面向移动终端的微博信息推荐方法

宋双永 李秋丹

(中国科学院自动化研究所复杂系统与智能科学重点实验室 北京 100190)

摘 要 微型博客(简称"微博")以其简洁方便的交互方式,受到越来越多手机用户的喜爱。然而,微博数据量大、更新速度快以及手机屏幕小、登录网络服务速度较慢等原因,使得用户很难通过移动终端快速了解到近期内微博流行内容。提出一种基于相关主题模型(correlated topic model)的移动微博信息推荐方法,并基于此方法设计了一个可视化移动信息推荐系统。通过'用户-主题-词语'三维关联矩阵的建立,帮助用户快速了解最近一段时间内的热点主题,并查找与其感兴趣主题相关的其他用户作为备选好友,同时计算主题之间的关联关系,进行主题扩展。在微博代表性网站——Friendfeed数据集上进行的实验表明了该方法在移动微博信息推荐中的简洁性和有效性。

关键词 微型博客,相关主题模型,好友推荐,主题扩展,移动信息推荐

中图法分类号 TP391 **文献标识码** A

Micro-blogging Information Recommendation System for Mobile Client

SONG Shuang-yong LI Qiu-dan

(The Key Laboratory of Complex System and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract As a new type of online community, micro-blogging has gained more and more attention from mobile users. The most favorite person, as well as understanding interesting themes have become the main reason why users visit micro-blogging. In this paper, we proposed a correlated topic model-based approach for information recommendation in micro-blogging system. The approach can automatically find the relationship among users, topics and words, from which we can get the hot topics in the most recent period, the most influential users about each topic, and the incidence relation among those topics. Experiment results show that the method can provide mobile users with useful information related to their interest.

Keywords Micro-blogging, Correlated topic model, Friend recommendation, Topic expansion, Mobile information recommendation

1 引言

微型博客(简称"微博")是近年来快速兴起的一种Web 2.0的网络媒体形式,以其方便快捷的信息传播模式成为备受关注的传播工具。随着移动终端的普及应用以及移动互联网技术的高速发展,越来越多的用户选择用短信、彩信或手机客户端随时随地更新自己的微博,追随自己喜欢的用户[5]。但是,移动设备本身在显示与处理能力上的有限性,使得手机用户无法方便快速地通过浏览大量微博帖子来了解最近一段时间内的微博热点,这就要求面向移动终端的微博应用能够为用户提供一种更为简洁的信息浏览模式。

目前在微博信息推荐相关应用与研究中,好友推荐与主题扩展是非常受关注的研究内容。WhoShouldiFollow[6]是一个以推荐好友供用户追随的应用网站。通过用户已经开始追随的好友,查找与之具有相似个人信息的其他用户,作为备选好友推荐给用户。WeFollow[7]是另一个以推荐微博好友为目的的网络应用,与WhoShouldiFollow不同的是,该应用以用户搜索的关键词为出发点,寻找标签中包含此关键词的其他用户,按照相同tag标注次数以及追随者数量共同对相关人物进行排序,将所得结果推荐给用户。Weng等提出了一种按照不同主题对相应用户进行排序的方法[4],以用户之间的发帖内容相似性作为用户之间'关联',经过PageRank算法多次迭代,得到相关主题下的用户排序。Grosseck和Holotescu[12]提出一种基于tag对twitter内容进行文本分类的方法,以查找与给定主题相关的微博内容进行主题扩展。Sakaki[13]等人以地震为主题,搜集微博中的相关内容,将语义分析与微博数据特性相结合,以达到对事件实时动态地跟踪以及地震发展情况预测之目的。

本文提出一种新的基于相关主题模型(correlated topic model,CTM)的移动微博信息推荐方法,以满足手机用户在浏览微博信息时对信息形式简洁性和信息内容全面性的综合要求。本文方法运用CTM模型,从最新一段时间内微博帖

到稿日期:2010-12-22 返修日期:2011-02-23 本文受973国家重点基础研究发展计划(2007CB311007),国家自然科学基金(60703085)资助。

