实时个性化微博推荐系统

刘慧婷 程 雷 郭孝雪 赵 鹏
(安徽大学计算机科学与技术学院 合肥 230601)

摘 要 目前很多社交网络服务对用户的个性化需求考虑得不充分,并且社交网络服务由于需要处理海量数据而难以保障服务的实时性。为了实时响应用户在微博推荐中的个性化请求,提高推荐的效率和质量,提出了一种基于LDA主题模型和KL散度相结合的RPMPS微博推荐模型。RPMPS推荐模型不但通过文档-主题概率分布矩阵获得了用户信息与待推荐微博的主题相似性,而且还通过文档-词来对词频概率进行统计,从而获得用户信息与待推荐微博的内容相似性。最后,基于RPMPS推荐模型构建实时个性化微博推荐系统,并在数据处理过程中对微博进行过滤以缩短系统的响应时间。通过真实数据集验证了系统可较好地满足用户的实时个性化需求。

关键词 社交网络,微博,RPMPS推荐模型,推荐系统
中图法分类号 TP319 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.09.042

Real-time Personalized Micro-blog Recommendation System

LIU Hui-ting CHENG Lei GUO Xiao-xue ZHAO Peng
(School of Computer Science and Technology, Anhui University, Hefei 230601, China)

Abstract At present, many social networking services do not fully consider the personalized needs of users, and it is difficult to guarantee the real-time services because social networking services need to deal with massive amounts of data. A micro-blog recommendation model called RPMPS based on LDA topic model and KL divergence was proposed to respond to users' personalized request in micro-blog recommendation in real time and improve the efficiency and quality of recommendation. RPMPS model not only uses the document-topic probability distribution matrix to get the similarity between the topic of user information and the topic of candidate micro-blog, but also obtains the similarity between the content of user information and the content of candidate micro-blog by utilizing the document-word to count the probability of the word frequency . At last, the real-time personalized micro-blog recommendation system based on RPMPS model is constructed, and micro-blog is filtered during the course of data processing to shorten the system response time. Experimental results on real-world datasets demonstrate that the system can meet the real-time personalized demands of users.

Keywords Social network, Micro-blog, RPMPS recommendation model, Recommendation system

1 引言

微博是Web 2.0[1]时代最典型的社会网络服务媒体,用户可以通过微博随时随地向公众发布信息以及获取信息,是实现社会交往的重要途径[2]。然而,由于微博拥有数量庞大的用户,他们每天发布数以万计的微博信息,如果让人们在海量的微博信息中判断自己感兴趣的微博无疑是很困难的,况且很多微博并没有实际意义[3]。因此,如何及时地在海量信息中为用户提供个性化的服务已成为社会网络服务研究的热门领域。

推荐系统主要通过分析用户、项目(物品)的信息以及其他相关信息,来获得用户对项目(物品)的偏好,并根据偏好特征为用户推荐项目(物品)[4]。当前被广泛使用的推荐算法[5]主要包括:基于协同过滤的推荐算法[6-7]和基于内容的推荐算法[8]。基于协同过滤的推荐算法需要计算用户之间特征的相似性,对于新加入的用户,由于无法提取其特征,因此该类算法存在冷启动问题[9]。基于内容的推荐算法通过用户的历史数据构建用户特征,并进一步与项目(物品)特征进行比较。由于该类算法依赖用户的历史数据,如果不充分考虑时间因素,当用户的兴趣突然发生变化时,将导致推

到稿日期:2017-08-29 返修日期:2017-11-16 本文受国家自然科学基金资助项目(61202227,61602004),安徽高校自然科学研究项目(KJ2018A0013)资助。

刘慧婷(1978-),女,博士,副教授,硕士生导师,CCF会员,主要研究方向为数据挖掘、机器学习,E-mail:htliu@ahu.edu.cn(通信作者);**程 雷**(1990-),男,硕士生,主要研究方向为数据挖掘和推荐系统;**郭孝雪**(1991-),女,硕士生,主要研究方向为数据挖掘和推荐系统;**赵 鹏**(1976-),女,博士,副教授,硕士生导师,主要研究方向为智能信息处理、机器学习。

荐的结果不够准确[10]。

对于微博推荐,不仅需要考虑冷启动和用户兴趣变化的问题,还应考虑实时性以及个性化的问题。例如,Busch 等[11]根据微博的发布时间和受欢迎度对微博排序,实现用户实时检索微博的功能;Otsuka 等[12]根据 TF-IDF 提出了针对微博话题的推荐方法;高明等[13]将 LDA 主题模型与滑动窗口相结合,提出了一种实时个性化微博推荐方法,但是该算法没有考虑用户兴趣随时间变化的特点以及用户的个性化需求;Qiu 等[14]根据用户的微博关系网,提出了一种在社区中基于协同过滤算法的推荐方法;Chen 等[15]把协同过滤与社交网络信息相结合,提高了推荐的质量;蒋超[16]和 Chen[17]等从用户兴趣的角度为用户推荐微博,根据用户兴趣,通过聚类算法对用户分组,以用户组作为个性化推荐对象,减少了数据的处理量。但是,以上算法没有考虑系统的实时性。

LDA 主题模型与 KL 散度在社区发现领域已有一定的应用。辛宇等[18]利用 LDA 主题模型获得节点语义空间,并进一步运用 KL 散度度量节点间语义坐标的相似性。受此启发,为实时响应用户的个性化需求,本文提出了一种基于 LDA 主题模型与 KL 散度的微博推荐模型(Real-time Personalized Micro-blog Recommendation System,RPMPS)。与文献[18]不同的是,RPMPS 模型通过 LDA 获得用户兴趣集与待推荐微博的主题相似度,再结合 KL 散度获得它们之间的内容相似性共同构成总体相似度。RPMPS 模型以用户在滑动时间窗口内发布的微博集合作为用户近期兴趣的组成部分,以解决当用户兴趣变化时推荐不准确的问题,并在数据处理阶段通过对微博进行过滤来提高时间效率,最终基于 RPMPS 推荐模型设计并实现了微博推荐系统。实验证明,该系统不仅保证了推荐质量,而且实现了对微博用户请求的及时个性化响应。

