社区中最具影响力博客的探测模型

卢 露[1,2] 丁才昌[1]

(长江大学计算机科学与技术学院 荆州 434023)[1] (武汉大学计算机科学与技术学院 武汉 430079)[2]

摘 要 虚拟社区中的一些博客通过发表高质量的博文,能够影响社区中其他成员的观点,指导其他成员的行动,具有一定的影响力。这些博客被认为是重要的信息源,探测这样颇具影响力的博客的活动对外部世界的决策有重要意义。利用博客特有的格式,提出度量影响力的相关标准,定量计算出博客在特定主题中的影响力。实验证明,该模型能够克服目前通过简单统计特征识别重要博客的缺点,是一种行之有效的方法。

关键词 LDA模型,热点主题,博客影响度

Model of Identifying the Influentials in Blog Community

LU Lu[1,2] DING Cai-chang[1]

(School of Computer, Yangtze University, Jingzhou 434023, China)[1]

(School of Computer, Wuhan University, Wuhan 430079, China)[2]

Abstract Some bloggers in the virtual community could impact other partners' opinion and guide the actions of other members by publishing high-quality blog post, which is considered to have high influence. We regarded them as important informative resources. Detecting the influential's activities gives assistance to external world's decision. We proposed some measurements of influence by taking advantage of the blog's unique form, and quantify influence with respect to topics. The experiment shows that model can overcome shortcomings of only using the simple statistical characteristics to identify important blogger, which is an effective method.

Keywords LDA model, Hot topic, Blog influence

1 引言

Web 2.0概念的出现使互联网新媒体的发展进入了新阶段,网络用户不再仅是信息的消费者,更成为信息的生产者。博客作为Web2.0下的一种新兴事物正处于高速发展时期,它一般被认为是一种网上在线日记,其内容由普通用户发表和维护。它由一系列带有时间戳的博文构成,还包括各种标记和超链接,人们在博客中能够自由地发表个人观点、新闻、知识及信息,使人们获得信息的来源不再仅仅限于传统的平面媒体、门户网站,更多的是来自博客站点。具有相同兴趣爱好的博客通过各种形式的超链接进行交流,进而容易形成虚拟社区。与真实社区一样,虚拟社区中的某些成员经常能够作为重要的信息源,提供有价值的信息;或是针对重大事件提出新颖独特的观点,吸引其他成员参与讨论,并能够影响后者的观点,我们称之为有影响力的博客。找出社区中的具有影响力的博客有重要的现实意义。它能帮助企业有效地分析用户偏好,发现用户的兴趣偏移;能够发现社区中人们对热点事件的主流观点和看法等等。那么如何找出社区中这样一些有影响力的博客呢?这一问题的关键在于如何度量博客的影响力,这是一个主观的行为,主要依赖于各种不同的需求。本文根据博客特有的格式,从内容和链接的角度提出了有影响力的相关标准:

1. 博客必须参与社区中热点主题的讨论。

2. 博客发表的博文经常作为重要信息源,能够吸引大量的读者,使其参与主题的讨论。

3. 博客发表的博文观点新颖,能够经常被社区中同主题的其他博文所引用,受到其他博主的认同和好评。

我们对上述标准进行了形式化定义,定量计算出各个博客的影响度值,并通过在真实数据集上进行测试,找出具有影响力的前top-n个博客。试验证明该模型能够克服目前通过简单统计特征识别重要博客的缺点,是行之有效的。

2 相关工作

博客技术飞速发展,使得博客空间中的知识发现和数据挖掘受到越来越多的关注,它能重塑商业模式,激发虚拟市场,提供趋势分析和销售预测,目前成为获取信息源的一个重要平台。Lin[1,2]等人利用个人blog行为和语义链接结构,使用相互感知特性和基于排序的社区抽取算法发现博客社区。Kiriakopoulos[3]利用网页排序的策略来发掘重要的博客站点,他们利用主题的相似性增加隐含边,使博客空间这个社会网络变得更加稠密,然后利用HITS[4]和PageRank[5]算法计算重要性值。Gruhlet[6]利用级联模型研究各种主题在博客空间中的传播现象。Mei[7]等人基于概率统计理论的无指导学习方法挖掘博客时序、空间的主题模式。Durant[8]利用贝