宋双永(1986—),男,博士生,主要研究方向为信息检索等,E-mail:shuangyong.song@ia.ac.cn;李秋丹(1976—),女,博士,副研究员,主要研究方向为信息检索、数据挖掘、移动电子商务等。

子数据里挖掘出其中隐含着的主题及其概率分布，对每个主题下相应的用户进行排序，并发现各个主题之间的相关性。与前面提到的好友推荐与主题推荐方法不同的是，基于相关主题模型得到的微博主题结果，能够反映相应时间段内微博帖子内容的整体概括信息，而不仅仅是某些比较热点的单一话题。用户能够通过此结果，快速了解到微博最近一段时间内帖子中包含的主要内容，根据自己的喜好选择相应内容加以了解，并追随相应的好友。同时，用户可以发现与其感兴趣主题相关联的其他信息，了解更多兴趣范围内的内容。

本文第 2 节介绍基于 CTM 模型提出的微博好友推荐方法；第 3 节描述实验数据、实验中参数的设定过程、实验结果及其分析；最后总结全文。

2 基于相关主题模型的微博信息推荐

近年来，LDA[2] 和 CTM[1] 等主题模型（Topic models）被广泛应用于文本分类和信息检索等领域。这些模型从文本中提取词与词之间的相关信息，并将一系列词语的分布概率称为主题（Topic）。本文提出的基于相关主题模型（Correlated Topic Model，CTM）的微博好友推荐及主题扩展方法，针对用户感兴趣的主题，为其推荐好友及主题相关内容，提高微博在用户信息服务方面的整体质量。

CTM 是一种用于从大量文本数据中检测隐含主题的非监督机器学习方法[1]。在 CTM 中，主题服从 Logistic 正态分布。Logistic 正态分布有两组参数，分别是均值向量和协方差矩阵。均值向量用以表示隐含主题的相对强弱；而协方差矩阵描述的是每对隐含主题之间的关联程度。因此，利用 CTM 不仅可以分析文本集合的隐含主题构成，而且可以考察隐含主题之间的联系[8]。

在微博中，用户发表的帖子内容能够反映其个人兴趣[4]，此内容可表示为 $U_i=\{w_{i1},w_{i2},\cdots,w_{iN}\}$，其中 w_{ij} 代表词语 w_j 在用户 U_i 发表帖子中出现的次数。在一段时间内所有用户发表的帖子信息中存在若干隐含主题，如图 1 所示。

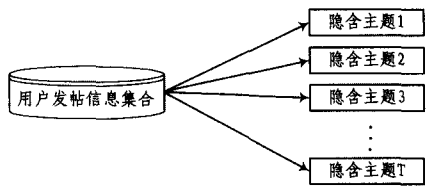

图 1　文档集合与隐含主题之间的包含关系

基于 CTM 模型，可通过‘用户-主题-词语’之间深层关联关系的获取，自动发现该时间段内的微博热点主题、各主题下用户的影响力排序以及主题之间的关联关系，从而为用户提供有效的信息服务。该模型为词层、用户层和用户发帖信息集合层（以下简称“信息集合层”）组成的有向概率图模型，如图 2 所示。信息集合包含 D 个用户以及 N 个不同的词。(μ,Σ,β) 是信息集合层的参数，其中 μ 和 Σ 用于描述信息集合中隐含主题间的相对强弱，隐含主题自身的概率分布用 β 表示。随机变量 η 是用户层参数，其分量代表目标用户帖子中每个隐含主题的权重。(Z,w) 是词层的参数，Z 表示目标用户帖子中的隐含主题在每个词上的份额，w 是目标用户的特征词向量。给定信息集合，CTM 模型将用户数据表示成主题 $t_{1:T}$ 与多元高斯参数 $\{\mu,\Sigma\}$ 的组合函数，并通过变化的 expectation-maximization（EM）方法来进行参数估计。需要注意的是，在 CTM 中，隐含主题抽取的数目 T 需要人工进行指定。