本文的主要工作和创新如下:

1)定义了用户的兴趣集,在一定程度上解决了冷启动的问题;设置了一个时间窗口,可以更好地代表用户近期的兴趣分布,解决了复杂兴趣推荐问题;以微博发布者的标签作为微博的初始分类,减少了数据处理的工作量,可以较为及时地获得推荐结果,从而保证了系统的实时性。

2)通过 RPMPS 推荐模型,可在文档-主题层获取用户信息及待推荐微博潜在的主题分布,以计算它们之间的相似度,并在文档-词层得到基于内容相似度的结果。结合主题相似度和内容相似度,将得到的微博集合进行综合排序后作为给用户推荐的微博内容,体现了对用户请求的个性化响应。实验表明,所提方法能够获得更理想的推荐结果。

2 系统模型

2.1 相关定义

为更好地理解本文提出的推荐模型,本节先给出相关的定义。

定义 1(用户兴趣集) 用户兴趣集是用户兴趣的体现,本文把用户兴趣集定义为 $U_m = U_{\text{Label}} + U_{\text{Mess}} + U_{\text{blog}}$。其中,$U_m$ 代表用户的兴趣集;U_{Label} 表示系统中用户的标签;U_{Mess} 代表系统中用户的个人简介,个人简介是用户在系统中注册时的必填项,用以解决冷启动问题;U_{blog} 表示在时间窗口 T 内当前用户微博集合的内容和主题信息。

定义 2(微博的类号) 本文把微博的类号定义为 $D_{\text{Label}} = U_{\text{Label}}$。其中,$D_{\text{Label}}$ 代表微博的类号,U_{Label} 代表发布该微博用户的标签号。微博本身虽然没有标签,但是微博用户自身带有标签,并且用户更倾向于发布与自身标签相关的微博,提高了系统的实时性,减少了数据的处理量。本文为微博指定类号,微博的类号与发布该微博用户的标签号一致。

定义 3(微博的重要性) 微博重要性 $I = co + re$。其中,I 代表微博的重要性,co 代表该微博的评论数,re 代表该微博的转发数。因为在特定时间段内微博的转发数和评论数可以很好地反映该微博的流行度,所以本文用微博的转发数和评论数作为微博重要性的体现。

定义 4(基于 RPMPS 模型的相似度) $similar = (1-\lambda)S_{\text{Topic}} + \lambda S_{\text{Content}}$。其中,$similar$ 为基于 RPMPS 模型的用户兴趣集与待推荐微博的总体相似度;S_{Topic} 为在 RPMPS 模型中用户兴趣集与待推荐微博的主题相似度;S_{Content} 为 RPMPS 模型中用户兴趣集与待推荐微博的内容相似度;λ 是权重因子,用于调节两者的影响程度。

2.2 系统架构

基于 RPMPS 模型的微博推荐系统如图 1 所示,该系统主要包括微博数据源模块、数据处理模块、推荐引擎模块、推荐结果处理模块等。当用户登录系统选择微博推荐功能后,系统首先结合当前用户的信息从微博数据源模块获取微博,并通过数据处理模块得到待推荐微博;然后进一步通过推荐引擎模块的 RPMPS 推荐算法获取推荐结果;最后通过推荐结果处理模块生成 TOP-K 推荐列表并返回给用户查看。当用户退出系统时,系统会自动清空用户的推荐列表。

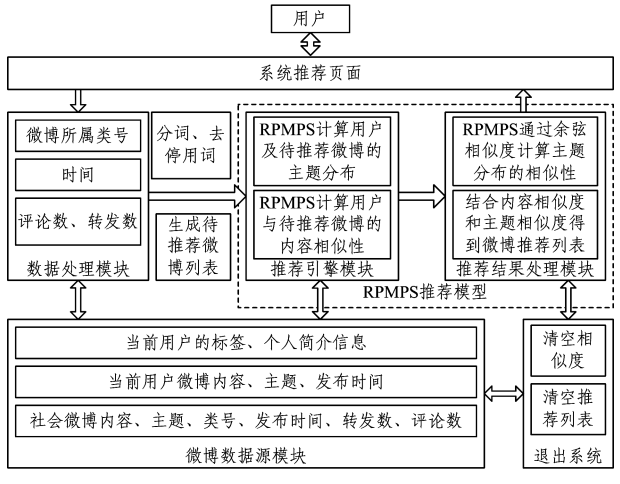

图 1 基于 RPMPS 模型的微博推荐系统的基本架构

Fig. 1 Basic architecture of micro-blog recommendation system based on RPMPS model

1)数据源模块:该模块主要包含了数据的来源及信息的格式,包括当前用户的标签和个人简介,当前用户的微博内容、主题、发布时间以及社会微博内容、主题、类号、发布时间、

转发数、评论数。该模块主要为推荐系统提供数据来源。

2）数据处理模块：数据的处理直接影响着推荐的效果，该模块通过类号、用户登录系统的时间间隔以及评论数、转发数等因素的共同调节获得较高质量的待推荐微博集合，并对获得的微博内容进行分词、去停用词等预处理操作。

3）推荐引擎模块：该模块是系统的核心组成部分，是 RPMPS 模型的主要实现部分。通过 RPMPS 模型，可以得到用户和待推荐微博的主题分布以及用户和待推荐微博之间的内容相似性。

4）推荐结果处理模块：在此模块中，RPMPS 模型通过余弦相似度计算用户和待推荐微博的主题分布，从而得到用户与待推荐微博的主题相似度，然后再结合 RPMPS 模型得到内容相似度，最终生成 TOP-K 微博推荐列表返回给当前用户。