卢 露(1981—),男,博士生,讲师,主要研究方向为人工智能、社会网络分析,E-mail:lulugraphics@163.com;丁才昌(1980—),男,博士生,讲师,主要研究方向为人工智能、Web数据挖掘。

叶斯和支持向量机分类器来挖掘政治博客的情感。

3 社区中热点主题的获取

博客具有影响力的条件之一是必须参与社区中各种热点主题的讨论，如何获得当前的热点主题是一个首待解决的问题。

我们利用爬虫获取社区中各个成员在一段时间内的博文，得到一个博文集合。然后利用产生式概率模型(LDA 模型)挖掘博文集中的热点，该模型是一种无监督机器学习的方法，不仅能够自动识别出博文集中的热点主题和热点词语，还可以将热点词语自动归类到对应的主题中，直观地反映了博文集中的热点。

3.1 LDA 模型[9]

LDA 模型是指由 Blei 等人在 2003 年提出的一个三层贝叶斯产生式概率模型，该模型假设文档由一系列潜在主题混合而成，主题由词汇表中所有的词汇混合而成，不同文档的主要区别在于它们的主题混合比例不同。

概率主题模型按如下方式进行形式化，记 $P(z)$为文档中主题 z 的分布，$P(w|z)$为给定主题 Z 后词语的分布概率。那么，文档中每一个词语 w 的产生过程为，首先从 $P(z)$中获取一个主题，然后根据 $P(w|z)$来得到一个词语 w。记 $P(z_i=j)$是第 j 个主题的概率，$P(w_i|z_i=j)$为主题 j 中词语 w_i 的概率。那么概率主题模型指定了文档中词语的概率分布：

$$P(w_i)=\sum_{j=1}^{T}P(w_i|z_i=j)P(z_i=j)$$

式中，T 是主题的个数。为简单起见，记 $\phi^{(j)}=P(w|z=j)$为主题 j 中词语的多项分布，$\theta^{(d)}=P(z)$为文档 d 中的主题多项分布。另外，记文档集合为 D，每个文档 $d\in D$ 包含 N_d 个词语，所有短语的总数为 $N=\Sigma N_d$。参数 ϕ 和 θ 分别表示了主题中词语的重要程度和文档中主题的重要程度。模型采用 Dirichlet 分布作为多项式分布 ϕ 和 θ 的共轭先验，简化了模型的统计推导。超参数 α、β 通过 Dirichlet 分布 $Dir(\alpha)$、$Dir(\beta)$控制主题分布 θ 和主题上的词语分布 ϕ。θ 和 ϕ 进一步决定了文档中词语的 w。在此定义下，包括文档上的主题分布 θ、主题上的词语分布 ϕ、主题 z 和 w 单词在内的各变量之间的联合概率密度为：

$$p(w_d,z_d,\theta_d,\phi|\alpha,\beta)=p(\phi|\beta)\prod_{n=1}^{N_d}p(w_{d,n}|\phi^{(z_{d,n})})p(z_{d,n}|\theta_d)p(\theta_d|\alpha) \quad (1)$$

消除变量 θ_d、ϕ 和 z_d，可以得到文档的单词 w_d 的似然值：

$$p(w_d|\alpha,\beta)=\iint p(\theta_d|\alpha)p(\phi|\beta)\prod_{n=1}^{N_d}p(w_{d,n})|\phi^{(z_{d,n})})p(z_{d,n}|\theta_d)\mathrm{d}\phi\mathrm{d}\theta_d \quad (2)$$

由于同时存在多个隐变量，上式不能直接通过推理求出，因此我们采用 Gibbs 抽样估计当前采样词 w 的主题 z 的后验分布，然后得到模型的参数 θ 和 ϕ。

Gibbs 抽样是一种马尔科夫链蒙特卡洛方法(Markov chain Monte Carlo，MCMC)，它能够非常有效地从大规模文集中抽取主题，而且实现起来比较简单。因此，Gibbs 抽样算法成为目前最流行、最常用的 LDA 模型抽取算法。

