

微博个性化标签图形化 RTM 模型 Gibbs 采样推荐

刘真臻, 徐东平

(武汉理工大学 计算机科学与技术学院, 湖北 武汉 430063)

摘 要: 为提高个性化标签推荐方法性能, 提出基于 Gibbs 采样推理的微博个性化标签隐含关系主题模型 (Relation Topic Model, RTM) 推荐算法. 首先, 利用图形化形式对微博中的潜在局部信息进行表达, 对用户主题分布为代表的用户进行 top-k 相似用户发现, 然后计算出现在这些用户中的所有标签的频率, 并推荐与用户最相关的标签. 其次, 为挖掘潜在主题信息, 利用带惩罚项的增强型余弦相似度 RTM 模型对微博标签进行命名, 大大提高联合建模对潜在主题生成标签的影响, 并可发现全局标签和主题之间的关系; 最后, 通过真实的实验结果显示, 所提推荐方法要优于选取的 TF-IDF、RTMSA 等几种经典标签推荐算法, 验证了算法有效性.

关键词: Gibbs 采样; 微博标签; 关系主题模型; top-k 算法

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2017)12-0138-07

Gibbs Sampling Inference Based RTM Model for Micro-blog Personalized Label Recommendation

LIU Zhen-zhen, XU Dong-ping

(College of Computer Science and Technology, Wuhan University of Technology, Wuhan 430063, China)

Abstract: In order to improve the performance of personalized tag recommendation method, the Gibbs sampling inference based RTM model for Micro-blog personalized label recommendation was proposed. Firstly, we used the user topic distribution as the representative of the user with Top-k similar for the potential local information in micro-blog. Then we calculated the frequency of all tags appearing in these users, and recommended the most relevant tags with the user. Secondly, in order to explore the potential theme information, we named the micro-blog tags by using the implicit relation model, which could greatly improves the influence of the joint modeling on the latent topic generation tags, and finds out the relationship between the global tags and the topic; Finally, the experimental results show that the proposed method is superior to the selected TF-IDF, RTMSA and other classical label recommendation algorithm, and verify the effectiveness of the proposed algorithm.

Key words: gibbs sampling; micro-blog tags; relational topic model; top-k algorithm

1 引言

微博是一种日益普及的基于用户信息共享、传播和获取的平台, 用户可以通过网络、WAP 等客户端建立个人社区. 随着云计算技术的发展, 各个国家建立起了几个百微博网站, 如推特, Tumblr, Friend-Feed, cif2.net、搜狐微博等^[1-2]. 近年来, 微博大受欢迎, 以 Twitter 为例, 它拥有 5 亿多个用户, 其中 2016 年 12 月活跃用户数超过了 2.84 亿. 随着微博

用户的激增, 微博平台上的信息量也是水涨船高^[3]. 基于云计算, 微博创造了海量信息. 然而, 由于信息碎片化、云冗余等固有的局限性, 使得微博平台中隐藏的信息难以利用. 为了快速找到特定主题或内容相关信息, 有关学者提出了一些有影响力的推荐方法, 包括有影响力的用户推荐、粉丝推荐、微博推荐、标签推荐等. 这其中, 标签推荐是当前研究的热点和难点^[4].

当前, 已有许多较为成熟的标签推荐算法, 主要

可分为两种类型:(1)基于关键词的标签推荐方法,如例如文献[5]提出一种标签的词频(TF)推荐方法,其采取的策略是直接文档描述中进行最高词频提取,将所获得的关键词进行标签推荐;文献[6]在 TF 方法基础上提出改进的基于逆文档频率的词频标签推荐算法(TF-IDF),其利用文档描述中的逆文档词频进行最接近的 APIs 选取,并将其用于标签推荐;(2)基于主题方法,如文献[7]提出基于 LDASA 模型算法,其基于隐含狄利克雷分配模型(Latent Dirichlet Allocation,LDA)进行算法设计,在进行最接近 APIs 选取后,利用排序算法进行标签推荐;文献[8]提出 RTMSA 算法,其利用 RTM 模型进行微博文档 APIs 的隐含主题训练,RTM 模型与 LDA 模型相比,其考虑到了文档链接关系,该方法其余步骤与 RTMSA 算法类似;文献[9]提出 RTMCS 算法,同样适用 RTM 模型对 APIs 隐含主题进行训练,并对相似 APIs 进行合并排序推荐,可实现若干相关标签的推荐。

为提高个性化标签推荐方法性能,提出基于 Gibbs 采样推理的微博个性化标签隐含关系主题模型推荐算法.利用用户主题分布为代表的用户进行相似用户发现,然后计算出现在这些用户中的所有标签的频率,并推荐与用户最相关的标签.并利用 RTM 模型对微博标签进行命名,实现了对微博个性化标签的有效推荐.

