

## 基于话题标签的微博主题挖掘

李 敬, 印 鉴, 刘少鹏, 潘雅丽

(中山大学信息科学与技术学院计算机科学系, 广州 510006)

**摘 要:** 随着互联网的发展, 微博已成为人们获取信息的主要平台, 为从海量微博中挖掘出有价值的主题信息, 结合微博中的会话、转发和话题标签, 将微博划分为用户兴趣、用户互动和话题微博 3 类, 提出基于作者主题模型 (ATM) 的话题标签主题模型 HC-ATM, 使用 Gibbs 抽样法对模型进行推导, 获取微博主题结构。在 Twitter 数据集上的实验结果表明, 与 ATM 模型和基于潜在狄利克雷分布的微博生成模型相比, HC-ATM 模型的主题困惑度更小、差异度更大, 并且能有效挖掘出不同微博类型的主题分布。

**关键词:** 主题挖掘; 微博; 社交网络; 话题标签主题模型; 作者主题模型

**中文引用格式:** 李 敬, 印 鉴, 刘少鹏, 等. 基于话题标签的微博主题挖掘[J]. 计算机工程, 2015, 41(4): 30-35.

**英文引用格式:** Li Jing, Yin Jian, Liu Shaopeng, et al. Microblog Topic Mining Based on Hashtag [J]. Computer Engineering, 2015, 41(4): 30-35.

## Microblog Topic Mining Based on Hashtag

LI Jing, YIN Jian, LIU Shaopeng, PAN Yali

(Department of Computer Science, School of Information Science and Technology,  
Sun Yat-sen University, Guangzhou 510006, China)

**【Abstract】** With the development of the Internet, microblog has become a major platform for people to obtain the information. In order to mine useful topic from microblog, based on the futures of microblog that having conversation tags, retweet tags and hashtags, this paper divides microblog into three kinds. They are microblogs about users' interest, users interaction and hashtag-related. It designs a novel hashtag topic model named Hashtag Conversation Author Topic Model (HC-ATM) based on Author Topic Model (ATM), and uses Gibbs sampling implementation for inference of this model. Experiments on Twitter dataset show that HC-ATM outperforms the ATM and MicroBlog Latent Dirichlet Allocation (MB-LDA) in terms of both perplexity and KL-divergence. Besides, HC-ATM can mine topic distribution of different kinds of microblog effectively.

**【Key words】** topic mining; microblog; social network; hashtag topic model; Author Topic Model (ATM)

**DOI:** 10.3969/j.issn.1000-3428.2015.04.006

### 1 概述

微博是一种新兴的社交网络服务, 用户可以通过登录微博客户端来发布短文本消息, 同时可以附带链接、图片和视频等多媒体资源, 与好友进行信息的及时分享。海量微博数据蕴含着丰富的信息, 从中挖掘出有效的微博主题, 对信息提取与用户分析具有重要意义。

微博与传统文本(新闻、论文等)不同, 作为短文

本消息, 微博中词项的共现信息匮乏, 并且没有特定的语法结构, 因此简单地套用传统文本主题挖掘的方法很难挖掘有用的微博主题<sup>[1]</sup>。针对微博的稀疏性和非结构化特点, 近年来学者在研究潜在狄利克雷分布 (Latent Dirichlet Allocation, LDA)<sup>[2]</sup> 的基础上引入联系人的关系<sup>[3]</sup>、新闻链接<sup>[4]</sup> 和时间信息<sup>[5]</sup> 等外部信息来挖掘微博主题, 并取得了较好成效。研究表明, 微博本身隐含的用户信息与社交网络信息可以促进微博主题挖掘。然而目前研究并没有明确地从用户兴趣、用户互动和话题微博 3 种微

**基金项目:** 国家自然科学基金资助项目 (61033010, 61272065); 广东省自然科学基金资助项目 (S2011020001182, S2012010009311); 广东省科技计划基金资助项目 (2011B040200007, 2012A010701013)。

