博
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2017)04-0898-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.04.019

Personalized Microblogging Recommendation Based on Dynamic Interests and Social Networking of Users

CHEN Jie[1], LIU Xue-jun[1], LI Bin[1], ZHANG Wei[2]
(1. *Department of Computer Science and Technology, Nanjing Tech University, Nanjing, Jiangsu* 211816, *China*;
2. 73677 *PLA Troops, Nanjing, Jiangsu* 210016, *China*)

Abstract: To recommend useful microblogs that match users' interests and likes effectively, an approach in which the dynamic interests and social networking (DISN) of users are seamlessly integrated based on LDA model is proposed. The approach infers the interest vector of users better by using time function and groups the new published microblogs by clustering method and gets the best matching groups with users' interest vector. Then DISN traverses the selected groups by grid querying approach and matches the microblogs with publishers' probabilities of being followed and sorts the result. Finally the personalized microblogging recommendation is achieved. Experimental results show that DISN is more effective and efficient than the traditional models.

Key words: dynamic interests; social networking; LDA; grid querying; personalized recommendation; microblog

1 引言

随着 web 2.0 技术的不断发展与成熟,新兴社交媒体,如 tweet、新浪微博等,已逐渐发展成为人们传播、分享信息的重要平台,并吸引了数以亿计的用户. 根据新浪微博《2014 年微博用户发展报告》,新浪微博月活跃用户数已达到 1.67 亿. 如此庞大的用户群,每日将产生大量的数据,用户很容易被海量的数据所淹没. 因而如何从海量的数据中选取并推荐用户感兴趣的内容变得越来越重要. 然而,用户的兴趣往往随着时间的推移而发生着变化;此外,微博推荐在一定程度上也是向用户推荐与其志趣相投的用户. 因此,实时而有效的个性化推荐显得尤为重要.

微博文本作为微博的信息载体很好地反映了用户的兴趣取向以及变化趋势,因而传统的微博推荐方法大都采用基于内容的推荐方法. 基于内容的推荐方法首要解决的问题是对用户进行建模,传统的方法是对用户发布的所有历史数据无偏重地进行处理,从而推断用户兴趣,如高明等提出的基于 LDA 模型以及滑动窗口的个性化微博推荐方法[1],Otsuka E 等人提出了基

收稿日期:2015-09-07;修回日期:2015-11-12;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61203072);江苏省重点研发计划(社会发展)(No.BE2015697)

于 TF-IDF 的微博话题推荐方法 HF-IHU[2],此类方法未考虑时间因素,因而无法反映用户的兴趣变化;此外,微博推荐在一定程度上,更希望用户能够关注微博的发布者,甚至与其成为好友,而传统的微博推荐方法,往往只关注用户的信息需求[1~3],侧重于内容的推荐,而未能考虑用户的社交需求.

针对以上问题,论文提出了基于用户动态兴趣和社交网络(DISN)的微博推荐方法. 该方法首先将新发布的微博数据基于其内容进行分层聚类,从而对微博进行分组. 利用 LDA(Latent Dirichlet Allocation)主题模型推断分组后各组内微博以及用户发布的历史微博的主题分布,引入时间函数,调整用户近期及远期微博数据的权重,从而动态地推断用户的兴趣取向,并以此为基础,选取与用户兴趣取向最匹配的微博分组. 进而以网格索引的方式遍历选定微博组中的各条微博,并依据用户与微博发布者的兴趣相似度和用户对微博发布者的信任度,计算微博发布者被用户所关注的可能性,并根据可能性的高低以及微博的热度进行排序,最终生成合适的微博推荐列表. 系统框架图如图 1 所示.

论文主要贡献如下:(1)引入时间函数,更好地提取用户兴趣取向,有效地解决了用户兴趣随时间推移而发生变化的问题;(2)通过对用户兴趣和用户社交网络的无缝整合,在推荐用户感兴趣内容的同时,增加用户关注微博发布者的可能性,从而更好地提升推荐效果;(3)提出了一个两阶段的微博推荐模型,第一阶段根据用户的兴趣取向选取与其匹配的微博分组,第二阶段遍历选定分组中微博,根据微博发布者被用户关注可能性的大小,选定最终推荐给用户的微博;(4)实验验证了该方法的有效性和高效性.

