

基于加权动态兴趣度的微博个性化推荐

陶永才^{1*}, 何宗真¹, 石磊¹, 卫琳², 曹仰杰²

(1. 郑州大学 信息工程学院, 郑州 450001; 2. 郑州大学 软件技术学院, 郑州 450002)

(*通信作者电子邮箱 icytao@zzu.edu.cn)

摘要:针对微博信息量大、用户兴趣随时间变化特征,提出一种基于加权动态兴趣度(WDDI)的微博个性化推荐模型。WDDI模型考虑微博转发特征,并引入时间因子,利用微博主题模型基于转发的狄利克雷分配(RT-LDA)对用户微博进行研究,建立用户对主题的个体动态兴趣模型。通过用户与其关注用户的相似度和交互频率获取用户的群体动态兴趣,将用户个体兴趣与群体兴趣加权结合得到加权动态主题兴趣模型。对用户接收的新微博按动态兴趣度降序排列,实现微博动态个性化推荐。实验表明,WDDI模型较之传统推荐模型,在微博服务中能够更准确地反映用户动态兴趣。

关键词:加权;主题模型;动态兴趣;个性化推荐

中图分类号: TP391 **文献标志码:** A

Personalized microblogging recommendation based on weighted dynamic degree of interest

TAO Yongcai^{1*}, HE Zongzhen¹, SHI Lei¹, WEI Lin², CAO Yangjie²

(1. School of Information Engineering, Zhengzhou University, Zhengzhou Henan 450001, China;

2. School of Software Technology, Zhengzhou University, Zhengzhou Henan 450002, China)

Abstract: On account of the features that the information in microblogging is enormous and the microbloggers' interests change over time, a personalized microblogging recommendation model based on Weighted Dynamic Degree of Interest (WDDI) was proposed. WDDI model considered the microblogging retweet features and the time factor of tweets, studied the tweets of microbloggers by exploiting the microblog topic model Retweet-Latent Dirichlet Allocation (RT-LDA) and built the individual dynamic interest model. Then WDDI got user's group dynamic interest by the similarity and the interacted frequency between users and their followee. Combining the user's individual interest and the group interest, the weighted dynamic degree of interest model was built. By ranking the new tweets that the user received in descending order by the degree of interest, the dynamic personalized microblogging recommendation was achieved. The experimental results show that WDDI is able to reflect the users' dynamic interest more precisely than the traditional models.

Key words: weighting; topic model; dynamic interest; personalized recommendation

0 引言

近年来,随着新兴社交媒体的流行,微博,如 Twitter、新浪微博等,已经成为人们获取和分享实时信息的重要方式。2006年7月,Obvious公司推出 Twitter 应用。截止到2013年底, Twitter 的注册用户已突破5亿,且每天新增消息达2亿多条^[1]。用户可以发布不超过140个字符的微博,也可以自动接收其关注对象发布的微博。目前,由于用户关注对象发布的微博都会自动更新到该用户的微博首页,且平均每天接收的微博高达上千条,导致用户花费大量时间和精力筛选感兴趣的信息。此外,用户的兴趣也随着时间的推移而发生改变^[2],如图1显示了一个普通用户在一个月内对3个主题(体育、娱乐、科技)的兴趣度随时间的变化。因此,有效的信息过滤和个性化的微博推荐尤为重要。

目前,关于微博推荐的研究已经有很多。常见的微博推

荐大多从网络结构、关注分类等出发,如好友推荐^[3]、热门话题推荐^[4]、新闻推荐^[5]等,但关于微博内容推荐的研究比较匮乏。在推荐算法中,比较成功的是协同过滤推荐算法^[6],但该算法侧重研究用户群体兴趣对目标用户的影响,对目标用户个体兴趣的挖掘不够充分。微博内容虽然简短却包含了大量的实时信息,作为微博信息主要的载体,反映了微博用户的兴趣变化。因此,对微博文本内容进行挖掘以发现用户潜在的个体兴趣尤为必要。传统的文本主题挖掘算法多采用文本聚类方法,将文本表示成单词向量,词汇相似的文本应该蕴含着相同的主题,但是基于单词向量的文本表示无法准确地描述文本的语义。微博信息简短,包含的单词数量有限,传统的文本挖掘方法并不能很好地适用于微博^[7]。适用于短文本的挖掘方法起初使用维基百科等知识库扩充词之间的语义关系,但知识库中词汇有限,之后引入“主题”概念,更好地表达文本的语义,即概率主题模型^[8]。近年来,概率主题模型