2 个性化标签推荐算法框架

为有效识别微博标签并推荐给特定用户微博,这里提出一种基于协同过滤推荐方法,充分利用潜在用户轮廓的协同过滤和明确用户轮廓的协同过滤.同时其还考虑了主题生成对于标签推荐的重要性.本文采用的个性化标签推荐系统体系结构如图 1 所示.

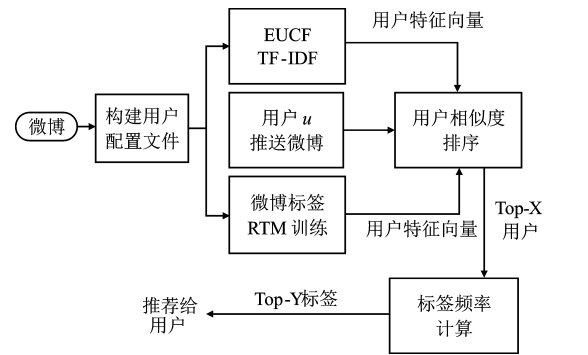


图 1 个性化标签推荐系统体系结构

通过用户 u 发布的微博构建用户 u 的配置文件.在进行个性化标签推荐前,需要根据训练数据对有关参数进行估计.两种方法来构建用户的使用特征向量,基于明确用户协同过滤(EUCF)方法(具体见文献[10]所示)和微博标签 RTM 方法(本文所提算法).

基于 EUCF 的方法的特征向量维度是整个训练数据的不同标签数量,因此,EUCF 的方法会产生高维向量,计算精度较高但计算复杂性也很高.与 EUCF 方法不同,微博标签 RTM 方法的特征向量维度与算法参数 T 相关,其所生成的是一种低维向量.当用户发表微博时,基于 EUCF 方法和基于微博标签 RTM 方法分别基于余弦相似度对用户间的相似度进行计算.然后,通过上述两种算法的线性组合进行用户相似性计算,并选取用户 u 的最接近的 X 个邻居点.

对于给定目标用户 u 和其他用户 u_i ,令 $Sim(u_i | EUCF, u)$ 表示利用 EUCF 方法计算得到的用户 u 和其他用户 u_i 之间的距离. $Sim(u_i | HashtagRTM, u)$ 表示利用微博标签 RTM 方法计算得到的用户 u 和其他用户 u_i 之间的距离.则用户 u 和其他用户 u_i 之间的线性组合方法如下:

$$Sim(u_i | EUCF + HashtagRTM, u) = \epsilon \cdot Sim(u_i | EUCF, u) + (1 - \epsilon)Sim(u_i | HashtagRTM, u) \tag{1}$$

式中, ϵ 取值区间是 $[0, 1]$,选择余弦相似度来评估用户之间的距离,微博标签推荐算法如下所示.

算法 1:微博个性化标签推荐算法

输入:用户 u 发送的微博 l ,相似用户数量 X ,用户推荐标签特定数 Y ;

输出:top- Y 推荐标签;

1. 构造用户轮廓特征向量 U ;
2. for $u_i \in U \& U \neq \emptyset$ do
3. 计算相似度 $Similarity(u_i | EUCF, u)$;
4. 计算相似度 $Similarity(u_i | HashtagRTM, u)$;
5. 计算相似度 $Similarity(u_i | EUCF + HashtagRTM, u)$;
6. endfor
7. 根据 $Similarity(u_i | EUCF + HashtagRTM, u)$ 选 top- X 用户;
8. 计算 top- X 用户中标签的出现频率;
9. top- X 频率标签集 $R = set$;
10. return $l = R$

获得上述 X 个邻居点后,计算出现在 top- X 用户配置文件中的微博标签出现频率,然后根据出现频率进行微博标签排序,最后对获得的 top- Y 最高

频率的微博标签进行推荐。

3 微博标签 RTM 模型

3.1 RTM 模型

Chang 等人最早提出了一种概率主题的层次模型,将其命名为关系主题模型,英文名称是 Relation Topic Model,可简记为 RTM 模型.该模型不但考虑到了所要推荐标签的内容,而且还在模型中纳入了标签关系链接,比如微博之间的相互转发和引用等关系链接,所提 RTM 模型见图 2 所示。