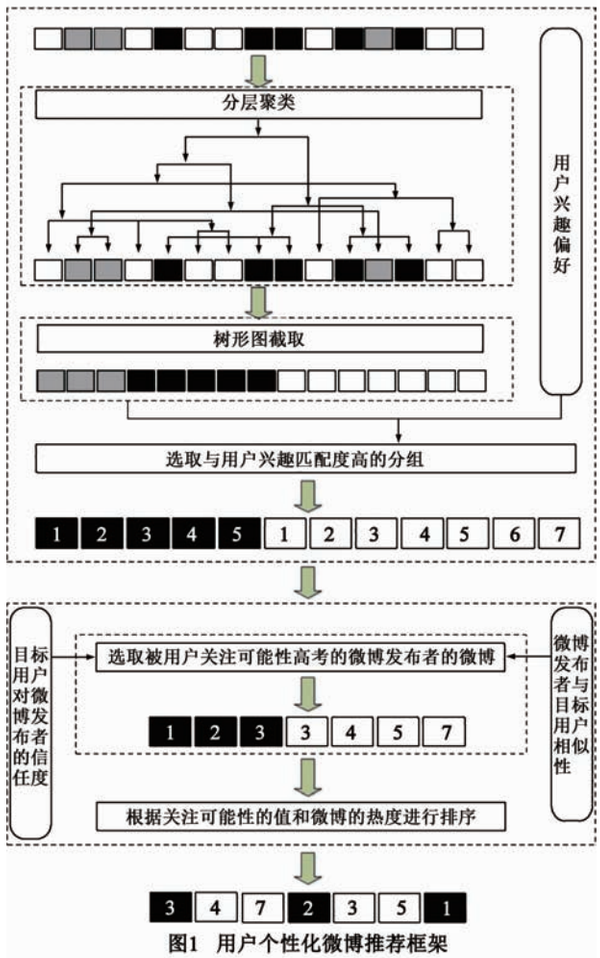

图1　用户个性化微博推荐框架

2　推断用户兴趣取向

2.1　LDA 主题模型

LDA 模型[4]由 Blei D M 在 2003 年提出,是一种非监督机器学习技术,可以用来识别大规模文档集(document collection)或语料库(corpus)中潜藏的主题信息. 其模型图如图 2 所示.

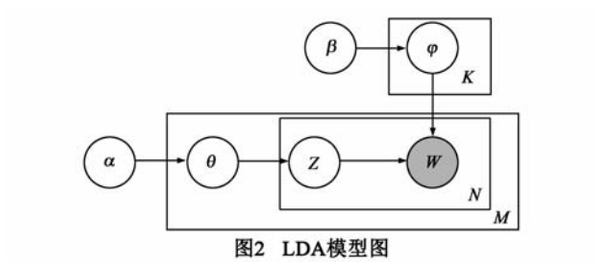

图2　LDA模型图

当给定一个有 M 篇文档的文档集合 D,共包含 K 个主题 z,N 个单词 w. 其中 $\boldsymbol{\alpha}$ 与 $\boldsymbol{\beta}$ 是服从狄力克雷分布的语料级别的参数,$\boldsymbol{\alpha}$ 是 $p(\boldsymbol{\theta})$ 的向量参数,用于生成一个主题 $\boldsymbol{\theta}$ 向量,$\boldsymbol{\beta}$ 是各个主题对应的单词概率分布矩阵 $p(w|z)$. 则文本的生成过程可描述如下:

(1)对每个文档 $d \in D$,从狄利克雷分布 Dir ($\boldsymbol{\alpha}$)中取样生成文档 d 的主题分布 $\boldsymbol{\theta}$;

(2)从主题的多项式分布 $\boldsymbol{\theta}$ 中取样生成文档 d 第 n 个单词的主题 z_n;

(3)从狄利克雷分布 Dir ($\boldsymbol{\beta}$)中取样生成主题 z_n 的词分布 φ_{z_n};

(4)从词的多项式分布 φ_{z_n} 中采样最终生成词 w_n.

M 个文档集合 D 用 LDA 生成的概率为:

$$\prod_{m=1}^{M}\int p(\theta_m|\boldsymbol{\alpha})\left(\prod_{n=1}^{N_m}p(z_n|\theta_m)p(\varphi_{z_n}|\boldsymbol{\beta})p(w_n|\theta_{z_n})\right)\mathrm{d}\theta_m \tag{1}$$

而其中比较重要的隐含变量:文档-主题分布 $\boldsymbol{\theta}$ 和主题-词分布可由 Gibbs Sampling 方法获得.

2.2　用户兴趣分布

论文通过使用 LDA 主题模型提取用户所发布的微博的主题分布,从而推断用户的兴趣取向.

考虑到用户的兴趣往往随着时间的推移而发生改变,论文将用户发布的历史微博按照时间段进行分割. 给定用户 U 在 T 时间段内所发布的微博集合 M,则 $M=\{M_{t_1},M_{t_2},\cdots,M_{t_n}\}$,其中 M_{t_i} 表示最近时间段内所有微博组成的文本. 由于单个微博内容较少,因而将 t_i 时间段内的所有微博内容组成一个文本,减少系统对文件的读取次数,提高系统效率.