收稿日期:2014-07-16;修回日期:2014-08-27。

基金项目:河南省教育厅自然科学基金资助项目(2011B520035);河南省教育厅科学技术研究重点项目(13A520651)。

作者简介:陶永才(1975-),男,河南武陟人,讲师,博士,主要研究方向:高性能计算、社交网络;何宗真(1987-),女,河南驻马店人,硕士研究生,主要研究方向:社交网络;石磊(1967-),男,河南郑州人,教授,博士,CCF会员,主要研究方向:高性能计算、社交网络;卫琳(1968-),女,河南郑州人,副教授,硕士,主要研究方向:Web挖掘;曹仰杰(1976-),男,河南郑州人,博士,主要研究方向:高性能计算。

被广泛应用于主题挖掘、文本分类、文本检索等领域。潜在狄利克雷分配主题模型(Latent Dirichlet Allocation, LDA)由 Blei 等^[9]提出,它将文本表示成多个主题的概率分布,将主题表示成多个单词的概率分布,在文本和单词之间增加了主题语义层,能够更好地挖掘文本的语义。Hong 等^[10]研究了三种在微博系统中使用数据集训练主题模型的方法,使其更好地适用于微博中的短信息。文献[11]主要研究单条微博的主题提取及分类。Ramage 等^[12]利用 Labeled-LDA 对 Twitter 的内容和用户建模,并应用于微博排序、用户推荐等,取得了不错的效果。以上研究侧重于微博文本这一显式特征,而没有考虑微博用户的行为及用户间关系等隐式特征。微博与普通文本不同,微博不仅内容简短,且微博之间还有着特殊的转发关系。微博用户的行为特征及用户间关系影响着用户的兴趣,导致用户的兴趣会随着时间的推移而发生相应的改变。

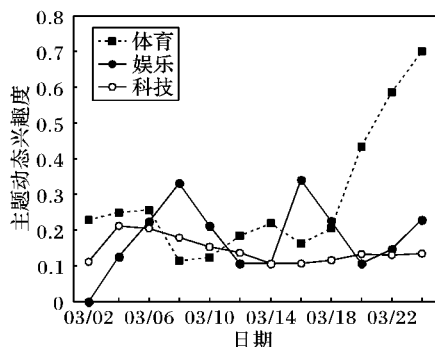


图1 2014年3月份用户u对主题的动态兴趣度

针对以上特点,本文利用微博主题模型挖掘微博文本的潜在主题语义,考虑用户间交互行为和用户关系,提出了一种基于加权动态兴趣度(Weighted Dynamic Degree of Interest, WDDI)的个性化微博推荐模型,研究随着时间的推移,用户的兴趣在不同主题上的变化,并在此基础上推荐用户感兴趣的新微博列表。

1 研究背景

1.1 潜在狄利克雷分配主题模型

潜在狄利克雷分配主题模型 LDA 是一个“文本-主题-词”的三层贝叶斯概率主题模型,如图2所示。该模型由参数 (α, β) 确定,给定一个有 M 篇文本的文本集合 D ,共包含 K 个主题 z , N 个单词 w 。那么文本的生成过程可描述如下:

1)对每个文本 $d \in D$,随机变量 θ 服从 Dirichlet 分布 $(\theta \sim \text{Dir}(\alpha))$,得到文本 d 上主题的多项分布参数 θ 。 θ 是一个 $M \times K$ 的矩阵,每一行的向量表示文本 d 上隐含主题的多项式条件概率分布。

2)对每个主题 $z_k \in z$,随机变量 φ 服从 Dirichlet 分布 $(\varphi \sim \text{Dir}(\beta))$,得到主题 z_k 上单词的多项分布参数 φ 。 φ 是一个 $K \times N$ 的矩阵,每一行向量表示主题 z_k 上单词的多项式条件概率分布。

3)对文本 d 中的第 n 个单词 w_n :

