计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20148133
52(11):2517-2526,2015

EMTM:微博中与主题相关的专家挖掘方法

张腊梅[1,2,3,4]　黄威靖[3,4]　陈　薇[3,4]　王腾蛟[3,4]　雷　凯[1,2]
[1](深圳市云计算关键技术与应用重点实验室(北京大学)　广东深圳　518055)
[2](北京大学信息工程学院　广东深圳　518055)
[3](高可信软件技术教育部重点实验室(北京大学)　北京　100871)
[4](北京大学信息科学技术学院　北京　100871)
(citlmzhang@163.com)

EMTM: A Method for Experts Mining in Micro-Blog with Topic-Level

Zhang Lamei[1,2,3,4], Huang Weijing[3,4], Chen Wei[3,4], Wang Tengjiao[3,4], and Lei Kai[1,2]
[1](*Shenzhen Key Laboratory for Cloud Computing Technology & Applications (Peking University)*, *Shenzhen*, *Guangdong* 518055)
[2](*School of Electronics and Computer Engineering*, *Peking University*, *Shenzhen*, *Guangdong* 518055)
[3](*Key Laboratory of High Confidence Software Technologies (Peking University)*, *Ministry of Education*, *Beijing* 100871)
[4](*School of Electronics Engineering and Computer Science*, *Peking University*, *Beijing* 100871)

Abstract So far, micro-blog has been one of the most popular platforms for people to access and share information. After long-term development, there are many experts with authoritative professional background knowledge. Mining experts in topic-level will contribute to the user recommendation and public opinion analysis in micro-blog. In micro-blog, experts in a topic are the users who have high influence on the topic, since they have authoritative professional knowledge and skills about the topic. High influence is a necessary condition for experts. Influence analysis belongs to subjective problems and need to be quantified objectively. In micro-blog, the probability of being retweeted is one of the most important indexes to measure the influence of users. So we can find out the high influencers by analyzing the retweet data. But, there are two kinds of retweet behaviors for the users in micro-blog: "topic-sensitive retweet" and "following retweet". Therefore, the users who have high influence because of being retweeted with high probability are not always experts. In this paper, we propose a probability generation model EMTM (experts mining topic model) which can find out the experts in topic-level by distinguishing two kinds of the retweet behaviors. We use Gibbs sampling for model inference. Our experiments on real Sina Weibo data show that our model EMTM is effective in mining experts in topic-level.

Key words experts; topic; micro-blog; retweet behavior; probability model

摘　要 目前,微博已成为人们获取信息、分享信息的最流行平台之一.经过长期的发展积累,微博中聚集了很多具有权威专业知识背景的专家,挖掘微博中与主题相关的专家有利于进一步地用户推荐、微博

收稿日期:2014-10-17;**修回日期**:2014-12-31
基金项目:国家"八六三"高技术研究发展计划基金项目(2012AA011002);国家自然科学基金项目(61300003);教育部高等学校博士学科点专项科研基金项目(20130001120001)
通信作者:陈　薇(pekingchenwei@pku.edu.cn)

舆情分析等工作.在微博中,与某个主题相关的专家是指因具有可靠的与此主题相关的专业知识或技能而在此主题下具有高影响力的用户.挖掘高影响力的用户可以通过分析微博的转发数据来进行,然而由于微博中用户的转发行为分为"主题相关转发"和"跟随转发"2 种,因此,因被转发概率高而具有高影响力的用户不一定是专家.EMTM(experts mining topic model)是一种基于主题模型的概率生成模型,通过区分微博用户的不同转发行为来挖掘微博中与主题相关的专家.模型采用 Gibbs 采样进行推理求解.在真实的新浪微博数据集上的对比实验表明 EMTM 能够有效地挖掘微博中与主题相关的专家.

关键词 专家;主题;微博;转发行为;概率模型

中图法分类号 TP311

近年来,微博作为当代最流行的基于关注机制的用户获取、分享以及传播信息的平台受到很多研究者的关注,如文献[1-2]关注发现微博中的影响力用户.本文关注于挖掘微博中与主题相关的专家.专家(expert)是指因具有可靠的专业知识或技能而在其专业领域具有高影响力的人①.相应地,微博中与某个主题相关的专家是指由于具有可靠的与此主题相关的专业知识或技能而在此主题下具有高影响力的用户.微博经过长期发展已经积累了很多有权威专业知识的专家,对于初次想要关注主题 T 信息的用户而言,更希望看到关于主题 T 有专业价值的微博,为他推荐与主题 T 相关的真正有权威的专家,将更符合他的兴趣.另外,当某个事件发生时,能够及时挖掘与事件主题相关的专家,将专家对事件的分析推荐给用户能够正确地引导舆论走向.因此,挖掘微博中与主题相关的专家是 1 个非常重要的问题,有利于进一步地用户推荐、社区中真正有影响力的权威用户发现和舆情分析等工作.

某用户 U 是与主题 T 相关的专家的 1 个必要条件是此用户 U 在主题 T 下有较高的影响力.影响力分析属于主观问题,需要进行客观性量化,被转发概率是传统的衡量用户影响力的重要指标之一[2-3],本文以用户 U 在主题 T 下被转发的概率来衡量用户 U 在主题 T 下的影响力.

然而,完全按照在某个主题下被转发概率的高低来判断某个用户是否是关于此主题的专家是不可行的.因为,在微博中,用户关注的单向性导致了用户的受欢迎程度是不平衡的,少量用户(大 V、名人等)拥有大量来自多个领域的关注者,如影视明星的关注者不只来自影视圈,还包含很多其他领域的用户.相应地,用户会有 2 种转发行为:1)微博的主题是用户和微博发布者都主要关注的主题,并且微博内容是有专业价值的,因此用户产生了转发行为,本文将此种转发行为称为"主题相关转发";2)微博内容并不具备与其主题相关的专业价值,但微博发布者的受欢迎程度高(如名人、大 V 等),用户出于对微博发布者的跟随心理而转发,本文将此种转发行为称为"跟随转发".