令 T 表示预先设定的 K 个主题的集合,$T=\{T_1,T_2,\cdots,T_k\}$. 论文中首次主题及其个数的确定,是通过 LDA 模型在训练过程中针对训练的语料库所得的语料库的主题向量来确定. 在系统的运行过程中,每隔一段时间,如系统维护时,会将近期的历史微博数据作为新的训练语料库,进而更新主题集合,以更好地适应微博平台的语义环境. 则给定 t_j 时间段内的微博文本 M_{t_j},其主题分布定义如下:

定义 1 主题分布. 对于 t_j 时间段内微博文本 M_{t_j},其主题分布可表示为

$$P_{t_j}=(p(T_1|M_{t_j}),p(T_2|M_{t_j}),\cdots,p(T_K|M_{t_j})) \quad (2)$$

其中 $p(T_i|M_{t_j})$ 表示用户在 t_j 时间段内发布的微博 M_{t_j} 属于主题 T_i 的概率.

定义 2 用户的兴趣取向. 用户的兴趣取向可以通过对各时间段内微博文本主题向量的加权来表示. 考虑到用户的兴趣取向往往随着时间的推移而发生着变化,论文定义了一个时间函数 $f(t)$. 用户的兴趣向量可以定义为

$$P_u=f(t_1)P_{t_1}+f(t_2)P_{t_2}+\cdots+f(t_n)P_{t_n} \quad (3)$$

由于用户最近所发布、评论或转发的微博往往反映了用户当前的兴趣关注点,因而需要增加近期微博文本主题向量的权重,所以时间函数 $f(t)$ 应是单调递减的,且其取值范围应在[0,1]内,受文献[5]的启发,论文选取指数函数来描述随着时间的推移,用户历史微博文本主题向量权重的递减. 时间函数为

$$f(t)=e^{-\lambda\cdot t} \quad (4)$$

其中 λ 作为一个可控变量,用于调整函数递减趋势的陡峭情况.

3 计算微博发布者被关注可能性

3.1 Wtf 问题

Wtf(Who to follows)问题,即“关注谁”的问题. 由于微博平台不仅仅是一个巨大的实时信息传播平台,更是一个庞大的社交平台. 因而微博推荐不仅仅是为用户推荐其感兴趣的内容,一定程度上更希望用户能关注微博发布者,甚至与其成为好友. 受文献[6,7]的启发,论文从两个方面来解决 Wtf 问题,推断用户潜在关注对象,即“用户信任度”和“用户相似性”.

3.2 发布者被关注概率

用户所关注的用户群,一般代表了该用户的兴趣取向以及社交关系、专业方向. 当用户 U 关注的所有用户中,有一定比例的用户同时关注了用户 Q,往往代表着用户 Q 在用户 U 的专业或兴趣方面的影响力越高,用户 U 对用户 Q 的信任度越高. 因而用户信任度定义如下:

定义 3 用户信任度. 对于用户 U,其对用户 Q 的信任度可表示为

$$C_{U,Q}=N_{U,Q}/N_U \quad (5)$$

其中 $N_{U,Q}$ 代表用户 U 所关注的所有用户中,关注用户 Q 的数量,N_U 代表用户 U 所关注的用户数量.

另一方面,用户往往更倾向于关注跟自己兴趣相似的用户,希望找到与自己志趣相投的朋友. 由于用户发布、转发或收藏的微博,往往代表了用户的兴趣取向. 以用户的发布、转发或收藏的微博文本主题分布 P_u 代表用户的兴趣取向,从而通过计算用户的兴趣向量的相似性来衡量用户间的相似性.

众所周知,KL(Kullback-Leibler)距离[8]是衡量概率分布间距离的有效方法,其公式如下:

$$D_{KL}(U\|Q)=\sum_{j=1}^{T}U_j\ln\frac{U_j}{Q_j} \quad (6)$$

然而 KL 距离并不是对称的,即 $D_{KL}(U\|Q)\neq D_{KL}(Q\|U)$,因而论文采用 JS 距离[9]来衡量用户 U 与用户 Q 兴趣差异,公式如下:

$$D_{JS}(U\|Q)=\frac{1}{2}(D_{KL}(U\|M)+D_{KL}(Q\|M)) \quad (7)$$

其中 $M=\frac{1}{2}(U+Q)$,D_{KL} 为 KL 距离计算公式. 由于要综合考虑用户信任度和用户相似性两方面因素,而用户信任度 $C_{U,Q}$ 的取值范围为[0,1],而 JS 距离取值范围为[0,∞),为统一二者的衡量标准,因而将用户相似性定义如下:

定义 4 用户相似性. 对于用户 U 与用户 Q,二者的相似性可定义为

$$S_{U,Q}=\frac{1}{D_{JS}(U\|Q)+1} \quad (8)$$

其中 $D_{JS}(U\|Q)$ 为用户 U 与用户 Q 兴趣向量的 JS 距离. $S_{U,Q}$ 值越大,则用户 U 与用户 Q 越相似.