3　模型研究

3.1　LDA 主题模型与 KL 散度

LDA(Latent Dirichlet Allocation)主题模型[19]是 Blei 于 2003 提出的，其在 PLSA 模型[20]的基础上引入了狄利克雷先验分布并加入了文档层。LDA 主题模型是基于词袋模型设计的，即忽略了词语之间的顺序，具有良好的降维功能，可以把高维度的、稀疏的文档-词分布转换为低维度的文档-主题概率分布。LDA 是一种非监督的主题模型，在自然语言处理中广泛应用于文档集或语料库中潜在主题信息的识别。

KL 散度(Kullback-Leibler Divergence)又称为相对熵，在信息检索中的应用较为广泛[21]。本文通过计算用户兴趣集的词频率分布 p 和待推荐微博信息的词频率分布 q，可进一步得到用户与待推荐微博之间的内容相似性。KL 散度值越小，代表相似度越高，其计算公式为：

$$KLdistance(p \parallel q)=\sum_{i=1}^{n} p(x)\log\frac{p(x)}{q(x)} \tag{1}$$

式(1)由于不具有对称性，因此在计算中需要计算 p 对 q 的相对熵以及 q 对 p 的相对熵，最后用两者的平均值来衡量文本之间的相似性。

$$Distance(p,q)=\frac{1}{2}Dis(p \parallel q)+\frac{1}{2}Dis(q \parallel p) \tag{2}$$

3.2　RPMPS 推荐模型

LDA 是一种无监督的主题模型，并不擅长处理短文本。为了在较少的时间开销下提高推荐的准确率，本文将 LDA 主题模型与 KL 散度相结合，提出了 RPMPS 推荐模型，如图 2 所示。其中，给定一个文档集合 $D=\{d_1,d_2,\cdots,d_M\}$，其中 M 是文档的个数，K 是主题的数目，N 是文档下词的数目，z 和 w 分别代表一个具体的主题和词。α 和 β 是狄利克雷先验参数，分别用来生成文档-主题矩阵 θ 和主题-词矩阵 φ，U_m 代表用户兴趣集，C_m 代表待推荐微博信息。S_{Topic} 为 U_m 与 C_m 的主题相似度，S_{Content} 为 U_m 与 C_m 的内容相似度，$similar$ 为 U_m 与 C_m 的总体相似度，R_d 为对当前用户推荐的微博列表。

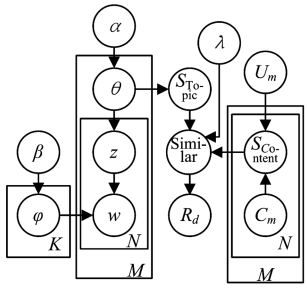

图 2　RPMPS 推荐模型

Fig. 2　RPMPS recommendation model

在 RPMPS 模型中，文档的生成过程如算法 1 所示。

算法 1　RPMPS 模型文档生成算法

1. for each k∈[1,k] do
2.　Sample mixture $\varphi_k\sim\text{Dir}(\beta)$
3. end for
4. for each documents m∈[1,M] do
5.　sample mixture $\theta_m\sim\text{Dir}(\alpha)$
6.　for each word n in m do
7.　sample topic index $z_{m,n}\sim\text{Mult}(\theta)$
8.　sample term for $W_{m,n}\sim\text{Mult}(\varphi_{z_{m,n}})$
9.　end for
10. end for

由算法 1 可知，RPMPS 模型生成 M 个文档集合 D 的概率为：

$$\prod_{m=1}^{M}\int p(\theta_m|\alpha)\Big(\prod_{n=1}^{N_m}p(z_n|\theta_m)p(\varphi_{z_n}|\beta)p(w_n|\theta_{z_n})\Big)\mathrm{d}\theta_m \tag{3}$$

然后使用吉布斯抽样可以完成对 RPMPS 模型的参数估计，从而得到词的主题(见式(4))；吉布斯采样后主题-词矩阵 φ 和文档-主题矩阵 θ 分别通过式(5)、式(6)进行计算。

$$p(z_i=k|\vec{z}_{-i},\vec{w})=\frac{n_{m,-i}^{k}+\alpha}{\sum_{k=1}^{K}n_{m,-i}^{k}+K\alpha}\times\frac{n_{k,-i}^{t}+\beta}{\sum_{t=1}^{V}n_{k,-i}^{t}+V\beta} \tag{4}$$

$$\varphi=\frac{n_{k,-i}^{t}+\beta}{\sum_{t=1}^{V}n_{k,-i}^{t}+V\beta} \tag{5}$$

$$\theta=\frac{n_{m,-i}^{k}+\alpha}{\sum_{k=1}^{K}n_{m,-i}^{k}+K\alpha} \tag{6}$$

图 1 所示的微博推荐系统中，推荐引擎模块及推荐结果处理模块应用了 RPMPS 模型。结合图 1 和图 2 可知，在推荐引擎模块通过 RPMPS 模型吉布斯采样可获得用户与待推荐微博的主题分布矩阵 θ，通过 RPMPS 模型相对熵可计算获得两者的内容相似性 S_{Content}。在推荐结果处理模块，RPMPS 模型根据矩阵 θ 可获得用户信息与待推荐微博的主题相似性 S_{Topic}，主题相似性与内容相似性相结合，得到基于 RPMPS 模型 U_m 与 C_m 的总体相似度，进一步对其排序可获得当前用户推荐的微博列表 R_d。

4　系统算法研究

根据 RPMPS 模型及系统架构设计出本系统对应的微博推荐的整体流程，如算法 2 所示。

算法 2 微博推荐算法

输入:用户信息 U_i 以及社会微博集合 C_d

输出:推荐给用户的 TOP-K 微博列表

1. C_m,U_m,C_1←calc(U_i,C_d)
2. θ←RPMPSGibbs(C_1)
3. $S_{Content}$←RPMPScalc(U_m,C_m)
4. S_{Topic}←RPMPScosine(θ)
5. R_d←TOP-K((1−λ)S_{Topic}+λ$S_{Content}$)