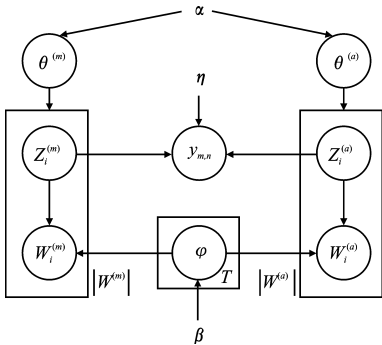


图 2 关系主题模型

图 2 中,标签主题向量分布为 $\theta^{(m)}$, API 标签主题向量分布为 $\theta^{(a)}$. API 标签同标签间的关系链接变量表示为 $y_{m,a}$,其可根据平滑参数 λ 和标签主题向量分布进行确定.训练过程中,RTM 标签推荐模型不但对 APIs 标签和标签 s 之间的信息关联进行考虑,同时还考虑了标签抽样过程中,与当前标签存在链接关系的主题标签向量分布情况。

为实现微博的 RTM 模型构建,在上述算法模型中,需要着重解决两部分内容:(1) 微博标签 RTM 模型形式,这里采用图形化形式进行表示,见 3.2 节;(2)RTM 模型的采样形式,这里采用 Gibbs 采样进行表达,见 3.3 节。

3.2 微博标签 RTM 模型描述

为了寻找可能影响标签和潜在主题之间关系的因素,提出了标签微博标签 RTM 模型,可通过潜在主题联合建模,增强主题的用户标签生成影响.假定 $Hashtag_LDA = \langle W, H, T, U \rangle$,其中 $W = (w_1, w_2, \dots, w_n)$ 表示微博词语, $H = (h_1, h_2, \dots, h_n)$ 表示微博中的微博标签, $T = (t_1, t_2, \dots, t_n)$ 表示潜在的主题, $U = (u_1, u_2, \dots, u_n)$ 表示用户配置文件, $N_u = (N_1, N_2, \dots, N_n)$ 是用户 $u \in U$ 发送的微博.微博标签 RTM 模型所使用的符号及其描述见表 1 所示.微博标签 RTM 模型的图形表示见

图 3 所示。

表 1 微博标签 RTM 模型符号

符号	描述
V	所有微博的独特词语, $ V $ 表示其数量
T	主题集合, $ T $ 表示其数量
U	用户集, $ U $ 表示其数量
N_u	用户 u 推送的微博
$L_{u,n}$	用户 u 推送的第 n 个微博词语数量
$K_{u,n}$	用户 u 推送的第 n 个微博标签数量
w	微博中的词语
z	特定主题
g	全局标签指示符
h	标签
$\rho, \lambda, \gamma, \alpha, \beta$	Dirichlet 或者 Bernoulli 先验参数
θ_u	用户 u 的主题分布
φ_t	主题 t 的词语分布
π_t	主题 t 的标签分布
σ	标签指示分布

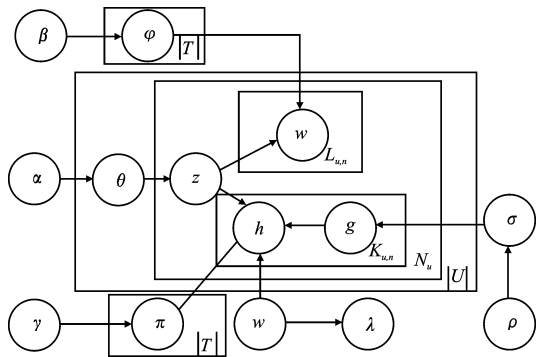


图 3 微博标签 RTM 模型的图形表示

图 3 中,每个节点代表一个随机变量,箭头代表节点之间的依赖关系.每个板块,标记有 $|N|$ 表示其有 $|N|$ 个这类节点.从 $Dir(\beta)$ 中采样的主题 $t \in T$ 具有多项式的词分布 φ_t 以及从 $Dir(\gamma)$ 中采样获得的多项式微博标签分布 π_t .对于用户 $u \in U$,从 $Dir(\alpha)$ 中采样获得多项式主题分布 θ_u .假定用户 u 的第 n 个微博主题为 t ,这个微博中的词只是由特定主题的词分布 φ_t 产生,但在微博的标签生成更复杂。

更具体地说,微博每个标签 h 由两部分的采样组成,即来自 $Dir(\lambda)$ 的全局多项标签分布 w 和特定主题的标签分布 π_t .全局分布 w 生成标签使用在微博长期存在的热点话题中.当进行微博标签选择时,利用从伯努利分布 σ 采样获得的指数 g 确定如何进行选择.当 $g_{u,n,h} = 0$ 时,微博标签在全局分布 w 中

采样获得,否则在 π_i 中采样获得. 标签生成过程是如算法 2 所示.