定义 5 用户被关注可能性. 对于用户 Q,其被用户 U 关注的可能性可用如下函数表示

$$P_{U,Q}=\sigma C_{U,Q}+(1-\sigma)S_{U,Q},\sigma\in(0,1) \quad (9)$$

其中 σ 为调和参数,调整 $C_{U,Q}$ 与 $S_{U,Q}$ 的比例.

4 微博的个性化推荐

4.1 微博文本预处理

对于新发布的微博数据,系统首先会对各微博数据进行分层聚类,从而实现对微博数据进行分组. 为了

提高效率,只对分词、去停用词之后的微博文本通过简单的余弦相似度进行聚类. 且去分词、去停用词也是接下来用 LDA 模型分析微博主题分布的必要步骤. 对于分组个数的选择,即聚类方法何时停止,论文采取 Dunn Index[10] 作为评价标准.

当微博分组完成后,论文使用 LDA 模型提取各组中微博的主题,并将各组中所有微博的主题向量相加并做归一化处理,则最后的聚类结果为一个主题向量集合 $\boldsymbol{C}=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_k\}$,其中 C_i 表示其中一个分组的主题向量,k 由 Dunn Index 确定.

4.2　用户分组的选取

当微博分组完成后,对于用户 U,设定阈值 M. 根据用户 U 的稳定兴趣向量与各分组的主题向量的相似度的高低来确定为用户 U 推荐的微博组,算法过程如算法 1 所示.

算法 1　选取为用户 U 推荐的微博组

输入:用户 U 的兴趣向量 $\boldsymbol{P}_u$、各微博组的主题向量集合 $\boldsymbol{C}$
输出:为用户推荐的微博组及其权重列表 H
1. 列表 H 为 $\boldsymbol{P}_u$ 与 $\boldsymbol{C}$ 中各元素相似度的值;sum = 0;
2. FOR EACH $\boldsymbol{C}_i$ IN $\boldsymbol{C}$
3. IF ($\boldsymbol{P}_u$ 与 $\boldsymbol{C}_i$ 的相似度 S > 阈值 M)
4. 将 S 加入列表 H;
5. ELSE
6. 将 0 加入列表 H;
7. FOR EACH H_i IN H
8. sum + = H_i;
9. //对 H 中元素进行归一化处理
10. FOR EACH H_i IN H
11. $H_i=\frac{H_i}{\text{sum}}$;
12. RETURN H

假设用户 U 与各微博分组的主题向量 $\boldsymbol{C}$ 的计算结果为 H,$H=\{H_1,H_2,\cdots,H_k\}$,其中 $\sum_{i=0}^{k}H_i=1$. 假设其中 H_1、H_2 的值不为 0,则为用户 U 推荐的微博组为 2 个,H_1、H_2 的值代表最终从其所在微博分组选取的微博占最终推荐的 Top-k 微博列表的比重.

4.3　为用户推荐个性化的 top-k 条微博

由于用户往往对热点事件有着更高的关注度,令微博 m 有 n 个特征,记为$(x_1,x_2,\cdots,x_n)$. 一般而言,被转发或者被评论次数越多的微博往往受关注程度越高,因而论文选取的微博特征为(ret,rep),其中 ret 代表微博 m 被转发的次数,rep 代表被评论的次数. 因而每条微博即可记为分布在二维空间内的点,其模型如图 3 所示.

因为微博的被转发次数或被回复次数服从长尾分布[11],所以左下角往往对应着关注度低的微博索引,且分布密集;右上角部分则对应着关注度较高的微博索引,且分布较稀疏.

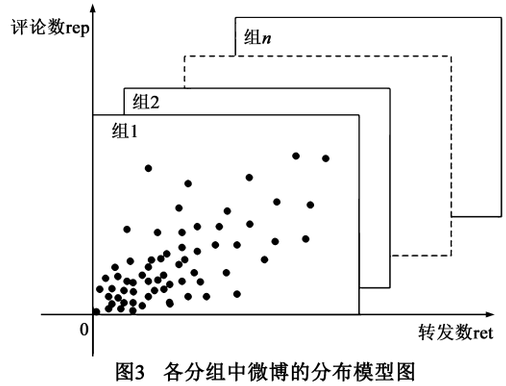

图3　各分组中微博的分布模型图

对为用户推荐的 Top-k 条微博的索引,可以用网格索引的方式来解决,具体访问过程如图 4 所示.