该推荐流程的整体思想为:根据用户 U_i 以及社会微博集合 C_d 获得用户兴趣集 U_m、待推荐微博列表 C_m,以及 U_m 和 C_m 的合集 C_1(算法 2 第 1 行,详细流程请参见算法 3);然后对 C_1 进一步运用 RPMPS 模型计算用户和待推荐微博的主题分布 θ,以及 U_m 与 C_m 之间的内容相似度 $S_{Content}$(算法 2 第 2−3 行,详见算法 4);最后利用余弦相似度计算 U_m 与 C_m 的主题相似性 S_{Topic},并结合内容相似度生成 TOP-K 推荐列表 R_d 并返回到用户显示页面(算法 2 第 4−5 行,详见算法 5)。当用户退出系统时,相似度以及推荐列表信息会自动清空。

4.1 数据源及数据处理模块算法

微博数据处理算法具体如算法 3 所示。

算法 3 微博数据处理算法

输入:用户 U_i 以及社会微博集合 C_d

输出:待推荐的微博集合 C_m,用户兴趣集 U_m 的合集 C_1

1. for login user U_i do
2. U_{ID},U_{Label},U_{Mess}←get(ID,Label,Message)
3. t_0←get(ULastLogin)//获取用户上次登录系统的时间
4. t←get(data)//获取当前时间
5. time←t−T//T 是区分用户近期兴趣的时间窗口阈值
6. U_{blog}←get(UID,time)
7. U_m←Nlp(U_{Label}+U_{Mess}+U_{blog})
8. if(t−t_0>T_1)//T_1 代表数据来源的最短时间间隔
9. C←get(U_{Label},t_0,co,re)
10. else
11. C←get(U_{Lable},t−T_1,co,re)
12. end if
13. C_m←Nlp(C)
14. C_1=U_m∪C_m

在数据库中查找登录系统的当前用户的 ID、标签 U_{Label} 以及个人简介信息 U_{Mess},在时间窗口 T 内查询当前用户发布的历史微博信息 U_{blog},共同构成用户微博集以代表用户的近期兴趣,再结合用户的标签构成用户的兴趣集 U_m(算法 3 第 1−7 行)。在社会用户发布的微博中选取与当前用户标签号(微博类号)一致的微博作为推荐给当前用户的微博来源。然后根据微博类号 U_{Label}、用户登录系统的时间间隔 $t-T_1$ 或 t_0,以及微博自身的转发数 re 和评论数 co 对微博进行排序,过滤得到待推荐微博列表 C_m,并求其与 U_m 的合集 C_1(算法 3 第 8−14 行)。同时,对用户兴趣集和待推荐微博列表进行中文分词、去停用词等预处理操作。

4.2 推荐引擎模块算法

推荐引擎模块算法具体如算法 4 所示。

算法 4 RPMPS 模型推荐引擎算法

输入:待推荐的微博集合 C_m 和用户兴趣集 U_m 的合集 C_1,先验参数 α 和 β,主题的数目 K,迭代次数 max

输出:文档-主题概率分布矩阵 θ,用户兴趣集与待推荐微博的内容相似度 $S_{Content}$

1. count=0,n_{mk}=0,n_m=0,n_{kt}=0,n_k=0
2. for each d_m∈C_1 do
3. for each w in d_m do
4. sample topic index $z_{m,n}$=k∼Mult(1/K)
5. w.value=w.count/d_m.total
6. n_{mk}+=1,n_m+=1,n_{kt}+=1,n_k+=1
7. end for
8. end for
9. while count<max do
10. for each d_m∈C_1 do
11. for each w in d_m do
12. n_{mk}−=1,n_m−=1,n_{kt}−=1,n_k−=1
13. sample topic index k∼P(z_i=k) //根据式(4)生成当前词的新的主题
14. n_{mk}+=1,n_m+=1,n_{kt}+=1,n_k+=1
15. end for
16. end for
17. if count>M
18. calculate θ //根据式(6)计算 θ
19. end if
20. count=count+1
21. end while
22. $S_{Content}$=infinitely
23. for each d_m∈C_m do
24. for each w in d_u do
25. if w in d_m
26. calculate accretion//根据式(2)计算词语的相对熵
27. end if
28. $S_{Content}$=$S_{Content}$+accretion
29. end for
30. end for

首先,初始化迭代计数 $count$、文档 m 下主题 k 出现的次数 n_{mk}、文档 m 的词数 n_m、主题 k 下词 t 出现的次数 n_{kt}、主题 k 的词数 n_k 为 0,然后对文档 d_m 的每个词在 K 个主题下随机分配词的主题 k,以及词在当前文档中出现的词频概率 $w.value$,同时更新 n_{mk},n_m,n_{kt},n_k 的值(算法 4 第 1−8 行)。在吉布斯抽样词主题的过程中首先去除当前词的主题,根据式(4)的参数估计重新分配主题并更新 n_{mk},n_m,n_{kt},n_k 的值(算法 4 第 9−16 行)。对每个词进行重复迭代计算,当迭代次数达到指定阈值 max 后,根据 n_{mk},n_m,n_{kt},n_k 的值利用式(6)可以得到文档-主题概率分布矩阵 θ(算法 4 第 17−21 行)。然后比较用户兴趣集与待推荐微博的相对熵,利用式(2)得到词语的相对熵,最终得到用户兴趣集与待推荐微博之间的内容相似度 $S_{Content}$(算法 4 第 22−30 行)。

4.3 推荐结果处理模块算法

利用余弦相似性来比较用户兴趣集和待推荐微博的主题相似性。余弦相似度的计算公式为:

$$cosine(A,B)=\frac{\sum_{1}^{K}(A_i\times B_i)}{\sqrt{\sum_{1}^{K}A_i^2}\times\sqrt{\sum_{1}^{K}B_i^2}} \tag{7}$$