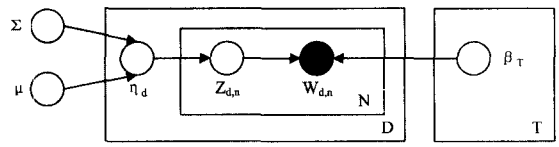

图 2　CTM 模型

基于上述各参数的确定，我们在图 3 中给出了整个算法的图示。其中，user-word 矩阵表示每个用户帖子中包含的词语及其概率，topic-word 矩阵表示每个隐含主题的内容，而 topic-user 矩阵表示每个主题下最为活跃的用户排序结果，topic-topic 矩阵表示主题之间的关联关系。通过隐含主题的抽取以及 topic-word 矩阵的建立，我们能够发现在这一段时间内微博用户所关心的主要内容。以此为基础，进一步建立 topic-user 关联矩阵，能够发现在每个隐含主题下的活跃用户，以进行基于主题的好友推荐，并通过 topic-topic 关联矩阵来实现主题扩展，帮助用户了解更多兴趣相关的内容。算法 1 所示为基于相关主题模型的微博信息推荐方法详细流程。

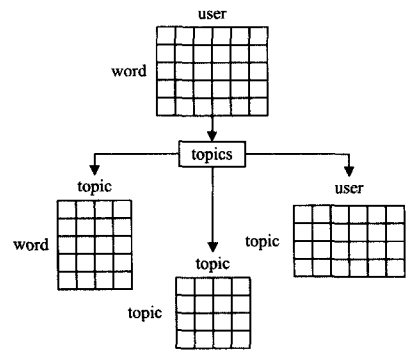

图 3　基于相关主题模型的微博信息推荐算法

算法 1　基于相关主题模型的微博信息推荐

输入：User-word 矩阵；待抽取的主题个数 T；EM 的收敛阈值 Ψ；迭代过程的最大循环次数 M_Ψ。

输出：Topic-word 矩阵；Topic-user 矩阵；Topic-topic 矩阵。

学习步骤如下：

for $n=1$ to M_Ψ

步骤 1　估计 CTM 模型的参数 μ，Σ 和 β：

$\vec{\mu}=\frac{1}{D}\sum_d\vec{\lambda}_d$，其中 $\vec{\lambda}_d$ 是指用户文档 d 在主题下的均值向量，D 为用户个数。

$\vec{\Sigma}=\frac{1}{D}\sum_d I\nu_d^2+(\vec{\lambda_d}-\vec{\mu})(\vec{\lambda_d}-\vec{\mu})^T$，其中 ν_d^2 表示用户文档 d 在所有主题下的方差。

$\vec{\beta}\propto\sum_d\varphi_{d,i}\vec{n_d}$，其中 $\varphi_{d,i}$ 表示用户文档 d 在主题 i 下的分布概率，$\vec{n_d}$ 表示用户文档 d 所对应的 N 维特征词向量。

步骤 2　利用隐含变量的迭代结果，计算目标文档的生成概率值 $p_{CTM}(x|\mu,\Sigma,\beta)$。

步骤 3　利用詹森不等式（Jensen’s inequality）[3] 计算得到结果的 EM 值。

步骤 4　若 EM$<\Psi$，停止迭代。

end for

步骤 5　通过生成的 Topic-word 矩阵，总结每个主题所包含的内容；从 Topic-user 矩阵中排序每个主题下的相关用户；以 Topic-topic 矩阵为基础，发现主题之间的关联关系。

3 实验结果及分析

3.1 数据描述

本文实验采用文献[10]中公开的 Friendfeed 网站的帖子数据集。该数据集包含了从 2009-09-01 到 2009-09-30 期间由 111284 位不同用户所发表的 1641531 条帖子内容。在实验中，假设用户查询时间为 2009-09-16。为保证查询结果的最新性，选取从 2009-09-13 到 2009-09-15 三天的帖子为最终实验数据，用以提取最新主题及相关信息。我们对数据做了以下预处理：1)删除发表帖子数量小于 3 的用户；2)对于出现次数小于 30 的词语，我们亦予以删除；3)将同一用户的所有帖子整理为一个文档；4)删除掉文档中所包含的网址内容(此类网址已经过 Friendfeed 系统自动编辑，不再包含其本来的真实域名)；5)删除掉文档中包含的标点符号；6)去除其中包含的非检索用词(stop words)[9]；7)统计词频，将每个文档表示成 CTM 模型中需要的词频向量形式。最终的数据包括 61934 个用户、29895 个词语。