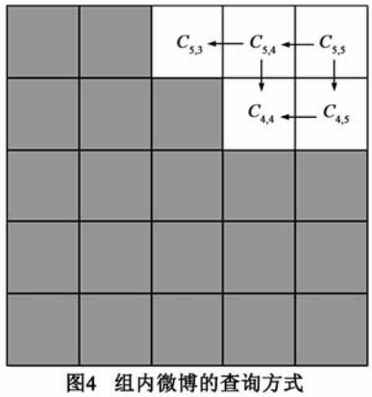

图4　组内微博的查询方式

因为 $C_{5,5}$ 中存储着关注度最高的微博,因而优先访问最右上角的 $C_{5,5}$,计算 $C_{5,5}$ 中各微博的主题向量与用户的近期兴趣关注点向量的相似度 S,将相似度最高的微博加入用户的个性化微博 Top-k 列表. 之后的访问次序是$\{C_{5,4}\rightarrow C_{4,5}\rightarrow C_{5,3}\rightarrow C_{4,4}\cdots\}$. 算法具体过程如算法 2 所示.

算法 2　为用户推荐个性化的微博 Top-k 列表

输入:为用户推荐的微博组及其权重列表 H,各微博组 G
输出:为用户推荐的个性化微博 Top-k 列表 $L(u)$
1. 将列表 $L(u)$ 初始化为空;
2. 列表 $L(u)$ 中各微博的发布者能被当前用户关注的可能性值的列表 S 初始化为 0;
3. FOR EACH H_k IN H
4. IF($H_k\neq 0$){
5. 将 H_k 对应微博组 G_k 最右上角格子 $C_{k_n,n}$ 加入队列 Q;
6. 微博组 G_k 与用户 U 的匹配最大值 $Gs_{\max}=0$;
7. REPEAT:队列 Q 不为空;
8. 弹出 Q 中的第一个元素 $C_{k_i,j}$;

```
9.    FOR EACH 微博 t IN C_{k_{i,j}}
10.     IF(被关注可能性(微博 t 发布者,U) > 阀值 M)
        //被关注可能性(微博 t 发布者,U)表示微博 t 发布者被用户 U 关注的可能性
11.     IF(被关注可能性(微博 t 发布者,U) > Gs_max)
12.     Gs_max = 被关注可能性(微博 t 发布者,U);
13.     IF(被关注可能性(微博 t 发布者,U) > S 中的最小值 &
        微博发布者未有微博位于 L 中)
14.     用微博 t 替换列表 L(u)中最小值 S_min 所对应的微博,同时更新 S,重新计算 S_min;
15.     END FOR
16.   IF(Gs_max < S_min && i == j){
17.     Count = 列表 L(u)内属于微博组 G_k 的微博个数;
18.     IF(Count < H_k · K){
19.     选取微博组 G_k 中"转发数" + "评论数"最高的(H_k · K - Count)条微博加入 L(u);
20.     BREAK;}
21.   }
22.   ELSE{
23.     IF(C_{k_{i-1,j}}不在队列 Q 中) 将其加入队列 Q;
24.     IF(C_{k_{i,j-1}}不在队列 Q 中) 将其加入队列 Q;
25.   }
26. }
27. END FOR
28. RETURN L(u);
```

算法首先遍历选定的微博分组(第 3 行),接着先将代表关注度最高的右上角网格 $C_{k_{s,s}}$ 加入队列 Q(第 5 行),弹出队列中元素并开始遍历元素中的所有微博(第 7 ~ 9 行),获取当前元素中最可能被用户关注的 $H_i \cdot K$ 个用户的微博并加入 $L(u)$ 列表(第 13 ~ 14 行),并记录当前元素的匹配最高值 $Gs_{\max}$(第 11 ~ 12 行). 微博发布者 Q 被用户 U 关注的可能性由上文所定义的"用户被关注可能性 $P_{U,Q}$来计算. 因为微博的被转发次数或被回复次数服从长尾分布,所以当 $i = j$ 时,$C_{k_{i,j}}$ 中的元素最多(第 16 行). 此时,若 $C_{k_{i,j}}$ 的 $Gs_{\max} < S_{\min}$,则选取微博组 G_k 中热度最高的微博填充 $L(u)$,即从微博组 G_k 中选取"转发数" + "评论数"最高的($H_k \cdot K$ - Count)条微博加入 $L(u)$,并跳出当前循环,开始遍历 H 内下一个元素(第 17 ~ 20 行);否则,将 $C_{k_{i-1,j}}$ 和 $C_{k_{i,j-1}}$ 加入队列 Q(第 22 ~ 25 行). 最终在遍历完 H 内元素后,返回推荐列表 $L(u)$(第 28 行).

5 实验分析

5.1 实验数据

本实验的数据由新浪微博开放平台提供的 API 自行抽取和采集,包含了 3082 个用户 2013 年 7 月 1 日至 2014 年 6 月 13 日的所有微博数据以及相互之间的关注信息. 论文获取了来自不同领域的认证用户数据,涉及了 8 个比较常见的领域,分别是科技、体育、房产、动漫、娱乐、健康、汽车和媒体. 由于微博数据来自于互联网,噪声大,需要做一定的预处理.