**作者简介:** 李 敬 (1988 -), 男, 硕士研究生, 主研方向: 文本挖掘, 机器学习; 印 鉴 (通讯作者), 教授、博士; 刘少鹏, 博士; 潘雅丽, 硕士研究生。

**收稿日期:** 2014-04-29    **修回日期:** 2014-06-09    **E-mail:** lijing337@mail2.sysu.edu.cn

博类型上将微博主题进行划分, 无法同时挖掘出不同微博类型下的主题分布。

本文结合用户兴趣、会话标签 (@)、转发标签 (RT @) 和话题标签 (含有 #hashtag), 对不同微博类型进行主题划分分析, 提出一个新的在微信平台下适用的话题标签主题模型 HC-ATM (Hash Tag Conversation Author Topic Model), 并给出与 HC-ATM 相对应的 Gibbs 抽样推导结果。

## 2 相关工作

主题模型是近年来在文本挖掘领域最受关注的方法之一, 是一种概率生成模型, 它常被用于挖掘大规模文档集的潜在主题。主题模型挖掘的主题与人类对文本的理解较接近, 体现出文本间的语义关系。

### 2.1 主题模型

LDA<sup>[2]</sup> 作为主题模型的典型代表, 避免了 pLSI<sup>[6]</sup> 中由于参数过多而导致的过拟合问题, 同时还可以对训练集之外的文档进行概率估计。LDA 将每篇文档看做是多个主题的概率分布, 而其中的每个主题则是多个单词的概率分布。在 LDA 中, 一篇文档内的单词是可交换的, 文档与文档之间也是条件独立同分布的。在给定主题个数  $K$  的情况下, 先生成每篇文档中单词的主题, 然后再由主题分布生成该单词。根据 LDA 对文档生成过程的假设, 可以使用参数估计方法反向推导概率模型, 求得每个主题下的词项分布和每篇文档下的主题分布, 进而揭示文档主题结构。常用的推导方法有变分贝叶斯<sup>[2]</sup>、Gibbs 抽样<sup>[7]</sup>、期望值传播<sup>[8]</sup>等。

随着对主题模型的深入研究, 衍生出了适用于各类具体应用的主题模型。文献[9]提出用于挖掘主题之间相关性的相关主题模型 (Correlated Topic Model, CTM), 文献[10]通过引入文档的作者信息, 提出作者主题模型 (Author Topic Model, ATM), 此外还有结合时间信息对文档进行建模分析的主题模型 ToT (Topics over Time)<sup>[11]</sup> 和 DTM (Dynamic Topic Models)<sup>[12]</sup> 等。

### 2.2 微博主题挖掘

相比于传统文本, 微博缺乏词项共现信息, 数据十分稀疏, 直接使用传统主题模型难以挖掘出有用的主题信息。研究者从不同角度出发, 提出了适用于不同应用场景的微博主题模型。Labeled-LDA<sup>[13]</sup> 是一个监督主题模型, 它将微博内容映射到 substance, style, status 和 social 4 个维度, 用以划分不同主题下的用户和微博。文献[14]结合用户兴趣、时间信息和背景信息, 构建出与时间片序列相关的主题分布, 将微博主题挖掘的结果用来分析、探测爆发性的新闻话题。微博 LDA (Microblog LDA, MB-

LDA)<sup>[3]</sup> 利用微博会话引入联系人关联关系, 用于改善微博的主题挖掘效果。文献[13]通过最小化消息序列主题的预测误差来求解主题转移矩阵, 用于预测用户未来微博的主题分布。文献[15]把每对词项的共现模式融入到文本生成过程中, 得到适用于挖掘问答系统和微博的主题模型。文献[16]利用微博主题对用户进行建模, 根据用户特征表示与微博特征的相似性程度对用户进行微博推荐。

由于微博的特殊性, 当前模型倾向于关注用户内在兴趣, 并没有充分利用微博中的会话、转发和话题标签, 无法同时得到用户兴趣、用户互动和话题微博下的主题分布。文献[3]考虑了会话、转发标签, 但是没有考虑话题标签, 无法挖掘话题标签下的主题分布, 也无法发现描述相同话题的话题标签。文献[13]主要是偏向于利用话题标签来对微博进行主题分析, 无法获取用户之间互动的主题分布。