根据 z_k 服从多项分布 $(z_k \sim \text{Mult}(\theta))$,得到主题 z_k ;

根据 w_n 服从多项分布 $(w_n \sim \text{Mult}(\varphi))$,得到单词 w_n 。

其中,参数 θ 和 φ 分别表示文本中各个主题的相对重要性和主题中各个单词的相对重要性。

LDA 模型对 M 个文本建模如下:

1)对主题变量 z_k 进行 Gibbs 抽样,经过参数估计,间接得到 θ 和 φ 。每个文档用 LDA 生成的概率为:

$$P(d | \alpha, \beta) = \int p(\theta | \alpha) \left(\prod_{n=1}^N \sum_{k=1}^K p(z_k | \theta) p(w_n | z_k, \beta) \right) d\theta \quad (1)$$

2) M 个文档集合 D 用 LDA 生成的概率为:

$$P(D | \alpha, \beta) = \prod_{m=1}^M \int p(\theta_m | \alpha) \left(\prod_{n=1}^{N_m} \sum_{k=1}^K p(z_k | \theta_m) p(w_{mn} | z_{mk}, \beta) \right) d\theta_m \quad (2)$$

3)输出文档的主题概率分布和主题的单词概率分布。

利用 Gibbs 抽样参数推理,简化文本数据,可以有效地实现降维,缓解微博数据的稀疏性。

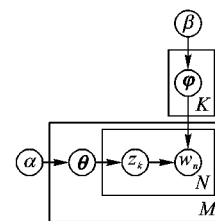


图2 LDA 主题模型

1.2 微博主题模型

用户发表的微博从时事新闻直到生活琐事,涉及到各种各样的主题。为了能够预测用户的兴趣集合,需建立用户对不同主题的兴趣度模型。本文选择潜在狄利克雷分配主题模型(LDA)对微博数据主题进行计算,得到用户感兴趣的主题以及每条微博所涵盖的主题比例。在微博主题模型中,每条微博被认为是一个文档,文档中所有互异单词构成词汇表,利用这些文档和单词可以提取微博的主题。

LDA 模型对单个文本建立文本-主题概率模型,并没有考虑文本之间的关系。而微博不同于一般的文本,它们之间存在着非常重要的关联关系即转发关系,记为 RT (retweet)。对于转发微博,转发部分和被转发部分的主题通常是相关的。如一条转发微博:“给力 RT@ 使徒子湖人摸黑战雷霆”,对转发部分的内容“给力”很难准确地提取主题,但通过转发关系,联系被转发部分的内容,可以推测转发部分是对一场篮球赛的评论。这种转发关系定义如下:

定义1 用户可以对某条自己感兴趣的微博 M1 进行转发并发表微博 M2,则定义 M2 为转发微博,M1 为被转发微博。RT 表示微博 M1 与 M2 之间的转发关系。

基于转发的狄利克雷分配 (Retweet-Latent Dirichlet Allocation, RT-LDA) 模型在 LDA 的基础上建模了微博文本之间的转发关系,形成适合于微博的主题模型。如图3所示,模型中各符号定义如表1所示。其中 r 表示转发关系,参数 λ 决定主题来自被转发微博还是转发微博本身的比例。图中微博的主题 z_k 取决于变量 r : 如果微博 M2 没有被转发微博,则 $r = 0$,表示微博为原创微博,该微博的主题由微博本身的主题先验概率 α 和微博本身的主题分布 θ 决定;如果 M2 有转发微博 M1,由参数 λ 来判断微博 M2 的主题 z_k 由参数 θ 或 θ_n 的多项式分布来决定。 $\lambda \in [0,1]$, r 服从参数为 λ 的二项分布,如果 $r = 1$,则微博 M2 与被转发微博 M1 相关,因此其主题完全由被转发微博 M1 的先验概率 α_n 和被转发微博本身