3.2 参数设定

本节首先讨论抽取主题数目 T 的设定。在 CTM 算法中，T 的取值是按照操作者经验来进行主观选取。在 CTM 模型中，各个主题之间存在关联关系，使得 T 的取值对模型构造结果的影响不大[11]。根据主题区分度和用户浏览方便程度等方面综合考虑，本文将 T 选取为 50。另外，EM 的收敛阈值 Ψ 以及迭代过程的最大循环次数 M_Ψ，根据原始程序默认设定为 10^{-3} 和 1000[1]。

在查找关联主题的部分，需要对主题的'关联'与否设定一个阈值。通过主题之间的协方差矩阵，能够计算出主题之间的关联程度。在对该关联程度值进行归一化之后，需要设定一个 0 到 1 之间的数 λ 作为判断主题关联与否的阈值，λ 取值越小，表明两个主题之间的关联程度越大。表 1 中给出了 λ 取不同数值时每个主题平均的相关主题的个数。根据经验，通常一个主题有 2 到 3 个关联主题是比较合适的，因此在这里将 λ 设定为 0.35。

表 1 λ 变化得到主题关联性的不同结果

λ	Related topics	λ	Related topics
0.05	0.2	0.30	2.0
0.10	0.4	**0.35**	**2.7**
0.15	0.7	0.40	4.0
0.20	1.2	0.45	5.3
0.25	1.5	0.50	6.9

3.3 微博中的信息推荐

本文基于 J2ME 技术，设计了一种移动微博信息推荐系统，系统能够对微博主题信息进行规整，提供给移动用户一种简洁方便的微博信息浏览方式，以适应移动终端显示屏幕小、信息浏览操作不便等特点。图 4(a)中给出了微博信息推荐的系统界面，用户可以输入查询时间、每个主题显示的相关词语个数和每个主题相关的用户个数。

通过处理用户输入时间段内的帖子数据，从中抽取出隐含主题，能够帮助用户了解最新的微博动态。图 4(b)中给出了其中几个主题的抽取结果，这几个主题分别描述在 2009 年 8 月份最后 3 天内微博帖子中所包含的不同方面的内容。以 topic 1 为例，通过列出的几个关键词，可以看出该主题主要讨论在移动社交网站 Jaiku 上发布图片、视频等内容，而 topic 2 则描述了美国总统 Obama 关于卫生医疗等方面的相关新闻报道。topic 3、topic 4、topic 5 也描述了其他方面的主题内容。

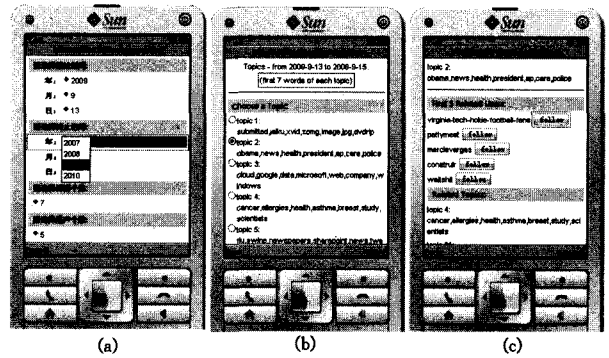

图 4 微博信息推荐系统显示

基于主题的微博好友推荐，旨在根据用户感兴趣的主题，寻找到近期发表过相关帖子的其他用户供其追随，方便实时地了解这些好友发表的最新内容。如图 4(c)所示，如果用户对 topic 2 感兴趣，想通过微博来关注此方面的信息，则可以通过追随 virginia-tech-hokie-football-fans，pattymeet，marcie-vargas 等用户来实现。

主题扩展部分，通过发现主题之间的关联，让用户了解更多与自己感兴趣主题相关的内容。如图 4(c)所示，topic 4 为系统检测出的与 topic 2 最为相关的一个主题。topic 4 主要描述内容为医疗健康等方面的研究学习，与 topic 2 内容有着很明显的联系。用户通过这两个主题之间的关联关系，能够对医疗健康方面的知识以及有关医疗健康的政策有更广泛的了解，知晓更多相关内容。