算法 2:RTM 模型计算流程

输入: APIs 文档、标签、API 标签及其关系链接;

输出: $\theta^{(m)}$, $\theta^{(a)}$, φ , $z^{(m)}$ 和 $z^{(a)}$ 等待模型估计参数;

1. 基于 Dirichlet 向量分布 $Dir(\beta)$, 对 API 文档或标签中存在的隐含主题进行近似, 获得其前面词向量分布的多项式表达形式 $\varphi \sim Dir(\beta)$.

2. 基于 Dirichlet 向量分布 $Dir(\alpha)$, 对标签描述进行近似, 获得其主题概率向量分布形式 $\theta^{(m)} | \alpha \sim Dir(\alpha)$.

3. 对所有存在的 API 文档, 循环执行同第 2 步程序.

4. 针对所有属于标签文档的标签词汇 $\omega_i^{(m)}$, 执行如下两步操作:

(1) 利用标签文档所呈现的主题向量概率分布 $\theta^{(m)}$, 对所推荐的标签词汇进行主题采样:

$$z_i^{(m)} | \theta^{(m)} \sim Multinomial(\theta^{(m)});$$

(2) 在上步获得的主题采样 $z_i^{(m)}$ 中, 进行标签词语的二次采样:

$$\omega_i^{(m)} | z_i^{(m)}, \varphi_{1:T} \sim Multinomial(\varphi_{z_i^{(m)}})$$

5. 针对所有属于 API 文档的标签词汇 $\omega_i^{(a)}$, 参照上步操作执行;

6. 对所有的 API 文档与标签之间存在的关系链接进行刻画:

$$y_{m,a} | z^{(m)}, z^{(a)} \sim \varphi(\cdot | z^{(m)}, z^{(a)}, \lambda)$$

$$z^{(m)} = \{z_1^{(m)}, z_2^{(m)}, \dots, z_I^{(m)}\}, z^{(a)} = \{z_1^{(a)}, z_2^{(a)}, \dots, z_I^{(a)}\}$$

算法 2 中, 如果 API 文档与标签之间有对应的链接关系属性, 则 $y_{m,a} = 1$. 函数 φ 的主要作用是对 APIs 文档与标签 s 存在的相似性进行测度和评价, 其值与 APIs 文档与标签 s 之间的主题标签的向量概率分布形式有关. 为了更准确的刻画 APIs 文档与标签 s 之间的的关系链接, 这里设计相似度余弦计算的增强公式, 具体为

$$\begin{aligned} \varphi(y_{m,a} = 1) = & \frac{\sum_{k=1}^T (z_k^{(m)} z_k^{(a)} / e^{\lambda | z_k^{(m)} - z_k^{(a)} |})}{\sqrt{(z_1^{(m)})^2 + \dots + (z_I^{(m)})^2} \sqrt{(z_1^{(a)})^2 + \dots + (z_I^{(a)})^2}} \end{aligned} \quad (2)$$

式(2)所示的相似度余弦指标的增强计算形式中, 增加了惩罚项, 作用是对向量主题 $z^{(m)}$ 和 $z^{(a)}$ 之间的元素差进行放大, 具有越大的元素差, 则式(2)所示的惩罚项的取值也越大. 利用惩罚参数 λ 对上述公式中的惩罚程度进行表征, 如果 $\lambda = 0$, 那么式(2)所示增强型相似度余弦指标计算公式转化成一般形式的相似度余弦指标计算形式. 利用 APIs 文档及其标签、标签 s 间存在的关系组合 y 当做 RTM 模型的训练输入, 利用 Monte Carlo-Markov chain 形式对其进行采样即可得到 $\theta^{(m)}$ 、 $\theta^{(a)}$ 、 φ 、 $z^{(m)}$ 和 $z^{(a)}$

等模型参数的训练结果.