在得到用户信息与待推荐微博的主题相似性后,再通过RPMPS模型得到用户信息与待推荐微博的内容相似性,进而得到用户信息与待推荐微博的整体相似性。

RPMPS模型结果处理及返回算法具体如算法5所示。

算法5 RPMPS模型结果处理及返回算法

输入:文档-主题概率分布矩阵 θ,用户兴趣集与待推荐微博的内容相似度 $S_{Content}$

输出:TOP-K 推荐列表

1. for each θ_{d_m} do
2. similar=infinitely
3. $S_{Topic}=\text{cosine}(\theta_{d_u},\theta_{d_m})$//根据式(7)计算主题相似性
4. $similar=(1-\lambda)S_{Topic}+\lambda S_{Content}$
5. end for
6. $C_2\leftarrow$Desc(similar)
7. $R_d\leftarrow$TOP-K(C_2)
8. User$\leftarrow R_d$

根据算法4得到的主题分布,利用余弦相似度计算用户兴趣集与社会微博的主题相似性 S_{Topic},结合内容相似度 $S_{Content}$ 形成用户兴趣集与待推荐微博的最终相似性 *similar*(算法5第1—5行),其中 λ 是调节因子,用来权衡主题相似性与内容相似性的影响程度。最后对待推荐微博进行相似性排序,得到 TOP-K 推荐列表 R_d 并将其返回到用户显示页面(算法5第6—8行)。在本微博推荐系统中通过RPMPS推荐模型计算用户与待推荐微博的主题相似度及内容相似度,可以在提升推荐结果准确性的同时保证时间效率,从而满足用户的实时性请求。

4.4 模型复杂度分析

在RPMPS模型中,假设迭代次数为 N,文档数量为 D,词的数量为 W。由于每篇文档的长度不同,假设文档的平均长度为 L,用户兴趣集的长度为 uL,主题的数量为 K。在RPMPS模型中吉布斯采样过程的时间复杂度为O($NDLK$),计算当前用户与待推荐微博之间相似性的时间复杂度为O($DuL+D$),因此RPMPS模型的整体时间复杂度为O($NDLK+DuL+D$)。由于算法需要维护规模为 DK 的文档-主题分布矩阵 θ、规模为 KW 的主题-词分布矩阵 φ、规模为 DL 的文档词的编号映射矩阵和规模为 DL 的所有词的词频概率矩阵,因此RPMPS模型的空间复杂度为O($DK+KW+2DL$)。

5 实验与分析

本文中的实时个性化微博推荐系统基于RPMPS推荐模型,使用Java语言在Windows 7平台的MyEclipse 10软件中搭建而成。实验环境如下:CPU为Intel Core i3-4170,内存为4 GB,OS为64位Windows 7,编程语言为Java,系统运行平台为MyEclipse 10,数据库为MySQL,服务器Tomcat 7.0。

5.1 数据集及实验参数的设置

个性化微博实时推荐系统为了实时响应用户的请求,需要不断地更新数据源,因此爬取新浪微博认证的知名博主的微博作为数据来源。为了降低微博数据的噪音,本文去除了词数量少于10的微博以及微博中对主题挖掘无用的一些特征,如表情符号、URL链接等,最后获得17406条微博数据(见表1)。由于微博本身没有标签,本文把微博发布者的标签号作为这条微博的所在类,再结合时间间隔、评论数和转发数,通过数据处理模块得到待推荐微博集合。实验以军事、体育类别为例,分别选取标签号为军事、体育的知名博主的微博信息,通过数据处理模块对其进行处理,将得到的相关数据作为待推荐微博集合。为了验证微博推荐的效果,本文聘请第三方人员对待推荐微博集合的真实类型进行人工标注[22]。在微博类型的人工标注过程中,请3位微博类型标注员对待处理微博进行类型标注并对标注结果的一致性进行检测,最终得到的微博类别信息如表2所列。由于通过吉布斯采样获取主题分布时,迭代200余次后,文档-主题分布即可趋于稳定,因此为了保证推荐的质量及效率,实验中设置吉布斯采样的迭代次数为300,主题数量为50,先验参数 α 为50除以主题数,β 值为0.01。

表1 微博数据集

Table 1 Micro-blog dataset

数据来源	原始微博数量	处理后的微博数量
新浪微博	22961	17406

表2 待推荐微博信息

Table 2 Candidate micro-blog information

(单位:%)

标签号	符合类别的微博比例	不符合类别的微博比例
体育	63.8	36.2
军事	52	48

5.2 评价指标

评价实时个性化微博推荐系统的标准一般有推荐的时间效率、推荐的质量以及推荐的多样性,它们分别反映了微博推荐系统的实时响应性、推荐性能和在推荐过程中用户的个性化请求。其中,推荐的质量可以用准确率和召回率来衡量。

准确率:排序后得到的TOP-K条推荐微博列表中推荐正确的微博所占的比例。其计算方式如式(8)所示:

$$Precision=\frac{\text{推荐结果符合用户要求的微博数量}}{\text{推荐的总数量}}=\frac{R_u\cap T_u}{R_u} \tag{8}$$

召回率:推荐结果中符合用户要求的微博数量占所有符合用户要求的微博数量的比例。其计算公式如式(9)所示:

$$Recall=\frac{\text{推荐结果中符合用户要求的微博数量}}{\text{符合用户要求的微博数量}}=\frac{R_u\cap T_u}{T_u} \tag{9}$$