结束语 网络以一种潜移默化的方式在人们的日常生活中蔓延，通过网络信息交互来完成日常工作、学习或是进行沟通和娱乐，已经成为现代人生活中必不可缺的一部分。本文基于微型博客这种新型网络交互环境，提出了一种基于相关主题模型的好友推荐及主题扩展的方法，能对用户之间的联系与交流起到很大的帮助。实验结果证明，此方法能在该问题上取得较好效果。但是，在主题抽取的结果中有一些无意义的词语出现，可能是在抽取词语的过程中出现了词语分割的错误，或者是微博用户自身的书写错误。所以，设计基于标准词典的新型信息抽取方法，能够使得主题抽取的效果得到更大的提高。未来将进一步通过考虑用户之间的追随信息，增加基于主题的好友推荐的实验效果。

参 考 文 献

[1] Blei D M, Lafferty J D. Correlated topic models [M]// Weiss Y, Schölkopf B, Platt J, eds. Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2006

[2] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022

[3] Hansen F, Pedersen G K. Jensen's inequality for operator and Löwner's theorem [J]. Math. Anal, 1982, 258: 229-241

[4] Weng J, Lim E P, Jiang J, et al. TwitterRank: finding topic-sensitive influential twitterers[C]// WSDM 2010. 2010: 261-270

[5] Song S, Li Q, Zheng N. A spatio-temporal framework for related topic search in micro-blogging[C]// International Conference on Active Media Technology. Toronto, Canada, August 2010: 28-30

(下转第 166 页)

$$=\rho_1(1-\rho_2)v_2+(1-\rho_1)\rho_2 v_1+\rho_1\rho_2 v_3-\rho_1\rho_2 u=0 \quad (3)$$

求解由式(1)、式(2)、式(3)构成的方程组，可得该博弈的混合战略纳什均衡是

$$\rho_1^*=\frac{2v_1}{u+3v_1-v_2-v_3},\rho_2^*=\frac{2v_2}{u+3v_2-v_1-v_3},$$
$$\rho_3^*=\frac{2v_3}{u+3v_3-v_1-v_2} \quad (4)$$

即在隐私保护分布式数据挖掘中假设参与者是准诚信攻击的前提下，参与者 P_1 以 $\rho_1^*=\frac{2v_1}{u+3v_1-v_2-v_3}$ 的概率选择共谋，参与者 P_2 以 $\rho_2^*=\frac{2v_2}{u+3v_2-v_1-v_3}$ 的概率选择共谋，参与者 P_3 以 $\rho_3^*=\frac{2v_3}{u+3v_3-v_1-v_2}$ 的概率选择共谋。

根据海萨尼(Harsanyi)[6]定理，该混合战略均衡等价于不完全信息下的纯战略纳什均衡，即在不完全信息下，所有参与者均为准诚信攻击的前提下，上述均衡式(4)给出了每个参与者选择共谋行为的概率。

4.2.2 讨论

根据上述分析可知，对于多方参与的隐私保护分布式数据挖掘博弈，如果假设所有的数据挖掘参与者都是准诚信攻击的，基于收益的最大化，参与者采取非共谋行为并不是最优纳什均衡策略。特别当参与者的隐私信息价值大时，如 v_1，v_2，$v_3\geqslant u$，共谋有可能得到更高的收益，所有参与者的纳什均衡策略是共谋，而使得合作无法进行下去。文献[9,10]均以安全求和为例，对隐私保护数据挖掘中参与者的共谋行为进行了博弈分析，指出非共谋行为并不是最优纳什均衡策略。我们的结论与文献[9,10]是一致的，而且本文给出了该博弈分析的混合战略均衡，其等价于不完全信息下的纯战略纳什均衡，更符合实际情况。因为在实际中，完全信息只是一个理想的情形，现实中所有参与者面对的更多的是采取混合策略的合作者。