(1)将回复数和转发数低于 10 的微博去除.

(2)清除掉以@ 开头的微博信息,此部分微博为用户私人对话,不宜用于推荐.

(3)除掉少于 3 个名词的微博.

(4)根据用户实际有效的微博数量,从每个领域中各选取 100 个用户. 选取的过程会过滤掉有效微博数量小于 10 条或关注列表大小小于 10 的用户.

(5)去掉微博数据中特有的一些对主题挖掘无用的特征,如表情符号、分享目标、URL 地址等.

(6)对微博数据进行分词、去停用词操作,根据词性保留对主题表达有作用的名词和动词.

5.2 参数设置和对比实验

实验中随机选取 50 名用户,根据其发布的微博数据推断其兴趣取向,其他用户发布的所有微博混合在一起形成论文的测试数据集.

5.2.1 评价标准

论文采用准确率(Precision)、成功率(Success)以及平均倒数排名 MRR 作为系统的优化和评价标准. 对于每个用户,分别计算:

(1)准确率 P_k,即排序后的 Top-k 条微博中正确推荐的微博所占的比例;

(2)成功率 S_k,即所推荐的 Top-k 条微博中发布者被用户所关注的微博所占的比例;

(3)平均倒数排名 MRR,即最终的 Top-k 推荐列表中第一条正确微博所在位置的倒数均值,MRR 值越高,则算法准确率越高,公式如下所示:

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i} \tag{10}$$

为保证实验的客观性,论文采用多人判断法,选取了 10 个志愿者对推荐结果进行评价,并对 10 个志愿者所得出的准确率(Precision)、成功率(Success)以及平均倒数排名 MRR 取平均值.

5.2.2 时间函数的选择

在 3.2 节中,论文对于时间函数的选择问题上,选取了指数函数来描述随着时间的推移,微博文本主题向量 P_{t_j}权重递减这一特性. 为更好地理解时间函数如何影响最终的推荐结果,论文分别选取了逻辑函数、阻尼函数、指数函数进行了性能比对试验. 在此实验中,衰减系数 λ 的值设为 0.3,调和参数 σ 的值设为 0.4,在实际情况下,λ 和 σ 的值会根据推荐结果进行调整. 在下两节中会阐述 λ 和 σ 设定的实验过程. 评价标准分别为准确率、成功率、平均倒数排名 MRR. 实验结果如图 5 所示.

从图5中可以看出,指数函数不论在准确率、成功率、平均倒数排名MRR指标上都明显优于其他两种函数. 因而验证了论文在时间函数选择上的正确性.

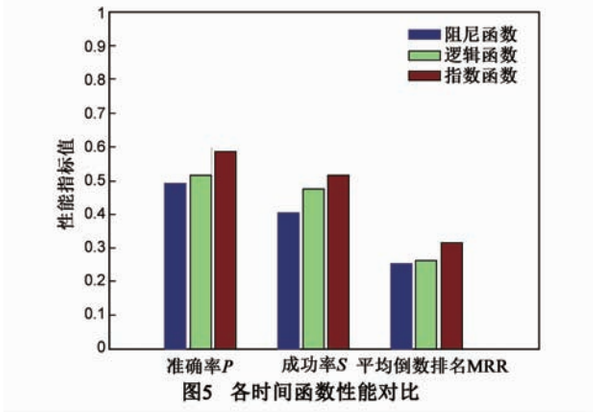

图5　各时间函数性能对比

5.2.3　衰减系数 λ 的调整

参数 λ 表示用户兴趣取向变化趋势的快慢,λ 的值越高,意味着用户发布时间越早的微博数据对于用户兴趣的提取影响越小;反之,λ 的值越小,也就意味着用户发布时间较早的微博数据对于用户兴趣的提取影响提高. 为计算参数 λ 对于推荐结果的影响,论文分别对不同 λ 取值下 Top_10 推荐结果进行了对比. 实验中控制定义5中用户被关注可能性函数 $P_{U,Q}$ 中的调和参数 σ 值不变. 评价标准分别为准确率、成功率、平均倒数排名 MRR. 实验结果如图6所示.