因此,在主题 T 下能够获得较高转发概率的用户尽管在主题 T 下是具有高影响力的,但不一定是与主题 T 相关的专家,因为知名度很高但并不具备与主题 T 相关的专业知识的用户(名人等)发布的与主题 T 相关的微博,由于其粉丝数量多,又有很多粉丝是"跟随转发",同样能够在主题 T 下获得很高的转发概率.同时,由于微博中用户受欢迎程度的不平衡性[4],会出现知名度很高的用户的微博被大量转发,导致与主题相关的一些知名度并不高的专家被淹没,该现象在本文的实验中得到了验证.

本文利用微博转发数据,基于主题模型提出一种能够挖掘微博中与主题相关的专家的概率生成模型 EMTM(experts mining topic model).该模型可以对微博主题语义和主题中用户被转发概率分布同时进行建模,并且通过区分微博用户的"主题相关转发"和"跟随转发"2 种转发行为,实现了微博中与主题相关的专家的挖掘.与传统的影响力用户发现方法 TwitterRank[1] 的对比以及自对比实验验证了 EMTM 的有效性,并从侧面说明了区分微博用户转发行为在解决与主题相关的专家挖掘问题中的重要性;同时,与 TwitterRank 方法的实验对比分析也显示了传统的影响力用户发现方法并不能直接用于解决微博中与主题相关的专家挖掘问题.另外,由于主题的语义建模效果会间接影响到与主题相关的专家发现的效果,因此本文将 EMTM 与传统主题模型 LDA 以困惑度为指标进行实验对比,说明了 EMTM 能够改善对微博主题的语义建模效果.

本文的工作与传统的影响力用户发现工作不

① https://en.wikipedia.org/wiki/Expert

同:传统的影响力用户发现工作目的是发现那些具有高影响力的用户;而本文的目的则是发现那些在某个主题 T 下由于"主题相关转发"而得到高影响力的用户,即与主题相关的专家.传统的影响力用户发现工作没有将用户的关注或者转发行为进行区分,因此知名度高的用户倾向于在多个主题下具有高影响力[1-2],若单纯将影响力排名在前 N 的用户当作专家,那么知名度不是很高但与主题相关的专家的排名将被知名度很高但并不是与此主题相关的专家的用户挤出"专家"的行列.因此不能将传统的影响力用户发现方法直接用于挖掘专家的问题上.

1 相关工作

1.1 主题模型

主题模型是一种对文档集中文本隐含语义主题建模的方法.常用的主题模型有 LSA 模型[5]、PLSA 模型[6]以及 LDA[7]模型,而目前应用最广泛的是 LDA 主题模型.LDA 主题模型假设每篇文章都有自己的主题分布,文章中的每个词都由这些隐含的主题生成,可以用来得到每个主题下的词频率分布,从而得到每个主题的抽象描述.

目前,利用主题模型对用户主题建模的研究主要基于 AT 模型[8],AT 模型在 LDA 的基础上引入作者及其主题分布变量,主要用来得到主题中词分布和用户的主题分布.CUT 模型[9]和 ART 模型[10]对 AT 模型进行扩展,分别用于社区发现和邮件分析系统中.尽管文献[8-10]已经可以解决各自所提出的问题,但是它们都是从用户集合中选择用户后由用户选择词,都是对用户的主题分布进行建模,并没有真正对用户的关系进行建模,直接将它们用于微博的转发数据中将不能体现用户的转发关系.

另外,还有一些基于主题模型单纯分析用户网络结构的研究,如文献[11-12],这些研究都只局限于对用户的链接关系建模,并没有加入语义分析.

1.2 影响力用户发现

近年来,影响力用户发现问题的研究主要分为 3 个方向:

1) 基于 PageRank[13]分析网络结构中影响力用户的研究.PageRank 最初应用在搜索引擎中,根据网页之间的超链接关系和互增强性质,利用随机游走思想对网页的重要性进行排序,之后研究者将 PageRank 算法应用于社会网络中,根据用户之间的链接关系对用户影响力进行排序.此类研究中较有代表性的发现微博中影响力用户的方法是 TwitterRank,TwitterRank 在利用用户间链接关系分析用户影响力的同时考虑了用户所发布微博的主题,它首先用 LDA 对微博主题进行建模,然后在用户的网络结构上基于 PageRank 分别计算每个主题下用户的影响力,用以发现与主题相关的影响力用户.

2) 基于用户属性的影响力用户发现研究.这类研究主要分析用户属性与用户影响力之间的关系,以发现更好地衡量用户影响力的指标,如文献[2].

3) 基于信息扩散的影响力用户研究.此类研究主要集中在解决用户影响力扩散最大化问题,即如何选择合适的种子使得信息在网络中得到最大扩散,如文献[14].

本文关注的挖掘微博中与主题相关的专家的问题与影响力用户发现问题比较相近,但目的不相同.若将用户的转发关系看作用户的链接关系,那么用户影响力分析中的第 1 类研究与本文的研究最相近;但是这类研究与本文的不同点在于,此类研究并没有考虑到用户间关系形成的不同原因,而是统一将用户之间的关系归结为兴趣相同,这将对专家的挖掘形成干扰.本文区分了用户的"主题相关转发"行为和由名人效应引起的"跟随转发"行为,从而能够有效地挖掘与主题相关的专家.