的主题分布 θ_n 决定,如果 $r = 0$, 则微博 M2 与 M1 不相关,主题由 M2 本身的主题分布 θ 决定。

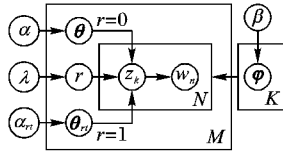


图3 微博主题模型 RT-LDA

表1 RT-LDA 模型中各符号含义

符号	定义	符号	定义
α	微博的主题先验概率	β	φ 的参数
α_n	被转发微博的主题先验概率	z_k	微博中的主题
θ	微博的主题概率分布	w_n	微博中的词
θ_n	被转发微博的主题概率分布	M	文本数
r	判断转发微博的布尔参数	N	词数
φ	主题上词的概率分布	K	主题数
λ	转发微博的权重参数		

微博 e 主题模型生成的概率如式(3)所示:

$$P(e | \lambda, \alpha, \beta) =$$

$$\int p(\theta | \alpha)^{1-r} p(\theta_n | \alpha_n)^r \left(\prod_{n=1}^N \sum_{k=1}^K p(z_k | \theta)^{1-r} p(z_k | \theta_n)^r p(w_n | z_k, \beta) \right) d\theta \quad (3)$$

微博主题模型 RT-LDA 中微博主题生成过程如算法 1 所示。

算法 1 RT-LDA 微博主题生成过程。

输入 超参数 α, α_n, β ; 微博文本 e 的集合。

输出 每个文本的主题概率分布 θ, θ_n ; 每个主题的单词概率分布 φ 。

- 1) for each topic $z_k, k \in \{1, 2, \dots, K\}$ do
- 2) draw $\varphi \sim \text{Dir}(\beta)$; /* 每个主题的单词概率 φ 分布服从参数为 β 的狄利克雷分配 */
- 3) endfor
- 4) for each microblog e do
- 5) for each word w_n do
- 6) if has “RT” do
- 7) draw $r \sim B(1, \lambda)$;
- 8) // r 服从参数为 λ 的二项分布
- 9) if $r = 1$ do
- 10) draw $\theta_n \sim \text{Dir}(\alpha_n)$;
- 11) // 转发微博与被转发微博相关
- 12) draw $z_k \sim \text{Multi}(\theta_n)$; /* 该微博文本的主题分布为被转发微博的主题分布 θ_n */
- 13) else
- 14) draw $\theta \sim \text{Dir}(\alpha)$;
- 15) // 转发微博与被转发微博不相关
- 16) draw $z_k \sim \text{Multi}(\theta)$;
- 17) // 该微博文本的主题分布为自身主题分布 θ
- 18) endif
- 19) else
- 20) draw $z_k \sim \text{Multi}(\theta)$;
- 21) endif
- 22) draw $w_n \sim \text{Multi}(\varphi)$;

19) endfor

20) endfor

算法首先对每个主题 z_k , 计算其单词的概率分布 φ 服从参数为 β 的狄利克雷分配, 记为 $\varphi \sim \text{Dir}(\beta)$ (第 1) 行至第 3) 行); 然后对每条微博 e , 首先判断 e 是否为转发微博, 如果是, 然后判断转发微博与被转发微博是否相关, 若相关, 则其主题概率分布等于被转发微博的主题概率分布 θ_n , θ_n 服从参数为 α_n 的狄利克雷分配, 记为 $\theta_n \sim \text{Dir}(\alpha_n)$, 每个主题 z_k 服从以 θ_n 为参数的多项分布, 记为 $z_k \sim \text{Multi}(\theta_n)$, 每个单词 w_n 服从以 φ 为参数的多项分布, 记为 $w_n \sim \text{Multi}(\varphi)$ (第 4) 行至第 10) 行); 若不相关, 则其主题概率分布为自身主题分布 θ , θ 服从以参数为 α 的狄利克雷分配, 记为 $\theta \sim \text{Dir}(\alpha)$ (第 11) 行至第 14) 行)。如果 e 不是转发微博, 则其主题概率分布 $\theta \sim \text{Dir}(\alpha)$, 每个主题 z_k 服从以 θ 为参数的多项分布, 记为 $z_k \sim \text{Multi}(\theta)$, 每个单词 w_n 服从以 φ 为参数的多项分布, 记为 $w_n \sim \text{Multi}(\varphi)$ (第 15) 行至第 20) 行)。