其中，R_u 表示微博推荐系统推荐给用户 u 的微博集合，T_u 表示在 R_u 中与用户 u 兴趣一致的微博集合。

5.3 实验方案

为了验证本文所提模型的推荐效果，将进行以下 4 组实验：

1)采用 LDA 主题模型得到个性化推荐列表；

2)采用 KL 散度得到个性化推荐列表；

3)采用基于用户长短期兴趣的微博推荐方法 LSI[10]得到个性化推荐列表；

4)采用本文提出的 RPMPS 推荐模型得到个性化推荐列表。

5.3.1 微博推荐系统的推荐质量

为了体现实时个性化推荐系统的推荐质量，本文以准确率、召回率作为评价标准，以军事类别为例(见表 3)，通过本文的实时个性化推荐系统以及 LDA 主题模型、KL 散度和 LSI 方法给当前用户推荐 50，100，150，200，250，300 条微博，并计算对应的准确率 P、召回率 R 以及在 RPMPS 模型中不同权重因子 λ 对应的推荐准确率。

表 3 准确率和召回率

Table 3 Accuracy and recall

评价指标	准确率						召回率					
推荐数量	T-50	T-100	T-150	T-200	T-250	T-300	T-50	T-100	T-150	T-200	T-250	T-300
RPMPS	0.86	0.76	0.70	0.67	0.64	0.63	0.17	0.29	0.40	0.52	0.62	0.73
KL	0.84	0.78	0.70	0.64	0.64	0.62	0.16	0.30	0.40	0.49	0.62	0.71
LSI	0.68	0.65	0.64	0.62	0.62	0.60	0.13	0.25	0.37	0.47	0.60	0.69
LDA	0.62	0.6	0.58	0.57	0.56	0.55	0.12	0.23	0.33	0.44	0.54	0.66

从表 3 可以看到，4 种模型得到的推荐准确率都是随着微博推荐条数的增加而不断降低，这主要是因为推荐结果以相似度排序，越靠后的微博的相似性越低。RPMPS 模型的推荐效果明显好于 LDA 主题模型得到的推荐结果，这是因为 LDA 主题模型是非监督主题模型，且不擅长处理短文本，因此得到的推荐精度较低。而使用 RPMPS 模型得到的推荐效果和使用 KL 散度得到的精度较为相似，KL 散度虽然可以较为准确地发现与用户兴趣相关的微博，但是由于其只是在文档-词层面进行统计，因此不能发现用户的潜在兴趣。RPMPS 模型不仅能在文档-词层面得到用户与待推荐微博的直接相似性，还联合了文档-主题充分挖掘用户隐含的主题信息。从图 3 中可以看出，权重系数 λ 越大，对应的推荐准确率越高。权重因子 λ 用于在 RPMPS 模型中调节内容相似性与主题相似性之间的权重，λ 越大，说明 RPMPS 模型的推荐结果更倾向于内容相似性，可获得较高的推荐准确率，但是推荐的多样性效果较差。多次实验验证表明，当 λ 取值为 0.8 时，既能获得较高的推荐准确率，又能保证推荐的多样性。在召回率方面，由于微博推荐条数不断增加，推荐正确的微博数量也随之增加，因此召回率不断提高。

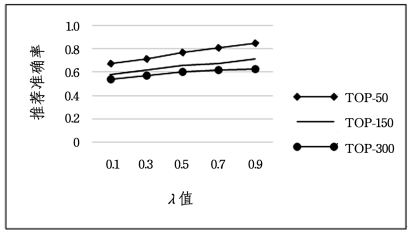

图 3 λ 值对准确率的影响

Fig. 3 Impact of λ value on accuracy

5.3.2 实时个性化微博推荐系统的时间效率

为了体现个性化推荐系统的实时性，本文将推荐流程的耗时作为评价标准。推荐流程的耗时为从用户点击“推荐微博”功能按钮开始到系统生成 TOP-K 推荐列表为止的时间间隔。通过在微博推荐系统程序中的相应位置加入时间统计函数，设置推荐微博数目为 300 条，统计得到数据处理模块、推荐引擎模块及推荐结果处理模块所需的时间分别为0.808 s，7.273 s，5.388 s，分别占整个推荐流程时间的 6%，54% 和 40%(见图 4)。RPMPS 模型分别与 LDA 主题模型和 KL 散度完成推荐的耗时对比如图 5 所示。

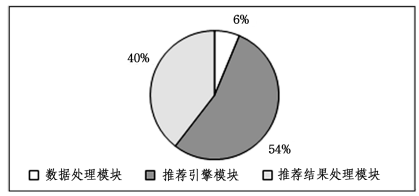

图 4 推荐模型中各模块的耗时比例

Fig. 4 Time-consuming ratio of each module in recommendation model

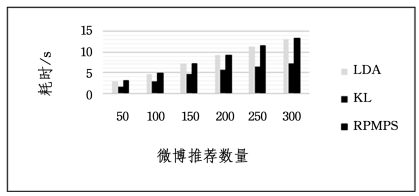

图 5 3 种模型在不同推荐数量下的耗时对比

Fig. 5 Time-consumtion comparison of three models under different recommendation quantities

从图中可以看到，3 种模型由于都需要过滤微博形成待推荐微博，在预处理后通过相应算法形成推荐结果，因此都需要消耗一定的时间。由于推荐数量不断增加，处理的数据也不断增加，因此需要更多的处理时间。由于 KL 散度直接在文档-词层进行处理，因此其耗时相对较少；而 LDA 主题模型由于需要迭代产生文档-主题矩阵，因此比 KL 散度的耗时长。使用 RPMPS 推荐模型的耗时和使用 LDA 主题模型的耗时相差不大，但是结合表 3 发现，使用 RPMPS 模型比使用

LDA 模型的推荐效果有明显的提升。如图 5 所示,针对 RPMPS 推荐模型而言,当推荐微博的数量为 50 条时,推荐流程的耗时为 3.009 s,当数量提升到 200 条时,推荐流程耗时为 9.309 s,基本满足了用户的实时性请求。