## 3 微博主题挖掘模型设计与实现

### 3.1 作者主题模型

作者主题模型<sup>[11]</sup> 认为每篇文章中的单词是由该篇文章的所有作者合力完成, 其中每个作者都有自己的研究领域, 因此不同作者拥有不同的主题分布。作者主题模型如图 1 所示。

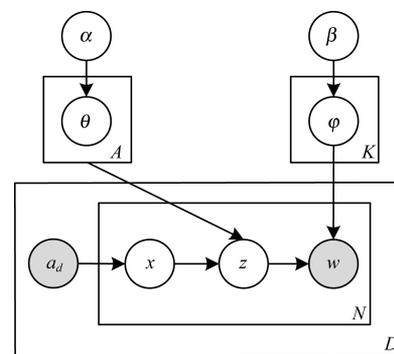


图 1 作者主题模型

在图 1 中,  $\alpha$  和  $\beta$  为 Dirichlet 超参数; 作者集合  $a_d$  和文档单词  $w$  为可观察的变量。在 ATM 的文档生成过程中, 首先生成每个主题  $z$  下的词项分布  $\varphi_z$ , 然后对于文档中的每一个单词  $w_i$ : 从文档合作作者集合  $a_d$  中以均匀分布选取出一个作者  $x$ , 然后再从该作者的主题概率分布  $\theta_x$  中抽取一个主题  $z_i$ , 接着由该主题下的词项分布  $\varphi_{z_i}$  生成一个单词  $w_i$ , 直至生成整篇文档。ATM 充分考虑了作者信息, 挖掘出的作者主题分布在一定程度上代表了该作者的兴趣爱好。

### 3.2 话题标签主题模型

微博主题模型 HC-ATM 以 ATM 作为基模型, 结合会话、转发标签和话题标签, 对微博生成过程进

行统一建模分析,构建出不同微博类别下的主题分布。

微博具有不同于传统文本的特性,其隐含的社交信息对微博主题发现具有一定的促进作用<sup>[3]</sup>。如含有会话标签(@)和含有转发标签(RT@)的微博:“@mashableHow much does the app charge?”与“RT@mashable the onion launches a new iphone app”。如果将2条微博分别作为独立的微博来分析微博主题,很难得出前一条微博中的“app”就是“a new iphone app”,而话题标签和转发标签所对应的用户 mashable 正是这2条微博的连接纽带。会话标签和转发标签揭示了微博之间的语义联系,能帮助主题模型更好地发现用户互动中的主题分布。

话题标签(#hashtag)是用户在发微博时给微博添加的自定义标记,表示该条微博的主题。话题标签具有一定的混淆性,不同标签可能指代相同的话题。不同用户给主题相同的微博添加的话题标签可以不同,因此提取出某一个无明显语义特征的话题标签(如话题“#tcot”)下的词项分布有助于发现具有相同语义信息的话题标签,进而可以对微博进行主题分析与标签推荐。

HC-ATM 将微博分为3类:用户兴趣微博,用户互动微博和含有话题标签的微博。HC-ATM 模型如图2所示,每条微博  $d$  中每个单词  $w_i$  的生成概率为:

$$p_d(w_i) = \lambda_d p(w_i | \theta_B) + (1 - \lambda_d) p(w_i | \varphi_{z_i}) p(z_i | \theta_d) \quad (1)$$

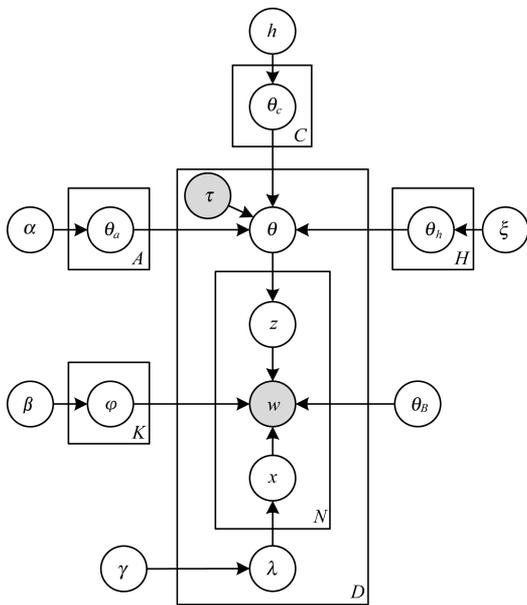


图2 HC-ATM 模型

在 HC-ATM 中,将词频作为模型的背景分布  $\theta_B$ ,用于平滑模型。用以  $\lambda_d$  为参数的伯努利分布来确定微博中每个单词是否由背景分布生成。当  $x_i =$