2 WDDI 个性化推荐模型

微博内容信息在一定程度上反映了用户的个体兴趣, 而用户的当前兴趣不仅取决于其个体兴趣, 同时还受其关注用户群体的影响^[13]。微博信息实时更新, 用户兴趣也会随着其关注用户更新的微博而变化。因此, 本文提出了基于加权动态兴趣度的个性化推荐模型 WDDI, 该模型不仅能够定性描述用户的兴趣, 还能定量地给出用户对某一微博感兴趣的程度。模型包含用户个体动态兴趣度、用户群体动态兴趣度、加权动态兴趣度和个性化推荐 4 个部分。图 4 所示为 WDDI 模型的流程。

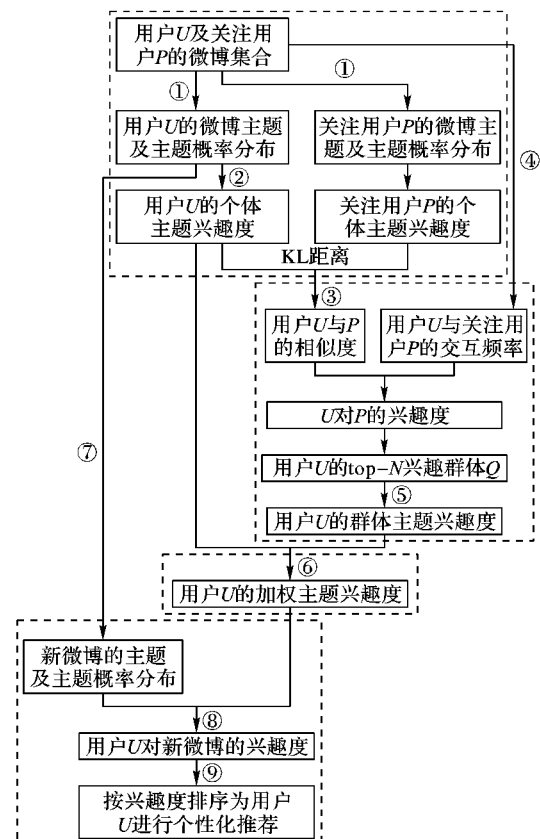


图4 WDDI 模型流程

第 1 步 模型采用微博主题模型 RT-LDA 计算目标用户

微博集合的潜在主题概率分布(①),引入微博时间因子得到目标用户对主题的动态个体兴趣度(②);

第2步 通过目标用户与其关注用户之间的相似度(③)和交互频率(④)得到用户感兴趣的群体用户,从而获取目标用户的动态群体兴趣(⑤);

第3步 将目标用户的个体动态兴趣度和群体动态兴趣度加权结合得到用户对微博主题的加权动态兴趣度(⑥);

第4步 根据新发布微博的主题概率分布(⑦)及目标用户对主题的加权动态兴趣度(⑥),计算可得目标用户对新微博的加权动态兴趣度(⑧)。将新微博按兴趣度降序排列,为用户个性化地推荐 TOP-N 个微博(⑨)。

2.1 个体动态兴趣模型

对每个用户 u , 获取用户 u 发布、评论、转发、收藏、赞的微博集合 E 。令 $E = \{e_1, e_2, \dots, e_m\}$, m 为微博个数。

定义2 动态主题分布。令主题集合 $T = \{t_1, t_2, \dots, t_k\}$, $e \in E$ 为一条用户微博, 则 $S_e[t_i]$ 表示该微博 e 属于主题 t_i 的概率, 由这 k 个主题概率组成的向量为 $S_e = \{S_e[t_1], S_e[t_2], \dots, S_e[t_k]\}$, 表示微博 e 的主题概率分布。对每条微博 e , 由 RT-LDA 模型计算 S_e 。

令微博 e 的时间因子 $w_{\text{time}}(u, e)^{[14]}$ 为:

$$w_{\text{time}}(u, e) = \exp\left\{-\ln 2 * \frac{\text{cur}_{\text{time}} - e_{\text{time}}}{L_u}\right\} = \exp\left\{-\ln 2 * \frac{\text{cur}_{\text{time}} - e_{\text{time}}}{\max_{\text{time}} - \min_{\text{time}}}\right\} \quad (4)$$

其中: \max_{time} 表示用户微博集合中最早的微博时间, \min_{time} 表示用户微博集合中最晚的微博时间, cur_{time} 为当前时间, e_{time} 为微博时间, L_u 表示用户 u 的微博集合的时间跨度。即离当前时间越近的微博, 其时间因子越大, 越能反映用户的兴趣。