2 EMTM 模型介绍

本文提出的 EMTM 模型是一种基于主题模型的概率生成模型.EMTM 模型假设每条微博有 1 个隐含主题,根据微博的隐含主题所对应的词概率分布生成微博中的词,同时根据其所对应的被转发用户概率分布生成微博的被转发者,因此,EMTM 可同时对微博主题中词概率分布和主题中用户被转发概率分布进行建模.另外,为了区分被转发用户是由于"主题相关"而被转发还是由于"跟随"而被转发,EMTM 模型中假设有 2 种用户被转发的概率分布:1)由于用户被"跟随转发"产生的用户被转发概率分布;2)由于用户被"主题相关转发"产生的与主题相关的用户被转发概率分布.EMTM 在生成微博的被转发用户时,首先需要对 1 个开关变量进行 1 次伯努利实验,伯努利实验的不同结果分别对应微博用户的不同转发行为("主题相关转发"和"跟随转发"行为),则在生成被转发用户时就从相应的不同概率分布中生成,从而达到区分微博用户的不同转发行为的目的.EMTM 模型可得到各个主题中被"主题

相关转发”的用户概率分布,将得到的每个主题 T 中所有被转发用户的概率从高到低排序,排名在前 N 的被转发用户为 EMTM 挖掘出的与主题 T 相关的专家.

2.1 相关术语解释及符号说明

主题 T 相关的转发微博 d:微博 d 是 1 条转发微博,并且微博 d 的主题是 T.

1) 转发微博 d 的被转发用户 a. 指微博 d 是用户 a 发布的微博,并且微博 d 被其他用户转发. 如转发微博“//@a:微博内容”,则 a 是被转发用户,本文不考虑转发链,1 条转发微博对应 1 个被转发用户.

2) 主题 T 中用户 a 的被转发概率. 指在所有与主题 T 相关的转发微博的被转发用户中,用户 a 的占比. 如与主题 T 相关的转发微博数是 1 000,这 1 000 条微博对应 1 000 个被转发用户,这 1 000 个被转发用户是有重复的,如果用户 a 有 20 条属于主题 T 的微博被转发,那么用户 a 将在这 1000 个用户中重复出现 20 次,即主题 T 中用户 a 的被转发概率是 20/1 000=2%.

3) 用户 a 被“主题相关转发”. 指用户转发用户 a 的微博 d 是一种“主题相关转发”行为.

4) 用户 a 被“跟随转发”. 指用户转发用户 a 的微博 d 是一种“跟随转发”行为.

EMTM 模型中的符号说明如表 1 所示:

Table 1 Symbol Table for EMTM

表 1 EMTM 模型中符号说明表

Symbol	Description
U	The set of the users who have retweet behavior. The size of the set is $\|U\|$.
D_u	Retweet set of the u-th user.
N_{ud}	Number of words in the u-th user's d-th retweet.
K	Topic set whose size is $\|K\|$.
V	Word set whose size is $\|V\|$.
A	The set of users who was retweeted by others. The size of the set is $\|A\|$.
w_{udn}	The n-th word in the u-th user's d-th retweet.
a_{ud}	The u-th user's d-th retweet retweeted from user a.
z_{ud}	The topic of the u-th user's d-th retweet.
$\boldsymbol{\Phi}_k$	Word distribution for the k-th topic. It's a vector which has $\|V\|$ dimensions.
$\boldsymbol{\Phi}_b$	Word distribution for background topic. It's a vector which has $\|V\|$ dimensions.
$\boldsymbol{\Pi}_k$	User distribution for the k-th topic. It's a vector which has $\|A\|$ dimensions. Notice that users here are the elements of set A.
$\boldsymbol{\Pi}_b$	User distribution for popular topic. It's a vector which has $\|A\|$ dimensions. Notice that users here are the elements of set A.
$\boldsymbol{\Psi}$	The parameter vector for word binary indicator distribution which has $\|V\|$ dimensions, and $\boldsymbol{\Psi}_w$ is the binary indicator distribution for word w.
$\boldsymbol{\Gamma}$	The parameter vector for user binary indicator distribution which has $\|A\|$ dimensions, and $\boldsymbol{\Gamma}_a$ is the binary indicator distribution for user a. Here, users are the elements of set A.
f_{udn}	Binary indicator for w_{udn}: if $f_{udn}=1$, w_{udn} is generated from $\boldsymbol{\Phi}_k$; if $f_{udn}=0$, w_{udn} is generated from $\boldsymbol{\Phi}_b$.
s_{ud}	Binary indicator for a_{ud}: if $s_{ud}=1$, a_{ud} is generated from $\boldsymbol{\Pi}_k$; if $s_{ud}=0$, a_{ud} is generated from $\boldsymbol{\Pi}_b$.
$\boldsymbol{\theta}_u$	Topic distribution for the u-th user. It's a vector which has $\|K\|$ dimensions.
$\boldsymbol{\alpha}$, $\boldsymbol{\beta}_b$, $\boldsymbol{\beta}$, $\boldsymbol{\rho}$, $\boldsymbol{\rho}_b$, $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_1$, $\boldsymbol{\delta}_0$, $\boldsymbol{\delta}_1$	Prior parameters.

2.2 模型生成过程描述

EMTM 模型的概率生成框架图如图 1 所示.

在 EMTM 模型中,对于每个用户 $u\in U$,生成 u 的主题多项式分布参数 $\boldsymbol{\theta}_u$,$\boldsymbol{\theta}_u$ 服从以 $\boldsymbol{\alpha}$ 为参数的狄利克雷分布. 对于用户 u 的每条转发微博 $d_u\in D_u$,根据 $\boldsymbol{\theta}_u$ 为这条转发微博 d_u 选取 1 个主题 z_{ud},根据主题 z_{ud} 所对应的词的多项式分布参数 $\boldsymbol{\Phi}_k$ 可生成微博 d_u 中的词语 w_{udn},根据主题 z_{ud} 所对应的被转

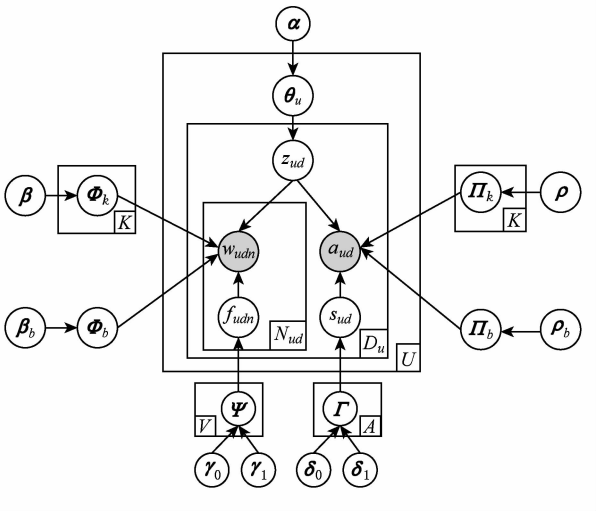


Fig. 1 Probability generation graph of EMTM.