1 时,则微博中的单词  $w_i$  由背景分布  $\theta_B$  生成;否则首先由微博主题分布  $\theta_d$  抽取出一个主题  $z_i$ ,然后再由主题  $z_i$  下的词项分布  $\varphi_{z_i}$  生成该单词  $w_i$ 。模型生成过程如下:

(1) 对每个主题抽样  $\varphi_k \sim \text{Dir}(\beta), k \in [1, K]$ ;

(2) 对每条微博  $d \in [1, D]$ :

1) 确定微博类型  $\tau_d$ ;

2) 确定微博主题分布  $\theta_d$ ;

①若  $\tau_d = 0$ ,则微博主题分布为  $\theta_d = \theta_h | \theta_h \sim \text{Dir}(\xi)$ ;

②若  $\tau_d = 1$ ,则微博主题分布为  $\theta_d = \theta_c | \theta_c \sim \text{Dir}(\eta)$ ;

③否则微博主题分布为  $\theta_d = \theta_a | \theta_a \sim \text{Dir}(\alpha)$ ;

3) 抽样  $\lambda_d \sim \text{Beta}(\gamma)$ ;

4) 对于每个词  $w_i \in [1, N_d]$ :

①抽样  $x_i \sim \text{Bern}(\lambda_d)$ ;

②若  $x_i = 0$ ,则从  $\theta_d$  中抽取出一个隐含主题  $z_i \sim \text{Multi}(\theta_d)$ ,生成单词  $w_i \sim \text{Multi}(\varphi_{z_i})$ ;

③若  $x_i = 1$ ,则生成单词  $w_i \sim \text{Multi}(\theta_B)$ 。

在 HC-ATM 的微博生成过程中,先生成每个主题  $z$  下的词项分布  $\varphi_z$ ;接着判断该条微博是否含有话题标签#hashtag,若含有话题标签,即  $\tau_d = 0$  时,该条微博的主题分布  $\theta_d$  由话题标签的主题分布  $\theta_h$  决定;否则判断该条微博是否含有会话标签(@)或转发标签(RT@),若含有会话转发标签,即  $\tau_d = 1$ ,该条微博的主题分布  $\theta_d$  由用户互动主题分布  $\theta_c$  决定;若不含有任何标签,即  $\tau_d = 2$ ,则该条微博的主题分布  $\theta_d$  由用户兴趣主题分布  $\theta_a$  决定。然后从  $\theta_d$  中抽取单词的主题  $z_i$ ,最后再从  $\varphi_{z_i}$  生成单词  $w_i$ 。

### 3.3 模型推导

Gibbs 抽样是马尔科夫链蒙特卡罗方法(Markov Chain Monte Carlo, MCMC)的特例,每次迭代只对联合分布中的一个维度进行抽样,而其他维度保持不变。Gibbs 抽样常被用于概率模型的参数估计。HC-ATM 的 Gibbs 抽样后验公式具体如下:

$$p(x_i = 0, \tau_d = 0, z_i = k | z_{-i}, w) \propto \frac{n_{0,-i} + \gamma}{n_{-i} + 2\gamma} \cdot \frac{n_{k,v,-i} + \beta}{n_{k,-i} + V\beta} \cdot \frac{n_{h,k,-i} + \xi}{n_{h,-i} + K\xi} \quad (2)$$