Gibbs 抽样算法与其他算法的不同之处在于，该方法并不直接计算每个文档的 θ 和 ϕ，而是先根据文档中可见的词语序列，通过计算 z 的后验分布，间接地统计出 θ 和 ϕ。每个词语 i 对应的主题变量 z_i 被赋予[1,2,…,T]中的某个整数 t，代表这个单词对应的是第 t 个主题。

具体地，从词汇对于主题的后验概率 $P(w|z)$出发，使用 Gibbs 抽样的关键是构造目标概率分布函数，这里只需要对变量 z_i 进行抽取样本。计算后验概率 $P(z_j=j|z_{-i},w_i)$的公式如下：

$$P(z_i=j|z_{-i},w_i)\propto\frac{n_{-i,j}^{w_i}+\beta}{n_{-i,j}^{(\cdot)}+V\beta}\cdot\frac{n_{-i,j}^{d_i}+\alpha}{n_{-i}^{d_i}+K\alpha} \quad (3)$$

式中，$z_j=j$ 表示主题 j 包含词语 w_i(w_i 是词汇 w 在所给的文档中位置权值)。z_{-i}表示所有 $w_k(k\neq i)$对该主题的分配。$n_{-i,\cdot}^{(d_i)}$ 是文档 d_i 中所有主题包含的词语个数。$n_{-i,\cdot,j}^{(d_i)}$是文档 d_i 中主题 j 包含的词语个数。$n_{-i,\cdot,j}^{(\cdot)}$是主题 j 包含除 $z_i=j$ 的所有词汇个数。$n_{-i,\cdot,j}^{(w)}$是主题 j 包含 w_i 的个数。进而可求得文档在主题上的分布 θ 和词汇在主题上的分布 ϕ。

$$\theta_j^d=\frac{n_j^d+\alpha}{n_{\cdot}^d+K\alpha} \quad (4)$$

$$\phi_w^j=\frac{n_j^w+\beta}{n_{j\cdot}^{\cdot}+V\beta} \quad (5)$$

式中，n_j^d 表示博文 d 中分配给主题 j 的词数，$n_{\cdot}^d$ 表示博文 d 中所有被分配了主题的词数之和，n_j^w 表示词汇 w 被分配给主题 j 的次数，$n_{j\cdot}^{\cdot}$ 表示被分配给主题 j 的所有词数。

3.2 基于 LDA 模型的热点主题分析

在社区的热点主题识别研究中，博客发布的博文集中的特征词是整个模型唯一可观察的变量，LDA 模型在已知主题数目的情况下，通过调节特征词语在潜在主题上的概率分布，完成博文的生成过程。在此过程中，可以获得每个特征词语在各个潜在主题上的概率分布以及每篇博文在潜在主题上的概率分布。如果一篇博文在某个潜在主题的概率分布值越高，那么成为该主题的可能性就越大。如果该潜在主题同时又是其他博文的主题，那么它就是整个博文集中的热点主题。我们把博文在潜在主题空间中概率分布值最高的前两位主题作为博文的主题，然后统计出每个主题所包含博文数，选择包含博文数最多的前若干位的主题作为该社区中的热点主题。

4 博客在社区中的影响度分析

利用 LDA 模型，我们从大量的博文中提取出热点主题集合 $T=\{T_1,T_2,\cdots,T_k\}$。计算博客在社区中的影响力必须以特定主题为依托，不能一概而论。比如博客 B 经常发表关于主题 A 的博文，吸引其他博主通过大量的评论信息与其交流，如其他博主在同主题的博文中也经常引用该博主的博文，说明博客 B 在该主题上有很强的影响力。反之博客 B 关于主题 C 的博文很少，也没有相关的反馈信息，那么他在主题 C 中的影响力就相对较低。所以我们应分别计算出博客在各个热点主题中的影响力，而博客在整个社区中的影响力为该博客在各个热点主题影响力之和。那么如何定量计算博客在特定主题上的影响力 $I(B,T)$?