图 1 EMTM 模型概率生成图

发用户的多项式分布参数 $\boldsymbol{\Pi}_k$ 可生成微博 d_u 的被转发用户 a_{ud},因此,EMTM 模型可同时建模微博主题的词概率分布以及主题的用户被转发概率分布.

为了区分用户的不同转发行为,EMTM 中存在 2 种用户被转发的概率分布.在生成转发微博 d_u 的被转发用户 a_{ud} 时,首先用开关变量 s_{ud} 控制用户 u 的转发行为,s_{ud} 由 1 次伯努利实验生成:$s_{ud}=1$ 表示用户 u 就微博 d_u 而言,对被转发用户 a_{ud} 的转发行为是"主题相关转发",a_{ud} 根据主题 z_{ud} 对应的用户被转发概率分布参数 $\boldsymbol{\Pi}_k$ 生成;$s_{ud}=0$ 表示用户 u 对被转发用户 a_{ud} 的转发行为是"跟随转发",即用户 a_{ud} 根据 $\boldsymbol{\Pi}_b$ 生成.其中,s_{ud} 根据 a_{ud} 对应的二项分布参数 $\boldsymbol{\Gamma}_a$ 生成,δ_{a0} 和 δ_{a1} 是 $\boldsymbol{\Gamma}_a$ 服从的 Beta 先验,且 $\delta_{a1}=\lambda_a/\ln(fn_a)$,$\delta_{a0}=\max(0,1-\delta_{a1})$,$fn_a$ 表示用户 a 的被转发总次数,即被转发总次数越多的用户被"主题相关转发"的概率越低,被"跟随转发"的概率相应越高.因为,被转发总次数多的用户倾向于是受欢迎的名人,其微博由于粉丝数量很多,更容易获得很高的被转发次数.

对于转发微博 d_u 的 N_{ud} 个词中的每个词 w_{udn},由于微博中存在大量如"关注""转发"等与主题无关的词语,因此在生成微博中的词时同样设置开关变量 f_{udn} 控制词 w_{udn} 是否根据相应的主题分布参数 $\boldsymbol{\Phi}_k$ 生成.其中,开关变量 f_{udn} 根据 w_{udn} 对应的二项分布参数 $\boldsymbol{\Psi}_w$ 生成,γ_{w0} 和 γ_{w1} 是 $\boldsymbol{\Psi}_w$ 服从的 Beta 分布先验参数,并且,$\gamma_{w1}=\lambda_w/\lg(n_w)$,$\gamma_{w0}=\max(0,1-\gamma_{w1})$,$n_w$ 表示词 w 在所有被转发微博中出现的总次数,即出现次数越多的词语与主题相关性越差.

假设整个转发微博集合共有 $|K|$ 个主题,EMTM 模型的生成过程如下(其中,在生成过程中如 $\boldsymbol{\theta}_u\sim Dir(\boldsymbol{\alpha})$ 表示 $\boldsymbol{\theta}_u$ 根据以 $\boldsymbol{\alpha}$ 为参数的狄利克雷分布生成,其他类似符号同理,Beta 表示贝塔分布,Multinomial 表示多项式分布,Binomial 表示伯努利分布).

Step1. 对于每个主题 $k\in K$:

选取 $\boldsymbol{\Phi}_k\sim Dir(\boldsymbol{\beta})$,$\boldsymbol{\Pi}_k\sim Dir(\boldsymbol{\rho})$;

Step2. 选取 $\boldsymbol{\Phi}_b\sim Dir(\boldsymbol{\beta}_b)$,$\boldsymbol{\Pi}_b\sim Dir(\boldsymbol{\rho}_b)$;

Step3. 对于词集合 V 中的每个词 $w\in V$:

选取 $\boldsymbol{\Psi}_w\sim Beta(\gamma_{w0},\gamma_{w1})$;

Step4. 对于被转发用户集合 A 中的每个被转发用户 $a\in A$:

选取 $\boldsymbol{\Gamma}_a\sim Beta(\delta_{a0},\delta_{a1})$;

Step5. 对于转发用户集合 U 中的每个用户 $u\in U$:

Step5.1. 选取 $\boldsymbol{\theta}_u\sim Dir(\boldsymbol{\alpha})$;

Step5.2. 对于用户 u 的每条转发微博 $d_u\in D_u$:

Step5.2.1. 选取主题 $z_{ud}\sim Multinomial(\boldsymbol{\theta}_u)$,主题 z_{ud} 对应主题集合 K 中的主题 k;

Step5.2.2. 对于被转发用户 a_{ud}:

选取 $s_{ud}\sim Binomial(\boldsymbol{\Gamma}_a)$,$a_{ud}$ 对应被转发用户集合 A 中的 a:

Step5.2.2.1. 如果 $s_{ud}=1$,则选取 $a_{ud}\sim Multinomial(\boldsymbol{\Pi}_k)$;

Step5.2.2.2. 如果 $s_{ud}=0$,则选取 $a_{ud}\sim Multinomial(\boldsymbol{\Pi}_b)$;

Step5.2.3. 对微博 d_u 的 N_{ud} 个词 w_{udn},逐个:

选取 $f_{udn}\sim Binomial(\boldsymbol{\Psi}_w)$,$w_{udn}$ 对应词集合 V 中的词 w:

Step5.2.3.1. 如果 $f_{udn}=1$,则选取 $w_{udn}\sim Multinomial(\boldsymbol{\Phi}_k)$;

Step5.2.3.2. 如果 $f_{udn}=0$,则选取 $w_{udn}\sim Multinomial(\boldsymbol{\Phi}_b)$.