3.3 微博标签 RTM 的 Gibbs 采样推理

通过 Gibbs 采样可实现在如主题模型的高维模型下的简单算法实现, 因此在本文微博标签 RTM 模型中也采用 Gibbs 采样方式进行算法简化. 在给定的 T, W, H 和 G 前提下, 需要导出潜变量的后验分布, W 表示数据集内所有词, T 表示数据集中的所有主题, H 表示数据集的所有微博标签, G 是指在数据集的所有微博标签的旗帜标志. 因此, 示例过程给出以下两种条件分布的近似. 此外, 以下公式不包括超参数. 潜变量 $t_{u,n}$ 的后验分布形式:

$$\begin{aligned} P(t_c = t | T_{-c}, W, H, G) \propto & \frac{n_{t,-c}^{t_c} + \alpha_{t_c}}{\sum_{t_1=1}^T (n_{t,-c}^{t_1} + \alpha_{t_1})} \times \\ & \frac{\prod_{w=1}^V \prod_{i=0}^{n_c^w} (n_{t,-c}^w + \beta_w + i)}{\prod_{k=0}^{n_c^w} (\sum_{w=1}^V (n_{t,-c}^w + \beta_w) + k)} \times \\ & \frac{\prod_{h=1}^H \prod_{i=0}^{n_{c,g=1}^h} (n_{t,-c,g=1}^h + \gamma_h + i)}{\prod_{h=1}^H (\sum_{h=1}^H (n_{t,-c,g=1}^h + \gamma_h) + k)} \end{aligned} \quad (3)$$

令 $c = \{u, n\}$ 表示用户 u 的第 n 个微博, T_{-c} 表示除用户 u 的第 n 个微博外的所有微博. n_c^w 表示 c 中词语 w 的出现次数. n_c^w 表示 c 中的单词总数. $n_{c,g=1}^h$ 指的是 c 中的主题微博标签次数, $n_{c,g=1}^H$ 指的是 c 中的主题微博标签总数. 潜变量 $g_{u,n,h}$ 抽样条件分布:

$$\begin{aligned} P(g_f = 0 | T, W, H, G_{-f}) \propto & \frac{n_{g_f}^{g_f} + \rho_{g_f}}{\sum_{g=0}^G (n_{g_f}^g + \rho_g)} \times \\ & \frac{n_{g=0,-f}^{h_{g=0,-f}} + \lambda_{h_f}}{\sum_{h=1}^H (n_{g=0,-f}^h + \lambda_h)} \end{aligned} \quad (4)$$

$$\begin{aligned} P(g_f = 1 | T, W, H, G_{-f}) \propto & \frac{n_{g_f}^{g_f} + \rho_{g_f}}{\sum_{g=0}^G (n_{g_f}^g + \rho_g)} \times \\ & \frac{n_{g=1,-f}^{h_{g=1,-f}} + \gamma_{h_f}}{\sum_{h=1}^H (n_{g=1,-f}^h + \gamma_h)} \end{aligned} \quad (5)$$

令 $f = \{u, n, h\}$ 表示用户 u 的第 n 个微博的第 h 个微博标签. 因为若 $g = 0$, 则在全局微博标签分布中进行微博标签采样; 若 $g = 1$, 则在主题微博标签分布中进行微博标签采样, 对两种情况进行分别讨论.

利用式(7) ~ (9), 可以通过 Gibbs 采样对未知参数进行估计. 模型参数 $\theta, \varphi, \omega, \pi$ 和 σ 可利用采样样

本估计, $\theta_{u,t}$ 表示用户主题分布, 其类似于 RTM 分别. 其用来表示推荐过程的用户特征向量:

$$\theta_{u,t} = \frac{n_u^t + \alpha_t}{\sum_{t_1=1}^T (n_u^{t_1} + \alpha_{t_1})} \quad (6)$$

式中, σ_g 是全局微博标签的概率, ω_h 是微博标签从主题微博标签分布中进行采样的概率. 可用来区分全局标签形式:

$$\omega_h = \frac{n_{g=0}^h + \lambda_h}{\sum_{h_1=0}^H (n_{g=0}^{h_1} + \lambda_{h_1})}, \sigma_g = \frac{n^g + \rho_g}{\sum_{g=0}^1 (n^{g_1} + \rho_{g_1})} \quad (7)$$

令 $\pi_{t,h}$ 表示主题微博标签分布. 利用其表示话题和微博标签之间的关系:

$$\pi_{t,h} = \frac{n^g + \rho_g}{\sum_{g=0}^1 (n^{g_1} + \rho_{g_1})} \quad (8)$$

3.4 计算复杂度分析

在所提的微博个性化标签图形化 RTM 模型 Gibbs 采样推荐算法中, 主要包含两部分计算过程:

(1) 关系主题 RTM 模型训练; (2) 两个 top-k 推荐过程 (标签频率计算和用户推荐). 其中, 在第一步中, 关系主题 RTM 模型训练采用的是 Gibbs 采样过程, 该过程中每次进行迭代计算的复杂度为 $O(T \cdot N)$, 微博的词汇总数为 N , 主题训练数量为 T , 如果关系主题 RTM 模型的迭代总数是 $Iter$, 那么可得 RTM 模型训练过程的复杂度是 $O(T \cdot N \cdot Iter)$. 一般在进行关系主题 RTM 模型训练过程中是离线方式进行计算, 因此其计算复杂度可不考虑在算法计算过程中, 而是通过离线形式进行训练, 对于推荐过程的计算时效性没有影响.