其中,  $V$  是词项个数;  $K$  是主题个数;  $z_{-i}$  表示除了单词  $i$  外所有单词的主题下标;  $w$  表示所有的单词;  $n_{0,-i}$  表示除了单词  $i$  外,属于主题分布  $\theta_d$  的单词个数;  $n_{-i}$  表示除了单词  $i$  外的单词个数;假设  $w_i = v$ ,则  $n_{k,v,-i}$  表示除了单词  $i$  外,词项  $v$  被分配给主题  $k$  的次数;  $n_{k,-i}$  表示除单词  $i$  外被分配给主题  $k$  的词总数;  $n_{h,k,-i}$  表示除单词  $i$  外,话题标签  $h$  中出现主题  $k$  的次数;  $n_{h,-i}$  表示除单词  $i$  外,话题标签  $h$  中出现的所有主题总和。

$$p(x_i = 0, \tau_d = 1, z_i = k | z_{-i}, w) \propto \frac{n_{0,-i} + \gamma}{n_{-i} + 2\gamma} \cdot \frac{n_{k,v,-i} + \beta}{n_{k,-i} + V\beta} \cdot \frac{n_{c,k,-i} + \eta}{n_{c,-i} + K\eta} \quad (3)$$

其中,  $n_{c,k,-i}$  表示除单词  $i$  外, 会话转发标签  $c$  中出现主题  $k$  的次数;  $n_{c,-i}$  表示除单词  $i$  外, 会话转发标签  $c$  中出现的所有主题总和。式(4)中的  $n_{a,-i}$  同理。

$$p(x_i = 0, \tau_d = 2, z_i = k | z_{-i}, w) \propto \frac{n_{0,-i} + \gamma}{n_{-i} + 2\gamma} \cdot \frac{n_{k,v,-i} + \beta}{n_{k,-i} + V\beta} \cdot \frac{n_{a,k,-i} + \alpha}{n_{a,-i} + K\alpha} \quad (4)$$

$$p(x_i = 1) \propto \frac{n_{1,-i} + \gamma}{n_{-i} + 2\gamma} \cdot p(w_i | \theta_B) \quad (5)$$

其中,  $n_{1,-i}$  表示属于背景分布中的单词个数。Gibbs 抽样迭代直至收敛后, 使用以下公式对  $\theta_h, \theta_c, \theta_a$  和  $\varphi_k$  进行估计:

$$\theta_{h,k} = \frac{n_{h,k} + \xi}{n_h + K\xi} \quad (6)$$

$$\theta_{c,k} = \frac{n_{c,k} + \eta}{n_c + K\eta} \quad (7)$$

$$\theta_{a,k} = \frac{n_{a,k} + \alpha}{n_a + K\alpha} \quad (8)$$

$$\varphi_{k,v} = \frac{n_{k,v} + \beta}{n_k + V\beta} \quad (9)$$

其中,  $n_{h,k}, n_{c,k}$  和  $n_{a,k}$  分别表示话题标签微博、会话转发微博和用户兴趣微博中出现主题  $k$  的次数;  $\varphi_{k,v}$  表示词项  $v$  被分配给主题  $k$  的次数;  $\theta_h$  代表话题标签微博下的主题分布;  $\theta_c$  代表会话转发微博下的主题分布;  $\theta_a$  代表用户兴趣微博下的主题分布。

## 4 实验结果与分析

### 4.1 数据集与数据预处理

本文使用2009年9月-2010年1月的Twitter数据集<sup>[17-18]</sup>, 含有3 845 624条用户微博。微博消息含有大量没有实际意义的停用词, 因此需要对微博数据集进行预处理, 先使用已有的停用词表对微博数据集进行去除停用词处理, 接着采用Snowball算法对单词进行词干提取, 将不同时态的单词归为统一的表示形式, 随后去除低频单词和含有单词数较少的微博。经过预处理后, 得到约  $1 \times 10^5$  条微博作为实验数据集, 数据集描述如表1所示。

表1 实验数据集

参数	参数值
微博数( $D$ )	102 531
词项数( $V$ )	41 051
用户数( $A$ )	4 294
会话转发标签数( $C$ )	13 388
话题标签数( $H$ )	15 510

### 4.2 实验环境与参数设置

实验环境为 Windows 7 操作系统, Intel Core 3.2 GHz 处理器, 内存容量为 8 GB。实验选取 ATM 和 MB-LDA 作为比较实验, 3 个模型的主题数  $K$  设为 100, 超参数  $\alpha$  和  $\alpha_c$  设为 0.5,  $\beta$  设为 0.01。HC-ATM 中的超参数  $\eta, \xi$  和  $\gamma$  都设置为 0.5。