2.3 Gibbs 采样

基于 Gibbs[15] 采样的参数推理方法实现简单且有效,因此目前对主题模型的参数求解一般均使用 Gibbs 采样算法.本文采用 Gibbs 采样来近似推导 EMTM 的参数,得到公式如下:

$$p(z_{ud}=k\mid\bullet)\propto\frac{\alpha+n_{u,k}^{-d_u}}{|K|\alpha+n_{u,*}^{-d_u}}\cdot\frac{\rho+m_{k,a}^{-d_u}}{|U|\rho+m_{k,*}^{-d_u}}\cdot\frac{\prod_{v=1}^{|V|}\left(\prod_{p=0}^{n_{ud}^{(v)}-1}(n_{k,v}^{-d_u}+p+\beta)\right)}{\prod_{b=0}^{n_{ud}^{(*)}-1}\left(\sum_{v=1}^{|V|}n_{k,*}^{-d_u}+b+|V|\beta\right)},$$

$$p(f_{udn}=t\mid\cdot)\propto\frac{n_{w,t}^{-d_u}+\gamma_{wt}}{n_{w,0}^{-d_u}+n_{w,1}^{-d_u}+\gamma_{w0}+\gamma_{w1}},$$

$$p(s_{ud}=t\mid\cdot)\propto\frac{n_{a,t}^{-d_u}+\delta_{at}}{n_{a,0}^{-d_u}+n_{a,1}^{-d_u}+\delta_{a0}+\delta_{a1}},$$

其中,$p(z_{ud}=k\mid\cdot)$是在迭代过程中为用户 u 的微博 d 选择主题为 k 的概率,公式中 $p(z_{ud}=k\mid\cdot)$的计算由 3 部分组成,分别对应 $\boldsymbol{\theta}_u$ 中主题 k 的概率、$\boldsymbol{\Pi}_k$ 中被转发用户 a 的概率、$\boldsymbol{\Phi}_k$ 中词 w 的概率,3 部分相互作用,尽管同一微博会被不同的用户转发,但该微博的主题的概率分布最终会采样迭代至平衡;$p(f_{udn}=t\mid\cdot)$是在迭代过程中词 w_{udn} 的开关变量选择概率,$t=1$ 表示 w_{udn} 属于主题相关词的概率,反之,$t=0$;$p(s_{ud}=t\mid\cdot)$是迭代过程中被转发用户 a_{ud} 的开关变量选择概率,$t=1$ 表示 a_{ud} 被"主题相关转发"的概率,反之,$t=0$. 所有公式中计数符号上标"$-d_u$"表示计数是在将当前微博 d_u 中的相应元素减去后的值. $n_{u,k}^{-d_u}$ 表示用户 u 中属于主题 k 的微博计数,$n_{u,*}^{-d_u}$ 表示用户 u 的微博个数. $m_{k,a}^{-d_u}$ 表示被转发用户 a 属于主题 k 的计数,$m_{k,*}^{-d_u}$ 表示属于主题 k 的所有被转发用户的计数. $n_{ud}^{(v)}$ 表示当前微博 d_u 中属于主题相关的词 v 的计数,$n_{ud}^{(*)}$ 表示当前微博 d_u 中所有属于主题相关的词的计数. $n_{k,v}^{-d_u}$ 表示词 v 属于主题 k 的计数,$n_{k,*}^{-d_u}$ 表示属于主题 k 的所有词的计数. $n_{w,t}^{-d_u}$ 表示开关变量 f_{udn} 控制的词 w_{udn} 在词表 V 中对应的词 w 被标记为 t 的计数,$n_{a,t}^{-d_u}$ 表示开关变量 s_{ud} 控制的被转发用户 a_{ud} 在被转发用户集合 A 中对应的被转发用户 a 被标记为 t 的计数. α,β,ρ 分别是先验参数向量 $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\rho}$ 的元素,主题模型中先验参数向量 $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\rho}$ 中的元素往往取相同的值.

3 实验与结果分析

本文所用实验数据为真实新浪微博数据,其中,数据中包含 3 万用户的转发微博共 544 479 条微博,利用分词工具[①] 进行分词,并且去除停用词与间接被转发的用户,实验数据中词表大小为 63 144,被转发用户个数为 22 510. 在提取被转发用户时,本文将新浪微博中以@开头的用户提取出来作为被转发用户,尽管在新浪微博中一部分被@的用户是被其他用户主动提及的用户,但是被提及也是一种影响,因此本文的处理方式是可行的.

通过实验,本文选取了经验值 $|K|=50$,$\alpha=0.1$,$\beta=\beta_b=0.01$,$\rho=\rho_b=0.01$,$\lambda_a=2.0$,$\lambda_w=1.0$.

本文设计 2 种实验:1)与传统影响力用户发现方法 TwitterRank 以及 EMTM 的变种模型 EMTM(−)进行对比实验,分析 EMTM 区分微博用户的转发行为后在挖掘微博中与主题相关的专家时的有效性,另外与 TwitterRank 的对比实验分析说明不能将传统的影响力用户发现工作直接用于解决微博中与主题相关的专家挖掘问题;2)由于主题的语义建模效果会间接影响到与主题相关的专家发现的效果,因此与传统 LDA 主题发现模型进行对比,分析 EMTM 的主题语义建模效果.