结束语 隐私保护的分布式数据挖掘问题是数据挖掘领域的一个研究热点。本文基于收益最大化，在完全信息静态博弈下研究了数据挖掘中参与者的策略决策问题，得出了如下结论：数据挖掘在满足一定的条件下(如当 $0\leqslant\theta_i\leqslant u$，$(1\leqslant i\leqslant n)$时)，参与者的准诚信攻击策略是一个帕累托最优的纳什均衡策略；当 $\theta_i\geqslant u\geqslant 0$，$(1\leqslant i\leqslant n)$时，博弈纳什均衡中所有参与者的最优策略均为恶意攻击策略；进一步，在准诚信攻击的假设下，参与者的非共谋策略并不是一个纳什均衡策略，而且给出了该博弈的混合战略纳什均衡，同时对上述问题进行了案例分析，从而对隐私保护的分布式数据挖掘中参与者的决策具有一定的理论和指导意义。

需要指出的是，本文只是在完全信息静态博弈下对隐私保护的分布式数据挖掘进行了研究。而现实的合作数据挖掘中，信息往往是不完全的，完全信息只是一个理想的情形，而且数据挖掘也具有动态的特性。本文主要是基于静态(Static)共谋的策略分析，而自适应(Adaptive)共谋也许更符合实际情况，这些都是今后需进一步研究的。

参 考 文 献

[1] Vaidya J S, Clifton C. Privacy preserving association rule mining in vertically partitioned data[C]// Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002:639-644

[2] Lindell Y, Pinkas B. Privacy preserving data mining [J]. Journal of Cryptology, 2002, 15(3):177-206

[3] Agrawal R, Srikant R. Privacy-preserving data mining[C]// Proceedings of the SIGMOD Conference on Management of Data. ACM Press, 2000:439-450

[4] Lindell Y, Pinkas B. Secure Multiparty Computation for Privacy-preserving Data Mining [J]. Journal of Privacy and Confidentiality, 2009, 1(1):59-98

[5] Clifton C, Kantarcioglu M, Vaidya J, et al. Tools for Privacy Preserving Distributed Data Mining [J]. ACM SIGKDD Explorations, 2002, 4(2):28-34

[6] 张维迎. 博弈论与信息经济学[M]. 上海：上海人民出版社，2004

[7] Kleinberg J, Papadimitriou C, Raghavan P. A microeconomic view of data mining [J]. Data Mining and Knowledge Discovery, 1998, 2(4):311-324

[8] Abraham I, Dolev D, Gonen R, et al. Distributed computing meets game theory: Robust mechanisms for rational secret sharing and multiparty computation[C]// Proceedings of the Twenty-fifth Annual ACM Symposium on Principles of Distributed Computing. New York, USA: ACM Press, 2006:53-62

[9] Kargupta H, Das K, Liu K. Multi-party, privacy-preserving distributed data mining using a game theoretic framework [J]. PKDD, 2007, 4702:523-531

[10] 张国荣，印鉴. 基于博弈论的安全多方求和方法[J]. 计算机应用研究，2009, 26(4):1497-1499

[11] Kantarcioglu M, Jiang Wei. Incentive Compatible Privacy-preserving Data Analysis[R]. UTDCS-29-08. UT Dallas Computer Science Department, 2008

(上接第 139 页)

[6] http://whoshouldifollow.com/

[7] http://wefollow.com/

[8] 李文波，孙乐，张大鲵. 基于 Labeled-LDA 模型的文本分类新算法[J]. 计算机学报，2008，31：620-627

[9] http://www.ranks.nl/resources/stopwords.html

[10] Celli F, Lascio F M L D, Magnani M, et al. Social network data and practices: the case of friendfeed[C]// International Conference on Social Computing, Behavioral Modeling and Prediction. Lecture Notes in Computer Science. Springer, Berlin, 2010

[11] 曹娟，张勇东，李锦涛，等. 一种基于密度的自适应最优 LDA 模型选择方法[J]. 计算机学报，2008，31(10)：1780-1787

[12] Grosseck G, Holotescu C. Can we use twitter for educational activities? [C]// the 4th International Sscientific Conference on elearning and software for education. Bucharest, Rumania, April 2008:17-18

[13] Sakaki T, Okazaki M, Matsuo Y. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors[C]// the 19th International World Wide Web Conference. Raleigh, North Carolina, USA, April 26-30