我们分析博客在该主题下发表博文的数量和质量，提出以下度量标准，并给出形式化定义。

1. 博客必须发表一定量与该主题有关的文章。

我们认为博客在某主题上具有影响力，必须对该主题有浓厚的兴趣，能够积极参与讨论，发文数量多，频度高。否则

他在此主题上的影响力很低。其形式化定义如下。

对于博客 B 和某一主题 T，$Post(B,T)$ 表示博客 B 中属于主题 T 的博文数。$m(B,T)$ 为博客 B 中主题为 T 的博文所占比例：

$$m(B,T)=\frac{Post(B,T)}{Post(B)} \tag{6}$$

$m(B,T)$ 的值能够定义标准一。

2. 博客必须在此主题下发表高质量的博文，这些博文可作为关于外界事物的重要信息源，或是对外界事物提出了自己独特新颖的观点，吸引社区中的博主或社区外的读者积极参与讨论，与其交流。博客提供了各种链接形式，作为博客之间相互交流的手段。这里我们主要分析博文中的评论链接和引用链接来度量博文的影响力。评论链接出现在博文的末尾，是读者关于此博文信息的意见反馈。如果博文的评论数很少，说明该篇文章的影响力很小。反之，如果博文后有大量的评论内容，说明受到的关注越多，影响力就越强。引用链接出现在博文中，是其他相关主题的博文对此博文的引用，表示对该博文观点的支持、借鉴等等。我们认为博客的博文能够被其他博文所引用，说明该博文观点新颖独特，具有一定的影响力。反之，如果博客引用他人的博文，说明他缺乏独创性，只是附和其他人的观点，影响力较小。这两种链接可以从侧面反映博文的质量和影响力，我们分别给出两种链接形式化的定义。

评论链接反映了读者的参与程度。对于博客 B 中属于主题 T 的任一博文 p，它含有的评论数 $comm(p)$ 越多，影响力就越大。因为每篇博文的评论数不同，简单起见，我们可设定阈值 $\tau_{\max}$，如果博文 p 的评论数超过 $\tau_{\max}$，就认为这篇博文的评论影响度为 1，否则为 $\frac{|Comm(p)|}{\tau_{\max}}$。则博客 B 在主题 T 上的评论影响度 $\psi(B,T)$ 为：

$$\psi(B,T)=\frac{\sum_{p\in Post(B,T)}\min(\frac{|Comm(p)|}{\tau_{\max}},1)}{|Post(B,T)|} \tag{7}$$

博文的引用链接反映了博文的新颖程度。如果在该博文中引用其他博文，且与引用博文的内容很相似，我们认为该博文可能只是对其他的博文观点的补充，新颖度不高。反之，如果该博文被其他博文引用，说明该博文可能是重要信息源或观点的发起者，有一定影响力。这里我们只从博文出现链接的角度来定义，如果博客 B 中的博文 p（博文 p 属于主题 T）引用其他博文的集合为 $out(p)$，则我们定义博文 P 的新颖度 $Nov(p)$ 为：

$$Nov(p)=\min_{q\in out(p)}(1-\cos(p,q)) \tag{8}$$

式中，$\cos(p,q)$ 表示博文 P 和博文 q 在文本上的余弦相似性。

博客 B 在主题 T 的创新性影响度 $\theta(B,T)$ 为：

$$\theta(B,T)=\frac{\sum_{p\in T}Nov(p)}{post(B,T)} \tag{9}$$

以上标准是从博文的角度来定义。从博客的角度分析，且博客 A 在某一领域是具有影响力的，同时他又是博客 B 的追随者，且博客 B 在该领域也发表了大量的博文，则我们可推断博客 B 在该领域也应该具有影响力。这种影响力之间的传递关系符合 Pagerank 算法的思想。Pagerank 算法利用网页之间的链接关系迭代计算网页的重要性值，在迭代初始为每个节点赋于均匀大小的初始值。在该算法的基础上做出改进：用上述提出的三个标准值的乘积作为迭代的初始值，而不是选用均匀大小的初始值，这样可以减少迭代次数，加快收敛时间，使最后的结果更加合理。计算博客在具体主题下的影响力迭代公式如下：

$$I_{i+1}(B,T)=1-\delta+\delta.\sum_{C\in IN(B)}\frac{I_i(C,T)}{OUT(C)} \tag{10}$$