3.1 主题相关专家挖掘有效性分析

根据 EMTM 模型得到的每个主题 k 下所有被转发用户概率分布 $\boldsymbol{\Pi}_k$,其中 $\boldsymbol{\Pi}_k$ 是 1 个维度为 $|A|$ 的向量,$\Pi_{k,a}$ 表示被转发用户 a 在主题 k 下被转发的概率,将 $\boldsymbol{\Pi}_k$ 中元素从高到低排序,排名在前 N 的被转发用户为 EMTM 挖掘出的与主题 T 相关的专家.

本文将 EMTM 模型的结果和 EMTM(−)模型以及 TwitterRank 的结果进行对比,用以说明 EMTM 在解决微博中与主题相关的专家挖掘问题时的有效性. 其中,EMTM(−)模型是 EMTM 模型去掉被转发用户的开关变量后不区分用户的转发行为的概率生成模型;TwitterRank 首先根据 LDA 模型得到每个用户的主题分布,然后在用户的网络结构上基于 PageRank 分别计算每个主题下用户影响力,本文在 TwitterRank 的实验中网络结构采用转发网络.

本文采用准确率 P(precision)指标衡量与主题相关的专家发现的结果好坏. 主题 k 的 Top-N 的准确率 P_k 等于在主题 k 中被转发概率排名在前 N 的用户中是专家的个数 M 的占比,即 $P_k=M/N$,则所有主题 Top-N 平均准确率 $P=\left(\sum_{k=0}^{|K|}P_k\right)\Big/|K|$. 其中,通过人工核对的方法确定 Top-$N$ 中的用户 a 是否是与主题 T 相关的专家(满足下列 2 个条件之一即可):1)如果用户 a 在微博中的粉丝数超过 1 万,并且检查其微博或者标签等信息,如果微博信息或者标签显示用户 a 主要关注主题 T 并且是关于主题 T 的专业人士;2)能够通过搜索引擎搜索用户 a

① http://ictclas.nlpir.org/

返回的前 50 个结果中查到用户 a 是关于主题 T 的专家的相关信息,则确定用户 a 是与主题 T 相关的专家.另外,对于每个主题本文抽取主题词概率分布从高到低排名前 20 的词代表此主题.如果 2 个主题的代表词中有 75%以上的重合率,则认为这 2 个主题相同.尽管不同的方法得到的主题不完全统一,但是通过实验可观察到,对于同一个数据集 EMTM,EMTM(−),TwitterRank 这 3 种方法分别得到的所有主题中有相同主题的概率为 90%以上.

本文将 EMTM,EMTM(−),TwitterRank 这 3 种方法的主题个数均设为 50,在 50 个主题中随机抽样其中 30 个主题的结果进行 Top-N 的平均准确率比较,其中 N 分别取 5,10,15,20,25,对比结果如图 2 所示:

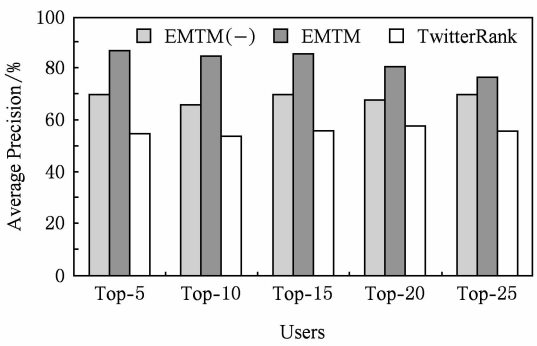


Fig. 2 Comparison of average precision for Top-N users being experts.

图 2 Top-N 用户中是专家的平均准确率对比

图 2 实验结果显示,EMTM 比不区分用户转发行为的 EMTM(−)和传统影响力用户分析 TwitterRank 方法在挖掘微博中主题相关专家的问题上都有较高的准确率.TwitterRank 的目的是发现主题下有影响力的用户,与本文的目的不同.EMTM 和 TwitterRank 的对比结果可以从侧面验证微博中影响力高的用户并不一定是专家的观点.

通过实例举例进一步分析,以挖掘出的主题集合中的“跑车”主题为例说明.EMTM,EMTM(−),TwitterRank 这 3 种方法中关于“跑车”主题的 Top-5 的结果分别见表 2～4(各表中的灰色行表示与“跑车”主题不相关的用户.各表中粉丝数等信息都是本文写作时通过新浪微博人工搜索得到的信息,粉丝数是写作当天查询到的结果).

由表 2～4 的实例对比结果可以看出,EMTM(−)和 TwitterRank 都会将与“跑车”主题不相关的用户(如“优酷网”和“新浪视频”)推到较高的排名,说明了 EMTM 模型中对用户转发行为进行区分的有效性.在不区分用户转发行为的方法中,粉丝数很高的大 V 用户对专家挖掘结果会造成影响,如“新浪视频”,由于其发布的关于跑车的新闻(如车展、跑车出交通事故等)得到了很多粉丝的“跟随转发”,同时跑车主题在整个数据集中占比较小,大 V 用户被大量跟随转发在跑车主题下被推高排名,使得与此主题更相关的用户排名降低,影响了真正专家的挖掘.

Table 2 Top-5 Experts Mined by EMTM for Sports Car Topic

表 2 EMTM 关于跑车主题得到的 Top-5 的专家

User Name	Followers	Description
跑车世界	130 000	The king of the micro-bloggers about car.
名车志	1 660 000	The office micro-blog for the famous magazine which is named《CAR AND DRIVER》.
新浪试车	220 000	Report authority car evaluation. Now its name is “汽车评测”.
宝马中国	840 000	The office micro-blog for BMW.
许群	590 000	Magazine publisher for magazines which are named《ramp 驾道》and《TAGET 商品评介》.

Table 3 Top-5 Experts Mined by EMTM(−) for Sports Car Topic

表 3 EMTM(−)关于跑车主题得到的 Top-5 的专家

User Name	Followers	Description
跑车世界	130 000	The king of the micro-bloggers about car.
新浪汽车	3 390 000	Report news about car. It is the office micro-blog for Sina Car.
名车志	1 660 000	The office micro-blog for the famous magazine which is named《CAR AND DRIVER》.
优酷网	none	This user is not the real user in Sina Weibo. It appears in micro-blog because the tweet is shared from the youku website. So it starts with @ in the tweets, and is extracted in the reason of being regarded as a user.
许群	590 000	Magazine publisher for magazines which are named《ramp 驾道》and《TAGET 商品评介》.