### 4.3 主题困惑度

困惑度 (Perplexity) 指标常被用于度量主题模型的性能, 表示模型预测数据时的不确定性。困惑度指标越小, 模型性能就越好。困惑度计算公式如下:

$$Perplexity(D) = \exp \left( - \frac{\sum_d \sum_i^{N_d} \ln p(w_{d,i})}{\sum_d N_d} \right) \quad (10)$$

其中,  $N_d$  表示微博  $d$  中的词数;  $w_{d,i}$  表示微博  $d$  中的第  $i$  个单词。ATM, MB-LDA 和 HC-ATM 的困惑度比较如图 3 所示。可以看出, 当迭代次数大于 300 时, 3 个模型的困惑度都趋于平稳状态, 并且 HC-ATM 具有比 ATM 和 MB-LDA 更小的困惑度, 证明了 HC-ATM 的可靠性。

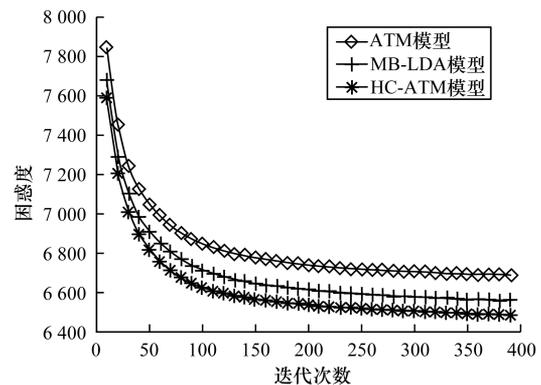


图3 模型的主题困惑度比较

此外, 主题模型的  $K$  值越大, 模型越容易识别不同含义的潜在主题, 具有更小的混惑度, 但  $K$  值也不宜设置过大, 否则会使模型挖掘出大量的垃圾主题, 不利于在整体上把握微博数据的主要内容。

### 4.4 主题差异性

主题差异性是指度量模型提取出的主题间的差异程度, 抽取出的主题两两之间的差异性越大, 说明主题越具有代表性。KL 距离常用于度量 2 个概率分布间的差异程度, 实验中使用 KL 距离来计算主题差异性。2 个主题的差异性越大, KL 距离越大; 反之, KL 距离越小。在极端情况下, 主题间的 KL 距离为 0, 表示 2 个主题完全一致。KL 距离计算公式为:

$$KL(\varphi_1, \varphi_2) = \sum_{w_{d,i}} p(w_{d,i} | \varphi_1) \log \frac{p(w_{d,i} | \varphi_1)}{p(w_{d,i} | \varphi_2)} \quad (11)$$

其中,  $w_{d,i}$  表示微博  $d$  中的第  $i$  个单词;  $\varphi_1$  和  $\varphi_2$  表示不同的主题。ATM, MB-LDA 和 HC-ATM 3 个模型中两两主题之间的 KL 平均距离比较如表 2 所示。

可以看出,HC-ATM 的 KL 值相对较大,具有更好的主题差异性,说明了对微博进行分类建模能够帮助主题模型发现更具代表性的主题。

表 2 模型 KL 距离比较

模型	KL 距离
ATM 模型	3.734
MB-LDA 模型	3.718
HC-ATM 模型	3.760

#### 4.5 主题有效性

微博主题挖掘的目标是挖掘出有用的主题,主题是否有效,与主题下的词项分布有关。

表 3 主题有效性比较

主题 1			主题 2			主题 3		
ATM	MB-LDA	HC-ATM	ATM	MB-LDA	HC-ATM	ATM	MB-LDA	HC-ATM
show	film	film	health	health	health	music	music	music
film	mark	movie	care	care	care	song	song	song
movie	movie	screen	bill	house	insurance	album	album	album
screen	sun	hollywood	vote	reform	bill	band	record	mp3
hollywood	hollywood	trailer	house	bill	house	listen	play	itunes