5.3.3　推荐的多样性

推荐的多样性反映了推荐系统对用户兴趣的识别程度,是对用户个性化请求的体现。本文以新浪微博中关于体育类别的数据为例(见表 2),分别使用 RPMPS 模型和 KL 散度得到相应的结果,如表 4 所列,由于体育类别中相关的微博数目比军事类别中的多(见表 2),因此体育类别的推荐准确率较军事类别也有一定的提升。从表 4 中看到,虽然使用 KL 散度得到的推荐结果的精确度和使用 RPMPS 模型得到的推荐精度相似,但是通过 KL 散度得到的推荐结果更倾向于体育的一个子集——篮球。在表 4 中,通过 RPMPS 模型得到的推荐结果中除篮球外属于体育的微博数量明显多于通过 KL 散度得到的结果,这主要是由于 KL 散度通过对用户信息和待推荐微博信息在文档-词层面进行词频概率的统计计算,可以较为准确地发现与用户兴趣直接相关的微博,但是并不能体现出用户兴趣的隐含信息。RPMPS 模型可以通过降维充分挖掘用户兴趣隐含的主题信息,因此既可以保证推荐结果的精度,又可以充分挖掘用户隐含的主题信息,使得推荐结果多样化,满足用户的个性化需求。

表 4　推荐的多样性

Table 4　Recommendation diversity

微博的不同推荐数量/条	推荐指标	KL	LDA+KL
50	推荐准确率	0.88	0.86
	推荐准确的数量	44	43
	推荐准确但不属于篮球的数目	9	12
100	推荐准确率	0.82	0.83
	推荐准确的数量	82	83
	推荐准确但不属于篮球的数目	19	24
150	推荐准确率	0.75	0.77
	推荐准确的数量	112	115
	推荐准确但不属于篮球的数目	30	38

结束语　微博作为一种通过关注机制分享简短实时信息的广播式的社交网络平台和典型的社会网络服务,每天产生着大量信息,使用户很难在海量的微博信息中及时找到自己感兴趣的内容,因此实时性、个性化是微博推荐的研究热点。本文提出了 LDA 主题模型结合 KL 散度的 RPMPS 推荐模型,并基于该模型构建了微博推荐系统。通过设置时间阈值确定用户的近期兴趣,并假设微博用户更倾向于发布与自身标签相符的微博,提高了对微博数据的处理效率,获得了更有价值的微博。由于使用 RPMPS 推荐模型不但可以通过文档-主题概率分布矩阵获得用户隐含的潜在兴趣,而且可以通过文档-词来对词频概率进行统计以获得用户信息与待推荐微博的直接相关性,因此结合两者的特点可以获得用户更满意的推荐结果。在真实数据集上的实验验证了该系统推荐微博的有效性。下一步将对模型进行进一步优化,考虑用户之间共同关注者的信息,进一步提升推荐的质量。

参 考 文 献

[1] CHEN J,LIU X J,LI B,et al. Personalized Microblogging Recommendation Based on Dynamic Interests and Social Networking of Users[J]. Acta Electronica Sinica,2017,45(4):898-905. (in Chinese)
陈杰,刘学军,李斌,等. 一种基于用户动态兴趣和社交网络的微博推荐方法[J]. 电子学报,2017,45(4):898-905.

[2] SHI L,TAO Y C,LI J Y,et al. Personalized and Real-time Recommendation Model for Microblogs[J]. Journal of Chinese Computer Systems,2016,37(9):1910-1914. (in Chinese)
石磊,陶永才,李俊艳,等. 个性化微博实时推荐模型研究[J]. 小型微型计算机系统,2016,37(9):1910-1914.

[3] YANG P,WANG D,ZHAO W B,et al. Research on Topic-Oriented Authoritative Information Retrieval Model in Microblog Site[J]. Journal of Frontiers of Computer Science & Technology,2013,7(12):1135-1145. (in Chinese)
杨平,王丹,赵文兵,等. 微博网站中面向主题的权威信息搜索技术研究[J]. 计算机科学与探索,2013,7(12):1135-1145.

[4] LÜ L,MEDO M,CHI H Y,et al. Recommender systems[J]. Physics Reports,2012,519(1):1-49.

[5] WEI C,HSU W,LEE M L. A unified framework for recommendations based on quaternary semantic analysis[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM,2011:1023-1032.

[6] SALAKHUTDINOV R R,MNIH A. Probabilistic Matrix Factorization[C]// Advances in Neural Information Processing Systems. 2007:1257-1264.

[7] GAO N,YANG M. Topic Model Embedded in Collaborative Filtering Recommendation Algorithm[J]. Computer Science,2016,43(3):57-61. (in Chinese)
高娜,杨明. 嵌入 LDA 主题模型的协同过滤推荐算法[J]. 计算机科学,2016,43(3):57-61.

[8] MOONEY R J,ROY L. Content-based book recommending using learning for text categorization[C]// ACM Conference on Digital Libraries. ACM,1999:195-204.

[9] LIN J,SUGIYAMA K,KAN M Y,et al. Addressing cold-start in app recommendation: latent user models constructed from twitter followers[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM,2013:283-292.

[10] CHEN J,LIU X J,LI B,et al. Personalized Microblogging Recommendation Based on Long-term and Short-term Interest of User[J]. Journal of Chinese Computer Systems,2016,37(5):952-956. (in Chinese)
陈杰,刘学军,李斌,等. 一种基于用户长短期兴趣的微博推荐方法[J]. 小型微型计算机系统,2016,37(5):952-956.

[11] BUSCH M,GADE K,LARSON B,et al. Earlybird: Real-Time Search at Twitter[C]// IEEE,International Conference on Data Engineering. IEEE Computer Society,2012:1360-1369.

(下转第 265 页)

[8] YANG Z M,QIAO L Y,PENG X Y. Research on datamining method for imbalanced dataset based on improved SMOTE[J]. Acta Electronica Sinica,2007,35(12):22-26. (in Chinese)
杨智明,乔立岩,彭喜元. 基于改进 SMOTE 的不平衡数据挖掘方法研究[J]. 电子学报,2007,35(12):22-26.