4 实验分析

4.1 实验设置

为实现对所提的微博标签推荐效果进行分析, 选取当前比较流行的几种算法进行对比, 见引言部分所列几种文献: TF 算法、TF-IDF 算法、LDASA 算法、RTMSA 算法和 RTMCS 算法, 上述所有算法均采用 Java 语言实现. 在本文算法的实现过程中, 采用 Phan 等学者编写的基于 Java 语言的 Gibbs 采样源代码 JGibbRTM. 实验硬件环境: i7-4720HQ 2.6 GHz, 内存大小 4 GB, 算法运行环境 Ubuntu GNU/Linux 10. 上述算法参数按照原文献中的描述进行描述.

实验数据选取: 选取搜狐微博抓取的数据集作为测试集, 该测试集保持了真实环境数据特点, 对于

验证算法在真实应用中的效果具有一定知道意义. 测试集构建方式: 对搜狐微博的最新列表进行监控, 并采用间歇方式进行数据抓取, 抓取时间是 2015 年 9 月~2016 年 9 月, 抓取的微博总数是 24 万条. 该测试集中的文章属性包含微博标签、创建时间、微博作者、微博正文以及标题等主要信息, 所抓取数据中每篇微博的平均标签数量是 5.32 个. 在进行测试集创建之初, 对用户要求进行互斥类别划分, 因此测试集中的微博被划分成 22 组具有互斥特性的类别.

虽然所提算法可实现微博标签的自动分类, 但是对算法性能进行评估需要人工进行, 需要非常大的计算量, 因此在实验过程中我们选取其中较小的数据子集进行测试. 在选择子集中, 所有的博文限定在 10 组分类组别中, 并对推送博文数量小于 60 的微博用户进行过滤, 然后通过随机方式选取约 50 名用户的 1 500 篇博文进行测试. 同时, 对博文的标签以及正文内容进行分词操作, 对使用次数不多的词汇标签进行事先滤除, 经上述操作后, 获得的最终测试集包含 32 组用户, 583 个微博标签, 1 456 个文章组成的测试集. 最后, 为降低微博标签推荐过程的计算量, 并提高 Gibbs 采样过程的计算速度, 测试过程中仅使用博文内前 450 组词作为标签资源替代描述内容.

4.2 算法性能分析

在进行微博标签 RTM 模型训练过程中, 有关参数设置: 迭代上限 $Iter$ 设置为 1 500, 标签主题数 T 设置为 8. 模型的超参数分别设置如下: $\alpha = 2.0$, $\epsilon = 0.5$, $\lambda = 0.3$ 以及 $\beta = \delta = \gamma = 0.1$. 对比算法如 4.1 节所述. 对相同的资源, 在结果推荐中将其无差别的进行合并, 并通过人工进行性能评估. 所采用的评估指标为准确率以及召回率形式为

$$\begin{cases} precision = TP / (TP + FP) \\ recall = TP / (TP + FN) \end{cases} \quad (9)$$

式中, TP 表示在人工算法性能评估与算法执行中认为均需要的标签总数; FP 表示算法执行中认为需要但人工评估认为不需要的标签总数; FN 表示算法执行中认为不需要但人工评估认为需要的标签总数. 同时, 为了综合上述两个指标, 还可对准准确率以及召回率指标进行合并操作, 获得微博标签推荐的 F1 指标形式:

$$F1 = (2 \times precision \times recall) / (precision + recall) \quad (10)$$

图 4~6 分别给出本文所提微博推荐方法与选取的 TF 算法、TF-IDF 算法、LDASA 算法、RTM-

SA 算法和 RTMCS 算法,在选取的微博标签推荐测试集上的算法推荐性能评估结果对比。

根据图 4~6 所示的算法推荐性能的评估结果可知,对于选取的准确率、召回率以及 F1 指标结果,如引言所述采用主题模型的微博标签推荐结果(RTMCS、RTMSA、LDASA 以及本文算法)要明显优于采用关键词方式获得的微博标签匹配推荐结果(TF-IDF 算法和 TF 算法)。产生上述现象的原因在于,微博标签通常呈现出两个明显特点:(1)微博标签标注的语义具有模糊属性;(2)微博用户在进行标签标注时具有较大随意性。这两个特点表明,通过语义匹配方式进行标签推荐,会产生较大的关键词的不匹配性,导致算法推荐效果不好,从实验结果看,在标签推荐设定数量为 1~6 情况下,TF 算法的推荐精度始终处于 20%以下,召回率指标处于 20%左右,F1 指标几乎处于 20%以下,TF-IDF 算法采用语义层次方式对其进行了改进,并且实现了算法性能的较大提升,但是效果也并不是很好,这主要是因为关键词匹配方式算法在微博推荐过程中存在的局限性所致。

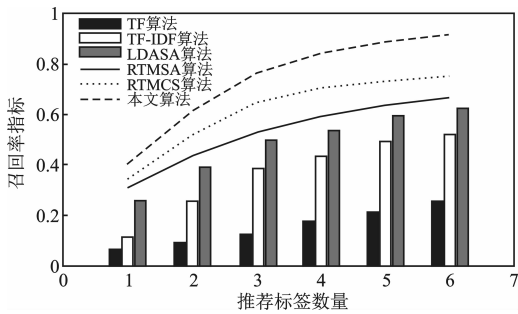


图 4 召回率比较

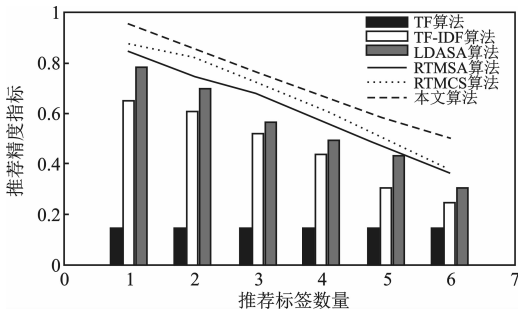


图 5 推荐精度比较

同时可知,在三种主题模型的微博标签推荐过程中,本文算法的准确率、召回率以及 F1 指标均要优于选取的对比算法。随着推荐标签设定数量的增加,选取的几种算法的召回率指标均随之增加,表明正确进行推荐的微博标签的数量也在增加。但是,微

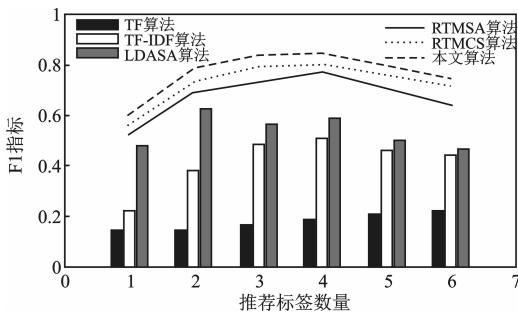


图 6 F1 指标比较

博标签数量上限是固定值,标签数量的增加所带来的召回率增加也会逐渐趋于稳定。而推荐精度指标与召回率指标情形正好相反,随推荐微博标签数量的增加,推荐的不准确的标签数量也在同步增加,从而导致推荐精度出现下降。召回率和推荐精度两种指标的一升一降,反应在 F1 指标中便是随着微博推荐标签设定数量增加,出现先增加后降低现象。从对比结果可知,本文算法的 F1 指标相对于选取的对比算法:TF 算法、TF-IDF 算法、LDASA 算法、RTMSA 算法和 RTMCS 算法,F1 指标分别提高 368%、75.3%、25.6%、16.7%和 8.5%,提现了所提算法在推荐性能指标上的优势。

4.3 算法参数影响

这里主要对 RTM 模型中平滑参数 λ 和式(1)中的指标权重取值进行实验分析。为了测试该参数变量的敏感度,本文设置 λ 的取值范围为 0 到 1,递增区间长度为 0.1,推荐标签数量设置为 6,随平滑参数 λ 不同取值,本文标签推荐算法的性能评估结果如图 7 所示。

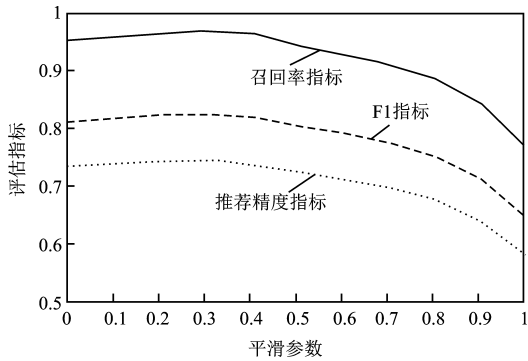


图 7 平滑参数影响实验

从图 7 中可以看出,随平滑参数 λ 取值增大,准确率、召回率以及 F1 指标结果均呈现出先增加后降低的情况,在平滑参数 λ 取值为 0.3 时,本文算法微博标签的推荐性能相对最好。

同时设置权重 ϵ 的取值范围为 0 到 1,递增区间

长度为 0.1,结果见图 8 所示.