#### 4.6 多类型主题分析

由 HC-ATM 分析出不同的主题类型如表 4 所示。通过分析话题标签下的主题可以发现微博中较热门的话题,并且可以得到相似主题的话题标签。不同用户的知识背景不同,对相同话题的标签描述页不同。如#tcot 和#hcr 都是属于医疗改革的话题标签,但仅从 2 个标签来看,无法确定这 2 个标签所

人工判断词项与描述主题的相关程度,是评价主题模型有效性的方法。限于篇幅,表 3 只给出 3 个模型中都能找到的 3 个相同主题:主题 1(电影相关),主题 2(医疗法案相关),主题 3(音乐相关),每个主题给出前 5 个出现概率最大的单词。可以看出,3 个模型都能挖掘出具有一定代表性的主题,但 HC-ATM 具有更加接近人类对主题理解的表示。在主题 1 中,ATM 得到概率最大的词项为 show,如果光从该词出发,并不能很好地判断主题 1 是与电影相关的主题,而 MB-LDA 中出现的 mark 同样也不能显著地表示出主题 1 与电影主题的关系。

描述的主题是相似的。

分析会话转发标签下的主题,可以得到用户之间互动的主题。如用户 jennaldewan 经常参与一些与电影相关的讨论,当有多条与用户 jennaldewan 进行互动的微博出现时,可以根据分析结果将与电影相关的微博评论排在靠前的位置,提高阅读评论的效率。

表 4 多类型主题分析

主题类型	所属主题	相关微博
#tcot (话题标签)	主题 2	Maybe they should name the new public health insurance company Fannie Health? #healthcare #obama #tcot #hc09
#hcr (话题标签)	主题 2	The Senate has begun working on their health care reform bill. #healthcare #hcr
@jennaldewan (用户互动)	主题 1	@jennaldewan: Get latest release data info, filming news, & photos 4m set of Jenna's movie 'Legends of Hell's Gate'. RT @jennaldewan: the next movie I'm going to do is called The Legend Hells gate. It is a western period film.
用户 ID:19671932 (用户兴趣)	主题 3	It's Monday.....new week.....new music.....speaking of.....what's your favorite NEW song out?

由用户兴趣的主题分布,可以挖掘出用户的兴趣爱好,如用户 ID:19 671 932 的兴趣爱好为音乐,对该用户进行与音乐相关的微博推荐,并结合用户互动主题分布,扩展用户兴趣爱好,提供更为广泛的微博推荐。