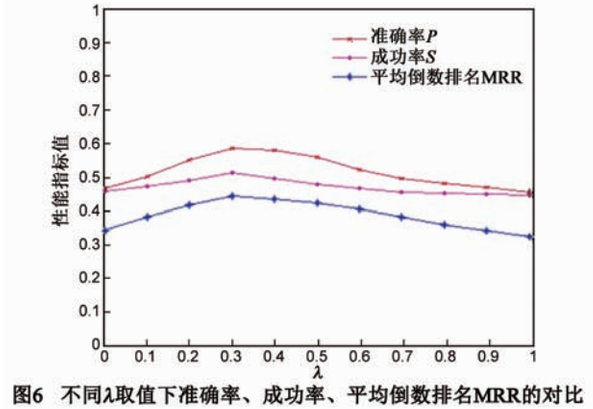

图6　不同λ取值下准确率、成功率、平均倒数排名MRR的对比

从图6中可以看出,当 λ 接近0.3时,推荐的准确率、成功率以及平均倒数排名MRR指标都最高,且在准确率、平均倒数排名 MMR 指标上尤为明显,因而实验中将 λ 设置为0.3. 且当 $\lambda=0$ 时,时间函数 $f(t)=e^{-\lambda\cdot t}=1$,即此时对用户的历史微博数据不做加权处理,直接得出用户兴趣向量;当 λ 逐渐增加时,即提高近期发布的微博数据所占权重时,推荐效果逐步提高;而当 $\lambda>0.3$ 时,推荐效果出现下降趋势. 因而对于用户兴趣的提取,应将长期和短期两个角度相结合,从而验证了论文引入时间函数对用户历史微博数据进行加权处理,进而得出用户兴趣向量的正确性.

5.2.4　调和参数 σ 的设定

在定义5中,调和参数 σ 的作用是调和用户信任度与用户相似性在计算用户被关注可能性时所占比重. σ 越大,则用户信任度所占比重越多;σ 越小,则用户相似性所占比重越多. 为计算参数 σ 对推荐结果的影响,论文分布对不同 σ 取值下 Top_10 推荐结果进行了对比. 实验中控制3.2节中时间函数的衰减系数 λ 的值不变. 评价标准分别为准确率、成功率、平均倒数排名MRR. 实验结果如图7所示.

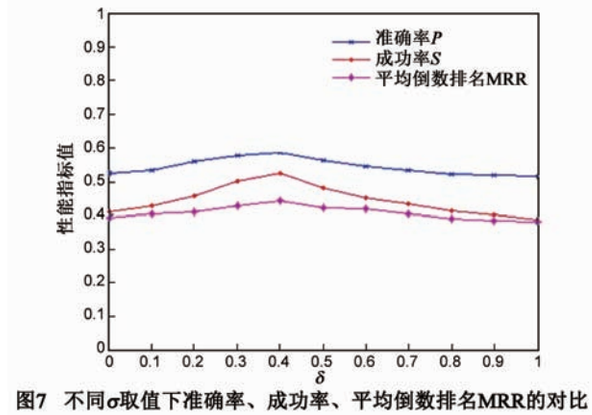

图7　不同σ取值下准确率、成功率、平均倒数排名MRR的对比

从图7可以看出,当 σ 取值接近0.4时,推荐的准确率、成功率、平均倒数排名 MRR 都最高,且在成功率指标上尤为明显,因而实验中将 σ 设置为0.4. 当 $\sigma=0$ 时,用户被关注可能性 $P_{U,Q}=\sigma C_{U,Q}+(1-\sigma)S_{U,Q}=S_{U,Q}$,即完全依靠用户相似性来确定用户被关注的可能性;而当 σ 值逐渐增加,即加大用户信任度的权重,推荐效果逐渐提高;而当 $\sigma>0.4$ 时,推荐效果出现下降趋势;且当 $\sigma=1$ 时,$P_{U,Q}=\sigma C_{U,Q}+(1-\sigma)S_{U,Q}=C_{U,Q}$,即完全依靠用户信任度来确定用户被关注的可能性. 因而解决 Wtf 问题,要将用户的兴趣取向与用户的社交关系相结合,从而验证了论文综合考虑用户相似性与用户信任度这两方面因素,进而计算用户被关注可能性的正确性.

5.2.5　与其他方法的比较

为验证论文所提出的方法的有效性和高效性,论文进行了如下3组实验:

(1)采用基于 LDA 模型的协同过滤 CF 算法[12]得到个性化推荐列表.

(2)采用基于用户动态兴趣度 LOI 模型[13]得到个性化推荐列表.

(3)采用论文提出的 DISN 方法得到个性化的推荐列表.

评价标准分别为准确率、成功率、平均倒数排名 MRR 和时间消耗. 实验结果如表1所示.