Table 4 Top-5 Influencers Mined by TwitterRank for Sports Car Topic

表 4 TwitterRank 关于跑车主题得到的 Top-5 的影响力用户

User Name	Followers	Description
跑车世界	130 000	The king of the micro-bloggers about car.
优酷网	none	This user is not the real user in Sina Weibo. It appears in micro-blog because the tweet is shared from the youku website. So it starts with @ in the tweets, and is extracted in the reason of being regarded as a user.
新浪试车	220 000	Report authority car evaluation. Now its name is "汽车评测".
新浪视频	6 810 000	Report all kinds of new things. It is the office micro-blog for Sina Video.
改装志	none	It is a famous magazine about modification for the sports car, but it has not office micro-blog. It appears in the dataset because of being mentioned by users initiatively in the way of @.

在 TwitterRank 的实验结果中,94%的主题中排名在 Top-5 的用户中都有用户"优酷网"存在,42%的主题中排名在 Top-5 的用户中都有用户"新浪视频"存在,这符合文献[1]中 TwitterRank 实验的结果,即有些用户会在多个主题中有较高的影响力,同时,说明了在某个主题下有影响力的用户不一定是此主题的专家.其次,"优酷网"尽管不是微博中真实的用户,但"优酷网"在多个主题下都能有高影响力是因为优酷网是目前国内最受欢迎的视频网站之一,很多用户都从优酷视频网站分享了各个主题的视频到微博中,但优酷网在各个主题中更偏向于新闻视频主题.若将"优酷网"看作 1 个非常受欢迎的用户,那么他就如"新浪视频"一样由于被大量粉丝在多个主题下"跟随转发",而在多个主题下被推到高排名的位置,影响这些主题下的专家挖掘效果.因此,TwitterRank 可以用来挖掘与主题相关的影响力用户,但并不能被直接用来解决与主题相关的专家挖掘问题.

接下来进一步说明 EMTM 通过区分用户的 2 种转发行为使得挖掘的与主题相关的专家与相应主题更有相关性.以实验结果中"优酷网"、"新浪视频"、"许群"、"宝马中国"4 个用户为例.

由区分 2 种转发行为的 EMTM 方法所得结果中"优酷网"、"新浪视频"、"许群"、"宝马中国"这 4 个用户被"主题相关转发"和被"跟随转发"的概率,如图 3 所示.概率值由实验 7 次取平均结果得到.由图 3 可以看出,由于"优酷网"和"新浪视频"是明星用户,使其被"主题相关"与被"跟随转发"的比例比用户"许群"和"宝马中国"的比例小.

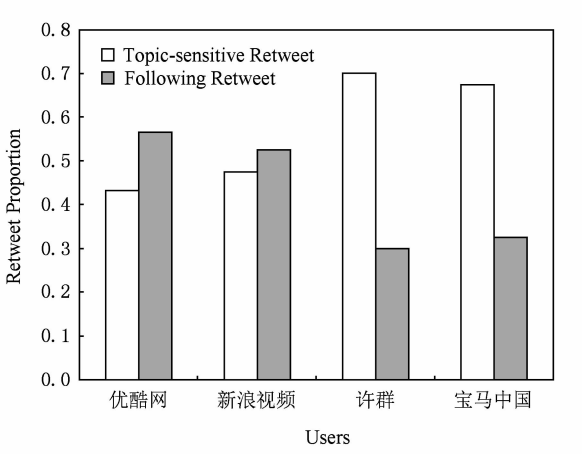


Fig. 3 Probability of being retweeted due to different retweet behavior.

图 3 用户被不同转发行为转发的概率

各用户在各个主题下被转发概率的分布比例如表 5 所示(每个用户只列出其相对比例高的主题和跑车主题).用户 a 在主题 k 下被转发的概率计算为 $\Pi_{ka}\Big/\sum_{k}\Pi_{ka}$,由表 5 可以看出在所有主题中"优酷网"、"新浪视频"约一半的概率是关于视频新闻主题的,而大约只有 1%是关于跑车汽车相关主题的;"许群"、"宝马中国"约 80%都是跑车主题.分析数据集中"优酷网"、"新浪视频"的微博发现其关于跑车的主题中约 85%以上都是关于车的新闻,如车展、车辆事故等.由此可以看出,在跑车主题下"许群"、"宝马中国"相比于"优酷网"、"新浪视频"更有相关性,对于想关注跑车主题的用户,为其推荐"许群"比为其推荐"新浪视频"将更符合其兴趣.

Table 5 Users' Topic Probability Distribution

表 5 用户在各主题下被转发的概率分布

User Name	Topic Key Words	Probability
优酷网	视频 独家 mv 分享 首播	0.464 4
	改装 奥迪 BMW F1 跑车	0.004 2
新浪视频	分享 视频 方舟 演唱会 五月	0.289 0
	中国 钓鱼岛 日本 美国	0.157 1
	比赛 视频 进球 中国 体育	0.100 5
	改装 奥迪 BMW F1 跑车	0.010 4
许群	改装 奥迪 BMW F1 跑车	0.866 7
宝马中国	改装 奥迪 BMW F1 跑车	0.714 3

3.2 主题建模效果分析

由于主题的语义建模效果会间接影响到与主题相关的专家发现的效果,因此本文以困惑度(perperlexity)[16]为指标进行模型 EMTM 与传统主题模型 LDA 的对比实验,以评估模型 EMTM 对主题语义的建模效果.