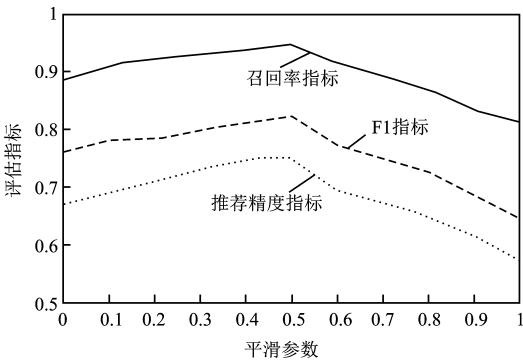


图 8 权重参数影响实验

从图 8 中可以看出,权重参数 ϵ 的影响情况与平滑参数 λ 的影响情况类似,准确率、召回率以及 F1 指标也呈现出先增加后降低的情况,当权重参数设置为 $\epsilon=0.5$ 时,本文微博标签的推荐性能最佳.

5 结束语

本文提出基于 Gibbs 采样推理的微博个性化标签隐含关系主题模型推荐算法,利用图形化形式对微博中的潜在局部信息进行表达,利用带惩罚项的增强型余弦相似度 RTM 模型对微博标签进行命名,实现了对全局标签和主题之间关系的有效挖掘.实验结果验证了所提算法的性能优势.今后研究重点主要有:(1)考虑非常规微博标签的推荐问题;(2)对所提算法在微博标签以外情形的推荐效果进行验证;(3)上述 RTM 模型是离线方式进行训练,考虑设计增量式的在线训练方式.

参考文献:

[1] 刘少鹏,印鉴,欧阳佳,等. 基于 MB-HDP 模型的微博主题挖掘[J]. 计算机学报, 2015, 38(7): 1408-1419.

[2] Wang L, Ranjan R, Kolodziej J, et al. Software tools and techniques for big data computing in healthcare clouds[J]. Future Generation Computer Systems,

2015, 43(2): 38-39.

[3] 杨长春,徐小松,叶施仁,等. 基于文本相似度的微博网络水军发现算法[J]. 微电子学与计算机, 2014, 31(3): 82-85.

[4] 邢千里,刘列,刘奕群,等. 微博中用户标签的研究[J]. 软件学报, 2015, 26(7): 1626-1637.

[5] Xiong X, Zhou G, Huang Y, et al. Dynamic evolution of collective emotions in social networks: a case study of sina weibo[J]. Science China Information Sciences, 2013, 56(7): 1-18.

[6] Wu Xinmiao, Wang Jianmin. Micro-blog in China: identify influential users and automatically classify posts on Sina micro-blog[J]. Journal of Ambient Intelligence and Humanized Computing, 2014, 5(1): 51-63.

[7] Ye Peng, Wang Changbo, Liu Yuhua, et al. Visual analysis of micro-blog retweeting using an information diffusion function[J]. Journal of Visualization, 2016, 19(4): 823-838.

[8] Zhu Jin-Qi, Li Lu, Ma Chun-Mei. From interest to location: neighbor-based friend recommendation in social media[J]. Journal of Computer Science and Technology, 2015, 30(6): 1188-1200.

[9] Xin Xin, Lin Chin-Yew, Wei Xiao-Chi, et al. When factorization meets heterogeneous latent topics: An interpretable cross-site recommendation framework[J]. Journal of Computer Science and Technology, 2015, 30(4): 917-932.

[10] Giovanni Stilo, Paola Velardi. Efficient temporal mining of micro-blog texts and its application to event discovery[J]. Data Mining and Knowledge Discovery, 2016, 30(2): 372-402.

作者简介:

刘真臻 女,(1992-),硕士研究生.研究方向为多媒体技术、文本挖掘、推荐算法. E-mail:liuzhenzhenang@sohu.com.

徐东平 男,(1958-),博士,教授.研究方向为多媒体技术.