## 5 结束语

本文对用户兴趣、用户互动和话题标签 3 种类

型微博进行统一建模,提出一个新的主题模型 HC-ATM。利用 HC-ATM 可以同时挖掘出不同微博类型下的主题分布,并能获得较好的主题质量。

由于本文主题个数确定,然而在实际应用中找到合适的主题个数需要一定人工经验,因此今后将对主题个数的自动获取进行研究,提高主题挖掘效率。

## 参考文献

- [ 1 ] Yan Xiaohui, Guo Jiafeng, Lan Yanyan, et al. A Bitern Topic Model for Short Texts [ C ] // Proceedings of the 22nd International Conference Companion on World Wide Web. Rio de Janeiro, Brazil: IW3C2 Press, 2013: 1445-1456.
- [ 2 ] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [ J ]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.
- [ 3 ] 张晨逸, 孙建伶, 丁轶群. 基于 MB-LDA 模型的微博主题挖掘 [ J ]. 计算机研究与发展, 2011, 48(10): 1795-1802.
- [ 4 ] Zhao Xin, Jiang Jing, He Jing, et al. Comparing Twitter and Traditional Media Using Topic Models [ C ] // Proceedings of the 33rd European Conference on IR Research. Berlin, Germany: Springer-Verlag, 2011: 338-349.
- [ 5 ] Hong Liangjie, Dom B, Gurumurthy S, et al. A Time-dependent Topic Model for Multiple Text Streams [ C ] // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2011: 832-840.
- [ 6 ] Deerwester S C, Dumais S T, Landauer T K, et al. Indexing by Latent Semantic Analysis [ J ]. Journal of American Society for Information Science, 1990, 41(6): 391-407.
- [ 7 ] Griffiths T L, Steyvers M. Finding Scientific Topics [ J ]. National Academy of Sciences of the United States of America, 2004, 101(S1): 5228-5235.
- [ 8 ] Minka T P, Lafferty J. Expectation-propagation for the Generative Aspect Model [ C ] // Proceeding of the 18th Conference on Uncertainty in Artificial Intelligence. Boston, USA: AUAI Press, 2002: 352-359.
- [ 9 ] Blei D M, Lafferty J D. Correlated Topic Models [ C ] // Proceedings of NIPS '05. Cambridge, USA: MIT Press, 2005: 147-155.
- [ 10 ] Steyvers M, Smyth P, Griffiths T. Probabilistic Author-topic Models for Information Discovery [ C ] // Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2004: 306-315.
- [ 11 ] Wang X, Mccallum A. Topics Over Time: A Non-Markov Continuous-time Model of Topical Trends [ C ] // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2006: 424-433.
- [ 12 ] Blei D M, Lafferty J. Dynamic Topic Models [ C ] // Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA: IEEE Press, 2006: 113-120.
- [ 13 ] Ramage D, Dumais S, Liebling D. Characterizing Microblogs with Topic Models [ C ] // Proceedings of the 4th International AAAI Conference on Weblogs and Social Media. Menlo Park, USA: AAAI Press, 2010: 130-137.
- [ 14 ] Diao Q, Jiang J, Zhu F, et al. Finding Bursty Topics from Microblogs [ C ] // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. New York, USA: ACM Press, 2012: 536-544.
- [ 15 ] Wang Y, Agichtein E, Benzi M. TM-LDA: Efficient Online Modeling of Latent Topic Transitions in Social Media [ C ] // Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining. Beijing, China: [ s. n. ], 2012: 123-131.
- [ 16 ] Khalid E A, Min X, Emily B F. Representing Documents Through Their Readers [ C ] // Proceedings of the 19th International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2013: 14-22.
- [ 17 ] Cheng Z, Caverlee J, Lee K. You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users [ C ] // Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2010: 759-768.
- [ 18 ] 王莎, 张连明. 基于标签的微博人脉网络挖掘算法和结构分析 [ J ]. 计算机工程, 2014, 40(5): 7-11.  
编辑 陆燕菲
- (上接第29页)
- [ 8 ] 王新芳, 张冰, 冯友兵. 基于粒子群优化的改进加权知心定位算法 [ J ]. 计算机工程, 2012, 38(1): 90-95.
- [ 9 ] 章磊, 段莉莉, 钱紫鹃, 等. 基于遗传算法的 WSN 节点定位技术 [ J ]. 计算机工程, 2010, 36(10): 85-87.
- [ 10 ] Kannan A A, Mao Guoqiang, Vucetic B. Simulated Annealing Based Localization in Wireless Sensor Network Localization with Flip Ambiguity Mitigation [ C ] // Proceedings of the 63rd IEEE Vehicular Technology Conference. Washington D. C., USA: IEEE Press, 2006: 1022-1026.
- [ 11 ] 郭永红, 万江文, 于宁, 等. 基于跳数的无线传感器网络定位求精算法 [ J ]. 计算机工程, 2009, 35(3): 145-147.
- [ 12 ] Ribeiro V J, Riedi R H, Baraniuk R G. Locating Available Band-width Bottlenecks [ J ]. IEEE Internet Computing, 2004, 8(5): 34-41.
- [ 13 ] 戴佩华, 薛小平, 邵玉华. 基于垂直平分线的区域定位算法 [ J ]. 计算机工程, 2009, 35(2): 105-108.
- [ 14 ] 周四清, 陈锐标. 无线传感器网络 APIT 定位算法及其改进 [ J ]. 计算机工程, 2009, 35(7): 87-89.
- [ 15 ] 汪炆, 黄刘生, 肖明军, 等. 一种基于 RSSI 校验的无线传感器网络节点的定位算法 [ J ]. 小型微型计算机系统, 2009, 30(1): 59-62.
- [ 16 ] 万国峰. 基于锚节点和高斯函数的测距算法 [ J ]. 计算机工程, 2013, 39(2): 73-76.  
编辑 金胡考