则微博 e 的动态主题分布为:

$$S_e' = w_{\text{time}}(u, e) * S_e \quad (5)$$

定义3 动态兴趣分布。令 $E = \{e_1, e_2, \dots, e_m\}$ 表示用户 u 的微博集合, 则该用户对主题 t_i 的个体动态兴趣分布 U_i^u 为:

$$U_i^u = \{U_{t_1}^u, U_{t_2}^u, \dots, U_{t_k}^u\}; U_{t_i}^u = \frac{1}{m} \sum_{j=1}^m S_{e_j}[t_i]' \quad (6)$$

其中 $U_{t_i}^u$ 表示用户 u 对主题 t_i 的个体动态兴趣度。

2.2 群体动态兴趣模型

用户的群体动态兴趣受其关注用户的影响。首先计算用户 u 对其关注用户 p 的兴趣度, 确定用户 u 的兴趣群体。用户 u 对其关注用户 p 的兴趣度由用户之间的相似度和交互频率的共同决定。

1) 用户与关注用户的相似度: 以用户 u 与关注用户 p 对微博主题的个体兴趣概率分布的相似性来度量。它可以用矢量之间的余弦相似性计算, 但如果使用余弦夹角计算相似度就失去了概率主题模型的优势^[15], 本文使用能够衡量两个概率分布向量相似度的函数, 即 KL (Kullback-Leibler) 距离函数^[16], 即:

$$KL(U_i^u, P_i^p) = \frac{1}{2} \left(\sum_{i=1}^k U_{t_i}^u \ln \frac{U_{t_i}^u}{P_{t_i}^p} + \sum_{i=1}^k P_{t_i}^p \ln \frac{P_{t_i}^p}{U_{t_i}^u} \right) \quad (7)$$

$$\text{sim}(U_i^u, P_i^p) = 1/KL(U_i^u, P_i^p) \quad (8)$$

其中: 微博主题集合 $T = \{t_1, t_2, \dots, t_k\}$; U_i^u, P_i^p 分别表示用户 u 和 p 对主题的个体动态兴趣分布向量。

2) 用户与关注用户的交互频率: 用户可以对感兴趣的关注用户的微博进行转发、评论、收藏、赞的操作, 也可以用微博 @ 其感兴趣的关注用户。本文选取了转发、评论、收藏、赞、提及 5 类交互行为来反映用户对关注用户的兴趣。用户 u 对其关注用户 p 的交互行为可以形式化表示为:

$$\text{InteractionFrequency}(u) = \{\text{Retweet}(u), \text{Comment}(u), \text{Favorite}(u), \text{Attitude}(u), \text{Mention}(u)\} \quad (9)$$

其中: $\text{Retweet}(u)$ 为用户的转发行为; $\text{Comment}(u)$ 为用户的评论行为; $\text{Favorite}(u)$ 为用户的收藏行为; $\text{Attitude}(u)$ 为用户的赞行为; $\text{Mention}(u)$ 为用户的提及行为。

用户 u 的转发行为用向量表示为:

$$\text{Retweet}(u) = (r_1, \dots, r_i, \dots, r_n) \quad (10)$$

其中: n 为关注用户总数, r_i 为用户 u 对关注用户 i 的微博的转发次数。

用户 u 的评论行为用向量表示为:

$$\text{Comment}(u) = (c_1, \dots, c_i, \dots, c_n) \quad (11)$$

其中: n 为关注用户总数, c_i 为用户 u 对关注用户 i 的微博的评论次数。

用户 u 的收藏行为用向量表示为:

$$\text{Favorite}(u) = (f_1, \dots, f_i, \dots, f_n) \quad (12)$$

其中: n 为关注用户总数, f_i 为用户 u 对关注用户 i 的微博的收藏次数。

用户 u 的赞行为用向量表示为:

$$\text{Attitude}(u) = (a_1, \dots, a_i, \dots, a_n) \quad (13)$$

其中: n 为关注用户总数, a_i 为用户 u 对关注用户 i 的微博的赞次数。

用户 u 的提及 @ 行为用向量表示为:

$$\text{Mention}(u) = (m_1, \dots, m_i, \dots, m_n) \quad (14)$$

其中: n 为关注用户总数, m_i 为用户 u 对关注用户 i 的 @ 次数。

那么用户与关注用户的交互频率 (Interaction Frequency) 可以表示为:

$$IF(u) = (IF_1, \dots, IF_i, \dots, IF_n) \quad (15)$$

其中: n 为关注用户总数, IF_i 为用户 u 对关注用户 i 的交互次数。

结合用户 u 与关注用户 p 的相似度及交互频率, 计算用户 u 对关注用户 p 的兴趣度:

$$R(u, p) = \text{Intersect}(u) = \alpha * \text{sim}(u, p) + (1 - \alpha) * IF(u) \quad (16)$$

其中: p 为用户 u 的关注用户, 令 $\alpha = 1/2$ 。取 $R(u, p)$ 值较大的关注用户构成用户 u 的兴趣群体集合 $Q (Q \subset P)$, P 为用户 u 的关注用户集合。

3) 计算用户 u 的群体兴趣。根据用户 u 对其兴趣群体 Q 的兴趣度 $R(u, p)$, 计算用户 u 对微博主题的群体兴趣分布 U_i^c :

$$U_i^c = \sum_Q R(u, p) * P_i^p \quad (17)$$

其中: P_i^p 表示用户 P 对主题的个体兴趣分布, Q 表示用户 u 的兴趣群体。

2.3 加权动态兴趣模型

将用户的个体兴趣与群体兴趣进行线性加权, 计算用户的加权动态兴趣度, 其计算如式 (18) 所示:

$$U_i = a * U_i^c + (1 - a) * U_i^l \quad (18)$$

其中: U_i^l 为用户 U 的个体兴趣; U_i^c 为用户 U 的群体兴趣; $a \in [0, 1]$, 可知 a 取 0 时模型得到的是用户的个体兴趣分布, a 取 1 时得到的是用户的群体兴趣分布。 a 的取值根据具体系统而定, 在本文中经过实验测试来确定。

2.4 个性化微博推荐

一段时间内用户 u 的关注用户发布的新微博集合记为 E_{new} , 对每条新微博 e_{new} , 计算用户 u 对该微博的加权动态兴趣度, 如式 (19) 所示。

$$\text{DegreeofInterest}(u, e_{\text{new}}) = \sum_{t \in T} (S_{e_{\text{new}}}^t \cdot U_t) \quad (19)$$

其中: T 为用户 u 感兴趣的主体集合, $S_{e_{\text{new}}}^t$ 由式 (5) 得到 e_{new} 的动态主题分布向量, 结合式 (18) 得到的用户对微博主题的加权动态兴趣分布, 得到用户 u 对新微博 e_{new} 的兴趣度。

当用户的关注对象发布或转发了新微博时, 对新微博集合按兴趣度降序排序, 将 TOP- N 个新微博推荐给用户。

3 实验及分析

3.1 实验设置

为了采集所需数据进行实验分析, 以便验证本文推荐模型的准确性和有效性, 本文通过新浪微博平台^[17]提供的 API 获取实际用户数据。

根据实验要求, 采集用户发布、转发、评论、收藏及赞的微博信息, 每条微博包括微博内容及微博时间。此外, 采集用户的关注用户 ID 及其微博信息。2014 年 3 月从新浪微博中随机选取 7 286 位用户及他们的前 200 条微博用于实验。按照模型要求, 去掉关注对象少于 10 个、微博数少于 200 个的用户, 经过筛选, 得到可用于模型的用户为 1 419 位。用户的关注用户总数为 21 585 位。最终, 得到 21 596 位不同用户及他们的微博。经过分词, 去掉停用词, 得到微博文本共 4 319 200 条。

在本文 WDDI 模型中, 选取参数 $\alpha = 50/k$, $\beta = 0.01$, 主题个数 $k = 10$, 其中 $\lambda = 1$ ^[18], 表示转发微博与被转发微博全相关。

3.2 评价标准

本文使用推荐准确率 (Precision)、召回率 (Recall)、F 值 (F-Measure) 和排序准确率 (Rank-Precision, Rank-P) 作为评价推荐模型的依据^[2]。