困惑度是度量主题模型效果的常用指标,表示预测数据时的不确定度,值越小代表主题建模的效果越好.困惑度计算公式为

$$Perperlexity(W) = \exp\left\{-\frac{1}{N}\sum_{n=1}^{N}\ln p(w_n)\right\},$$

其中,W 为测试数据集,w_n 为测试数据集中可观测到的词,N 为词个数.本文在实验数据集外,另选 6 000 个用户的所有转发微博作为测试数据集,在相同先验参数设置下计算困惑度.每次 Gibbs 采样迭代 100 轮,设置 $|K|=50$,共运行 10 次 Gibbs 采样得到 LDA 和 EMTM 模型的困惑度平均值如表 6 所示.通过对比困惑度可以发现,EMTM 对主题的建模比传统 LDA 对主题的建模效果好.

Table 6 Comparison of Perperlexity

表 6 困惑度对比结果

Model	Perplexity
EMTM	5 743.351
LDA	21 027.203

4 总结与未来工作

微博作为近年来人们获取信息、分享信息最流行的平台之一,挖掘微博中与主题相关的专家无论是在用户推荐还是舆情分析中都是很重要的.专家是有影响力的用户,但是由于微博中用户受欢迎程度非常不均衡,使得微博中有影响力的用户不一定是专家.本文提出了一种通过区分用户转发行为挖掘微博中与主题相关的专家的概率生成模型 EMTM,并且采用 Gibbs 采样方法对模型进行求解.在真实新浪数据集上的对比实验验证了 EMTM 在解决微博中与主题相关的专家挖掘问题时的有效性,以及 EMTM 对主题语义建模的有效性.

在将来的工作中,本文将进一步探索引入用户的个人标签等属性建模用户兴趣,用以辅助区分用户 A 对用户 B 的转发行为,进而能够更好地挖掘微博中与主题相关的专家.

参 考 文 献

[1] Weng J, Lim E P, Jiang J, et al. TwitterRank: Finding topic-sensitive influential Twitterers [C] //Proc of the 3rd ACM Int Conf on Web Search and Data Mining. New York: ACM, 2010: 261-270

[2] Cha M, Haddadi H, Benevenuto F, et al. Measuring user influence in Twitter: The million follower fallacy [C] //Proc of the 4th Int Conf on Web and Social Media. Palo Alto, CA: AAAI, 2010: 10-17

[3] Pal A, Counts S. Identifying topical authorities in microblogs [C] //Proc of the 4th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2011: 45-54

[4] Fan Pengyi, Wang Hui, Jiang Zhihong, et al. Measurement of microblogging network [J]. Journal of Computer Reasearch and Development, 2012, 49(4): 691-699 (in Chinese)
(樊鹏翼, 王晖, 姜志宏, 等. 微博网络测量研究[J]. 计算机研究与发展, 2012, 49(4): 691-699)

[5] Papadimitriou C H, Tamaki H, Raghavan P, et al. Latent semantic indexing: A probabilistic analysis [C] //Proc of the 7th ACM SIGACT-SIGMOD-SIGART Symp on Principles of Database Systems. New York: ACM, 1998: 159-168

[6] Hofmann T. Probabilistic latent semantic indexing [C] //Proc of the 22nd Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 1999: 50-57

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

[8] Steyvers M, Smyth P, Rosen-Zvi M, et al. Probabilistic author-topic models for information discovery [C] //Proc of the 10th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2004: 306-315

[9] Zhou D, Manavoglu E, Li J, et al. Probabilistic models for discovering e-communities [C] //Proc of the 15th Int Conf on World Wide Web. New York: ACM, 2006: 173-182

[10] McCallum A, Wang X, Corrada-Emmanuel A. Topic and role discovery in social networks with experiments on enron and academic email [J]. Journal of Artificial Intelligence Research, 2007, 30: 249-272

[11] Cha Y, Bi B, Hsieh C C, et al. Incorporating popularity in topic models for social network analysis [C] //Proc of the 36th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2013: 223-232

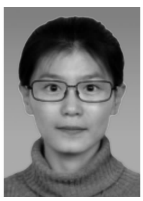
[12] Cha Y, Cho J. Social-network analysis using topic models [C] //Proc of the 35th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2012: 565-574

[13] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine [C] //Proc of the 7th Int World Wide Web Conf. New York: ACM, 1988: 107-117

[14] Barbieri N, Bonchi F, Manco G. Topic-aware social influence propagation models [J]. Knowledge and Information Systems, 2013, 37(3): 555-584

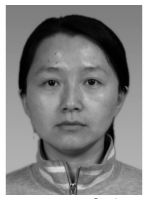
[15] Steyvers M, Griffiths T. Probabilistic topic models [J]. Handbook of Latent Semantic Analysis, 2007, 427(7): 424-440

[16] Wallach H M, Murray I, Salakhutdinov R, et al. Evaluation methods for topic models [C] // Proc of the 26th Annual Int Conf on Machine Learning. New York: ACM, 2009: 1105-1112



Zhang Lamei, born in 1988. Received her master degree from Peking University in July, 2015. Her main research interests include data mining, etc.

Huang Weijing, born in 1986. PhD candidate at Peking University. His main research interests include topic detection and tracking, machine learning and social networks.



Chen Wei, born in 1981. Received her PhD degree from the School of Electronics Engineering and Computer Science, Peking University in 2009. Assistant professor at the School of Electronics Engineering and Computer Science, Peking University. Her main research interests include artificial intelligence, social networks analysis and Internet public opinion, etc.

Wang Tengjiao, born in 1973. Professor at the School of Electronics Engineering and Computer Science, Peking University. Senior member of China Computer Federation. His main research interests include data warehousing, data mining and Web information processing.

Lei Kai, born in 1976. Received his MS degree from the Department of Computer, Columbia University in 1999. Now executive vice-director of the Center for Internet Research and Engineering, Shenzhen Graduate School, Peking University. Member of IEEE and ACM. His main research interests include database, Internet information processing and mining, etc.