式中，$I_{i+1}(B,T)$ 为迭代 $i+1$ 次后博客 B 关于主题 T 的影响力值。δ 为随机跳动因子，通常选用 0.85。设置常量 ε，当 $I_n(B,T)-I_{n-1}(B,T)<\varepsilon$，迭代结束。初始值 $I_0(B,T)$ 为：

$$I_0(B,T)=m(B,T).\psi(B,T).\theta(B,T) \tag{11}$$

最后博客总的影响力值为博客在每个热门主题上的影响力值之和：$I(B)=\sum_{n=1}^{k}I(B,T_n)$。

我们取社区中的影响度值最大的前 top-k 个博客作为社区中的具有影响力的博客。

5 实验及分析

我们利用博客爬虫从新浪的博客空间中爬取 5000 个博客及它们之间的链接关系，并设计社区发现算法[10]找出社区结构（有共同的兴趣爱好且交流频繁的博客组成）。我们选中其中一社区，包含 308 个社区成员及其 24025 篇博文，他们之间包含 3784 个评论链接和 947 个引用链接。针对每一篇博文进行分词和词性标注，考虑到热点大多表现为名词和名词短语，因此在词性标注后，只统计名词和名词短语的词频，通过设置频度阈值，过滤掉总频度低于 6 次以及所出现的博文数低于 10 篇的词汇，得到一张 3215 个词汇组成的词汇表。

由模型的求解过程可知，模型中存在 3 个可变量，Dirichlet 分布中的超参数 $\alpha\beta$ 和潜在主题数目 K，为了有效利用 Gibbs 抽样算法，需要确定这 3 个可变量的最佳取值。

本实验根据经验值确定 α,β 的取值，令 $\alpha=50/K$，$\beta=0.1$，$K=20$。模型中 3 个可变量确定后，结合输入向量运行 Gibbs 抽样，得到词汇表中的特征词语在 20 个潜在主题上的概率分布以及潜在主题在每篇博文中的概率分布。将每篇博文中概率值排在前两位的主题看作是该博文的主题。统计所有博文的主题，然后按照主题出现次数的多少判定该主题是否为主题热点。最终我们选取前 10 位的主题为热点主题。

为了判断所提出的探测模型是否合理有效，我们从最具影响力博客的内容和链接两个角度提出如下评估标准进行度量。

1. 覆盖面。具有影响力的博客，他们的观点应该尽可能多地传递给社区中其他成员，覆盖面广。我们可以统计社区中有多少成员可以直接或间接地与这些具有影响力的博客交流，从社会网络分析的角度，我们计算出有多少博客可以通过各种链接直接或间接地到达我们获取的有影响力的博客集中。

2. 涉及主题的多样性。我们利用模型探测出的有影响力的博客集中，应该尽量包含在各个热门主题中具有影响力的博客，保证主题的多样性。对于项集的多样性度量，采用如下的公式，V 代表所选取的博客，博客之间的余弦相似性为博客在主题空间上的相似性（利用 LDA 模型构造博客在主题空间的向量）。

$$Diversity(v_1,v_2,\cdots,v_n)=\sum_i\sum_j(1-\mathrm{Cos}(v_i,v_j))/n(n-1)/2$$

我们与如下 3 种方法在上述两种标准下进行比较。

1. Pagerank 方法：该方法纯粹以链接结构为基础计算博客的重要性。选取前 K 个重要度值最大的博客作为最具影响力的博客。

2. 随机取样(RS)的方法：随机选取 K 个博客作为最具影响力的博客。

3. 基于发文数排序方法：选取发文数最多的前 K 个博客作为最具影响力的博客。这是大多数网站常用的博客推荐方法。

如图 1、图 2 所示，我们分别比较了各种算法的博客直接覆盖数和间接覆盖数，我们的算法和 PageRank 算法在不同的 N 值上，覆盖面较高，且曲线越来越平滑。而随机取样算法和基于博文数统计的方法覆盖面较低，且直线陡峭，说明它们选取的点在影响力上是无序的排列，也能进一步说明发表博文数越多的博客，其影响力不一定越大，从而在大多数网站上所采用的基于博文统计数的博客推荐是不合理的。