表 1 各方法在 Top_10、Top_20 推荐性能上的比较

方法	Top_10				Top_20			
	P	S	MMR	Time	P	S	MMR	Time
LDA	0.398	0.329	0.261	210ms	0.402	0.390	0.310	227ms
LOI	0.501	0.404	0.357	232ms	0.536	0.427	0.391	250ms
DISN	0.556	0.510	0.429	175ms	0.577	0.542	0.479	191ms

从实验结果可以看出:(1)采用基于 LDA 模型的协同过滤 CF 算法在准确率、成功率以及平均倒数排名方面都明显低于其他两种方法,这说明在提取用户兴趣时,如果直接将用户的历史微博数据无偏重地用于用户兴趣的提取,就不能很好地对用户的兴趣偏好建模;(2)论文提出的 DISN 方法相对于 LOI 方法在准确率、成功率以及平均倒数排名方面都更优异,且在成功率上表现明显,因而验证了对新发布的微博数据进行分组,以用户兴趣选取分组,进而根据微博发布者被用户关注的可能性去获取 Top_k 列表这一方法的正确性;(3)在时间效率方面,论文提出的 DISN 方法较其他两种方法都有较大改进,因而验证了本文采用网格索引的方式对新的微博数据进行查询这一方法的正确性.

6 总结与展望

论文对微博个性化推荐方法进行了研究,提出了一种基于用户动态兴趣和社交网络(DISN)的微博推荐方法.通过时间函数对用户历史微博数据进行加权处理进而提取用户的稳定兴趣向量.基于微博内容对新发布的所有微博数据进行分层聚类分组,再由用户的兴趣向量选定分组.进而采用网格索引的方式访问组内微博,再通过与计算微博发布者被目标用户关注的可能性,最终生成推荐列表.实验表明,论文提出的方法在准确率、成功率以及执行效率上均有较优异的表现.然而,本实验中系统的主题集合只有在维护时才能变更,且微博推荐往往还面临着海量性、实时性的问题,因而对系统的复杂性和扩展性也有更高的要求.如何更准确地确定主题集合,并进一步提高方法的效率和扩展性是我们今后研究工作的重点.

参考文献

[1] 高明,金澈清,钱卫宁,等.面向微博系统的实时个性化推荐[J].计算机学报,2014,37(4):963-975.
Gao Ming, Jin Chen-qing, Qian Wei-ning, et al. Real-time and personalized recommendation on microblogging systems[J]. Chinese Journal of Computers, 2014, 37(4): 963-975. (in Chinese)

[2] Otsuka E, Wallace S A, Chiu D. Design and evaluation of a Twitter hashtag recommendation system[A]. Proceedings of the 18th International Database Engineering & Applications Symposium[C]. New York: ACM, 2014. 330-333.

[3] 张引,张斌,高克宁,等.面向自主意识的标签个性化推荐方法研究[J].电子学报,2012,40(12):2353-2359.
Zhang Yin, Zhang Bin, GaoKe-ning, et al. Autonomy oriented personalized tag recommendation[J]. Acta Electronica Sinica, 2012, 40(12): 2353-2359. (in Chinese)

[4] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(5): 993-1022.

[5] Ding, Yi, Xue Li. Time weight collaborative filtering[A]. Proceedings of the 14th ACM International Conference on Information and Knowledge Management[C]. Bremen, Germany: ACM, 2005. 485-492.

[6] Khater S, Elmongui H G, Gracanin D. Personalized microblogs corpus recommendation based on dynamic users interests[A]. Proceedings of the 2013 International Conference on Social Computing[C]. Washington: IEEE, 2013. 979-982.

[7] Wan S, Lan Y, Guo J, et al. Informational friend recommendation in social media[A]. Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. New York: ACM, 2013. 1045-1048.

[8] Kullback S, Leibler R A. On information and sufficiency[J]. Annals of Mathematical Statistics, 1951, 22(1): 79-86.

[9] Endres D M, Schindelin J E. A new metric for probability distributions[J]. IEEE Transactions on Information Theory, 2003, 49(7): 1858-1860.

[10] Dunn J C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters[J]. Cybernetics and Systems, 1973, 3(3): 32-57.

[11] Yang Z, Guo J, Cai K, et al. Understanding retweeting behaviors in social networks[A]. Proceedings of the 19th ACM International Conference on Information and Knowledge Management[C]. Toronto: ACM, 2010. 1633-1636.

[12] Wang C, Blei D M. Collaborative topic modeling for recommending scientific articles[A]. Proceedings of the 17th

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. San Diego, California: ACM, 2011. 448 – 456.

[13] Khater S, Elmongui H G. Tweets you like: Personalized tweets recommendation based on dynamic users interests [A]. Proceedings of the ASE Big Data/Social Informatics/PASSAT/Bio Med Com 2014 Conference[C]. USA: Harvard University, 2014. 1 – 10.

作者简介

陈　杰　男,1990 年生,江苏如皋人. 现为南京工业大学计算机科学与技术学院硕士研究生. 研究方向为个性化推荐、数据挖掘.
E-mail:chenjie _ nj@126. com

刘学军　男,1970 年生,江苏南京人. 现为南京工业大学计算机科学与技术学院博士、教授、硕士生导师,CCF 高级会员. 研究方向为数据库、数据挖掘、传感器网络、隐私保护等.
E-mail:lxj _ njgd@163. com