[9] WANG L Y. Research of boosting classification algorithm for imbalance data[D]. Harbin: Harbin Institute of Technology, 2013. (in Chinese)
王璐林. 面向不平衡样本的 Boosting 分类算法研究[D]. 哈尔滨:哈尔滨工业大学,2013.

[10] HU X S,WEN J P,ZHONG Y. Imbalanced data ensemble classification using dynamic balance sampling[J]. CAAI Transactions on Intelligent Systems,2016,11(2):257-263. (in Chinese)
胡小生,温菊屏,钟勇. 动态平衡采样的不平衡数据集成分类方法[J]. 智能系统学报,2016,11(2):257-263.

[11] GALAR M,FERNANDEZ A,BARRENECHEA E. Ordering-based pruning for improving the performance of ensembles of classifiers in the framework of imbalanced data sets[J]. Information Sciences,2016,354(C):178-196.

[12] KIM M J,KANG D K,HONG B K. Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction[J]. Expert Systems with Applications,2015,42(3):1074-1082.

[13] CHAWLA N V,BOWYER K W,HALLO L O. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research,2002,16(1):321-357.

[14] HAN H,WANG W Y,MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]// Proc. of International Conference on Intelligent Computing. 2005:878-887.

[15] DONG Y,WANG X. A new over-sampling approach: Random-SMOTE for learning from imbalanced data Sets[C]// Internation al Conference on Knowledge Science, Engineering and Management. 2011:343-352.

[16] WANG C X,PAN Z M,DONG L L,et al. Research on classification for imbalanced dataset based on improved SMOTE[J]. Computer Engineering and Applications,2013,49(2):184-187. (in Chinese)
王超学,潘正茂,董丽丽,等. 基于改进 SMOTE 的非平衡数据集分类研究[J]. 计算机工程与应用,2013,49(2):184-187.

[17] CRISTIANINI N,SHAWE T J. An introduction to support vector machines: and other kernel-based learning methods[M]. Cambridge University Press,2000.

[18] SCHOIKOPF B,MIKA S,BURGES C J C. Input space versus featurespace in kernei-based methods[J]. IEEE Transactions on Neural Networks,1999,10(5):1000-1017.

[19] TAO X M,ZHANG D M,HAO S Y. SVM classifier for unbalanced data based on spectrum cluster-based under-sampling approaches[J]. Control and Decision,2012,27(12):1761-1768. (in Chinese)
陶新民,张冬梅,郝思媛. 基于谱聚类欠取样的不均衡数据 SVM 算法[J]. 控制与决策,2012,27(12):1761-1768.

[20] ZENG Z Q,WU Q,LIAO B S. A classfication method for imbalance data set based on kernel SMOTE[J]. Acta Electronica Sinica,2009,37(11):2489-2495. (in Chinese)
曾志强,吴群,廖备水. 一种基于核 SMOTE 的非平衡数据集分类方法[J]. 电子学报,2009,37(11):2489-2495.

(上接第 259 页)

[12] OTSUKA E,CHIU D. Design and evaluation of a Twitter hashtag recommendation system[C]// International Database Engineering & Applications Symposium. ACM,2014:330-333.

[13] GAO M,JIN C Q,QIAN W N,et al. Real-time and personalized recommendation on microblogging systems[J]. Chinese Journal of Computers,2014,37(4):963-975. (in Chinese)
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.

[14] QIU H H,LIU Y,ZHANG Z J,et al. An Improved Collaborative Filtering Recommendation Algorithm for Microblog Based on Community Detection[C]// Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. IEEE,2014:876-879.

[15] CHEN X. A Hybrid Microblog Recommendation Model in Mobile Social Network[J]. Journal of Electronic Commerce in Organizations,2014,12(4):69-79.

[16] JIANG C. A Mircroblog Recommendation System Based on User Clustering and Semantic Dictionary[D]. Hangzhou: Zhejiang University,2013. (in Chinese)
蒋超. 基于用户聚类和语义词典的微博推荐系统[D]. 杭州:浙江大学,2013.

[17] CHEN L,JIANG C,WANG W. A Micro blog Recommendation System Based on User Clustering[C]// 2014 International Conference on Computer Science and Electronic Technology (ICCSET 2014). Atlantis Press,2015.

[18] XI Y,YANG J,TANG C H,et al. An Overlapping Semantic Community Detection Algorithm Based on Local Semantic Cluster[J]. Journal of Computer Research & Development,2015,52(7):1510-1521. (in Chinese)
辛宇,杨静,汤楚蘅,等. 基于局部语义聚类的语义重叠社区发现算法[J]. 计算机研究与发展,2015,52(7):1510-1521.

[19] BLEI D M,NG A Y,JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research,2003(3):993-1022.

[20] BERGROTH L,HAKONEN H,RAITA T. A survey of longest common subsequence algorithms[C]// International Symposium on String Processing and Information Retrieval, 2000 (Spire 2000). IEEE,2000:39-48.

[21] ZHANG J P,XIE J,YANG J,et al. A t-closeness privacy model based on sensitive attribute values semantics bucketization[J]. Journal of Computer Research & Development,2014,51(1):126-137. (in Chinese)
张健沛,谢静,杨静,等. 基于敏感属性值语义桶分组的 t-closeness 隐私模型[J]. 计算机研究与发展,2014,51(1):126-137.

[22] GAO C,MIAO D Q,ZHANG Z F,et al. A semi-supervised rough set model for classification based on active learning and co-training[J]. Pattern Recognition & Artificial Intelligence,2012,25(5):745-754. (in Chinese)
高灿,苗夺谦,张志飞,等. 主动协同半监督粗糙集分类模型[J]. 模式识别与人工智能,2012,25(5):745-754.