微博系统中只有用户发布、转发等行为信息, 并没有包含用户对微博真实喜好的数据。对于此类系统, 通常假设用户的行为代表他的喜好。在微博系统中, 可观察到的用户对微博的操作行为有: 发布、转发、评论、赞、收藏。将用户对每个微博的喜好程度划分为四个等级^[19]: 0、0.33、0.67、1。其中划分规则如下:

- 1) 若用户没有对微博进行任何操作行为, 则用户对该微博的喜好程度为 0。
- 2) 若用户对微博进行评论、赞或收藏行为之一, 则用户对该微博的喜好程度为 0.33。由于发布和转发行为更能表达用户的喜好^[20], 若用户发布或转发微博, 则对该微博的喜好程度为 0.67。
- 3) 若用户对微博进行转发并评论、评论并赞等两种及以上行为, 则用户对该微博的喜好程度为 1。

根据实验用户微博数据集, 设定用户微博喜好平均值为喜好阈值。对测试集数据, 若用户对微博的喜好程度 \geq 喜好阈值, 则用户喜欢该微博; 否则, 为不喜欢。

根据新浪微博首页实际微博个数, 设定推荐列表长度 $N = 10$, 将所有待推荐的微博按 WDDI 模型计算的加权动态兴趣度降序排序, 将 TOP- N 条微博推荐给用户。待推荐微博的可能结果如表 3 所示, 其中 a, b, c, d 表示微博数目。

表 1 待推荐微博的可能结果

用户喜好	推荐	不推荐
喜欢	a	b
不喜欢	c	d

对于用户 u , 其推荐准确率 P 、召回率 R 和 F 值 F 为:

$$P = \frac{\text{推荐列表中用户喜欢的微博个数}}{\text{模型推荐的微博个数}} = \frac{a}{a + c} \quad (20)$$

$$R = \frac{\text{推荐列表中用户 } u \text{ 喜欢的微博个数}}{\text{用户 } u \text{ 喜欢的所有微博个数}} = \frac{a}{a + b} \quad (21)$$

$$F = \frac{2PR}{P + R} \quad (22)$$

推荐列表中的顺序关系对用户的推荐效果是至关重要的, 本文采用排序准确率 (Rank-P) 来处理微博在推荐列表中的位置, 排序准确率为:

$$\text{Rank-P} = \sum_u \frac{1}{N} * \sum_{i=1}^N \frac{1}{\text{pos}_i} \quad (23)$$

其中: N 为用户喜欢的微博个数, pos_i 为用户喜欢的微博在推荐列表中的位置。Rank-P 越大, pos_i 越小, 表示用户喜欢的微博在推荐列表中的排名越靠前, 推荐效果越好。

3.3 实验结果分析

本实验把每个用户的微博数据集平均分成 10 份, 其中 9 份作为训练集, 1 份作为测试集, 每次使用不同的测试集重复 10 次。对于每组数据集, 得出每个用户的 Top-10 推荐列表。这里的推荐列表可以通过以下 3 组实验得到:

- 1) 采用基于 LDA 模型的协同过滤 CF 算法^[21]得到个性化推荐列表。
- 2) 采用基于 RT-LDA 模型的协同过滤 CF 算法得到个性化推荐列表。
- 3) 采用基于加权动态兴趣度 WDDI 模型得到个性化推荐列表。

将这 3 个推荐列表与测试集进行对比, 实验结果如图 5 所示。群体兴趣在不同权重值 a 下对 WDDI 模型的影响如图 6 所示。

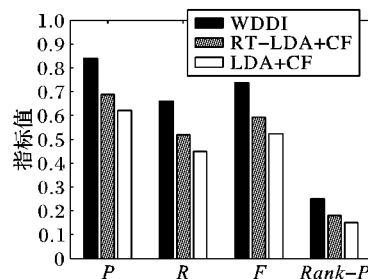


图 5 准确率、召回率、F 值和排序准确率对比

通过图 5 可知: 实验 1 和实验 2 对比说明采用基于传统 LDA 的 CF 推荐模型在推荐准确率和排序准确率方面稍低于改进后的 RT-LDA 模型, 说明微博的转发特征能够更好地反