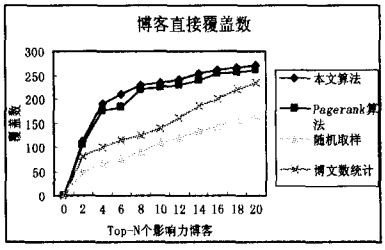

图 1　各种方法的博客直接覆盖数比较

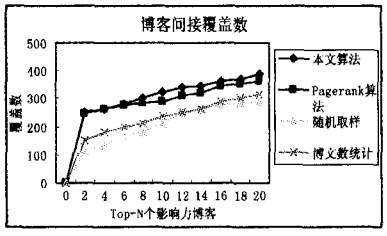

图 2　各种方法的间接覆盖数比较

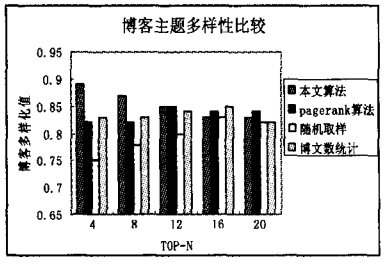

图 3　探测到影响力博客的主题多样性比较

如图 3 所示，我们选取的博客中能覆盖的主题相对较多，特别是在 N 值较小时能覆盖到更多样化的主题。当 N 值逐渐变大时，各种方法的主题覆盖率相对较为接近。

结束语　本文提出了一种探测影响力博客的模型，该模型根据博客特有的格式，不仅考虑了博客在热点主题中的博文数，还分析了引用链接、评论链接这些和影响力密切相关的博客信息，使其结果相比于通过简单统计特征识别重要博客更加合理。在以后的工作中，我们还应充分考虑链接下的语义信息，对各种评论链接、引用链接下的观点进行分析，即是赞同，是反对还是无倾向性的一般陈述，并由此对链接设定权值，计算博客的声誉度，使所提出的模型更加完善。

参 考 文 献

[1] Lin Y R, Sundaram H, Chi Y, et al. Discovery of blog communities based on mutual awareness[C]// Proc. of the World Wide Web 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. Edinburgh, 2006

[2] Lin Y R, Sundaram H, Chi Y, et al. Blog community discovery and evolution based on mutual awareness expansion[C]// Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. 2007: 48-56

[3] Kritikopoulos A, Sideri M, Varlamis I. Blogrank: ranking Weblogs based on connectivity and similarity features[C]// AAA-IDEA'06: Proceedings of the 2nd international workshop on Advanced architectures and algorithms for internet delivery and applications. 2006: 8

[4] Kleinberg J M. Authoritative sources in a hyper-linked environment[C]// SODA '98: Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms. 1998: 668-677

[5] Page L, Brin S, Motwani R, et al. The pagerank citation ranking: Bringing order to the Web[R]. Stanford Digital Library Technologies Project, 1998

[6] Durant K T, Smith M. Mining sentiment classification from political Weblogs[C]// Proc. of WebKDD workshop inconj with ACM SIGKDD. Philadelphia, PA, August 2006

[7] Gruhl D, Guha R, Liben-Nowell D, et al. Information Diffusion through Blogspace[C]// Proceedings of the 13th International Conference on World Wide Web. 2004: 491-501

[8] Mei Q, Liu C, Su H, et al. A Probabilistic Approach to Spatio-temporal Theme Pattern Mining on Weblogs[C]// Proceedings of the 15th International Conference on World Wide Web. 2006: 533-542

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

[10] Lu Lu, Zhu Fu-xi, Hu Bin. A novel method to detect latent community in blogspace[J]. Journal of Computational Information Systems, 2010, 7: 2151-2157

(上接第 160 页)

[3] Buyya R. High Performance Cluster Computing: Architectures and Systems, Volume1[M]. 郑纬民，石威，汪东升，等译. 北京：电子工业出版社，2002

[4] Buyya R. High Performance Cluster Computing: Programming and Applications, Volume2[M]. 郑纬民，石威，汪东升，等译. 北京：电子工业出版社，2002

[5] 陈庆奎，那丽春. 基于强化学习的多机群网格资源调度模型[J]. 计算机科学，2007(11)

[6] Chen Qing-kui, Wang Hai-feng, Wang Wei. Continuance parallel computation grid composed of multi-clusters[J]. Journal of Networks, 2010, 5(1): 3-10

[7] 王国俊. 计算智能-词语计算与 Fuzzy 集[M]. 北京：高等教育出版社，2006

[8] 陈水利，李敬功，王向松. 模糊集理论及其应用[M]. 北京：科学出版社，2005