DOI:10.11992/tis.201706096
网络出版地址:http://kns.cnki.net/kcms/detail/23.1538.TP.20171021.1349.004.html

基于用户查询日志的网络搜索主题分析

张森[1], 张晨[1,2], 林培光[1], 张春云[1],郭玉超[1], 任威龙[1],任可[2]
(1.山东财经大学 计算机科学与技术学院,山东 济南 250014;2. 香港科技大学 计算机科学及工程学系,香港 999077)

摘　要:网络搜索分析在优化搜索引擎方面具有举足轻重的作用,而且对用户个人搜索特性进行分析能够提高搜索引擎的精准度。目前,大多数已有模型(比如点击图模型及其变体),注重研究用户群体的共同特点。然而,关于如何做到既可以获取用户群体共同特点又可以获取用户个人特点方面的研究却非常少。本文研究了基于个人用户网络搜索分析新问题,即通过研究用户搜索的突发性现象,获取个人用户搜索查询的主题分布情况。提出了两个搜索主题模型,即搜索突发性模型(SBM)和耦合敏感搜索突发性模型(CS-SBM)。SBM 假设查询词和 URL 主题是无关的,CS-SBM 假设查询词和 URL 之间是有主题关联的,得到的主题分布信息存储在偏 Dirichlet 先验中,采用 Beta 分布刻画用户搜索的时间特性。实验结果表明,每一个用户的网络搜索轨迹都有多种基于用户的独有特点。同时,在使用大量真实用户查询日志数据情况下,与 LDA、DCMLDA、TOT 相比,本文提出的模型具有明显的泛化性能优势,并且有效地描绘了用户搜索查询主题在时间上的变化过程。

关键词:网络搜索;搜索引擎;自然语言处理;主题模型;文本挖掘;突发性;时间分析;参数估计
中图分类号:TP391　　**文献标志码**:A　　**文章编号**:1673-4785(2017)05-0668-10

中文引用格式:张森, 张晨, 林培光,等.基于用户查询日志的网络搜索主题分析[J]. 智能系统学报, 2017, 12(5): 668-677.
英文引用格式:ZHANG Sen, ZHANG Chen, LIN Peiguang, et al. Web search topic analysis based on user search query logs[J]. CAAI transactions on intelligent systems, 2017, 12(5): 668-677.

Web search topic analysis based on user search query logs

ZHANG Sen[1], ZHANG Chen[1,2], LIN Peiguang[1], ZHANG Chunyun[1],
GUO Yuchao[1], REN Weilong[1], REN Ke[2]
(1. School of Computer Science & Technology, Shandong University of Finance & Economics, Jinan 250014, China; 2.Department of Computer Science & Engineering, Hong Kong University of Science and Technology, Hong Kong 999077, China)

Abstract: Web search analysis plays a critical role in improving the performance of contemporary search engines. In addition, search engine accuracy can be improved by analyzing the individual search properties of users. Most existing models, such as the click graph and its variants, focus on the common characteristics of the group. However, as yet, there has been little investigation of a model that would obtain both the collective group characteristics and the unique characteristics of individual users. In this paper, we investigate user-specific web search analysis, whereby we obtain the topic distributions of the search queries of individual users by determining the burstiness of user searches. We propose two topic models, i.e., the search burstiness model (SBM) and the coupling-sensitive search burstiness model (CS-SBM). The SBM adopts the assumption that the query words and URL are topically independent, The CS-SBM supposes that the query words and URL are topically relevant. The obtained topic distribution information is stored in skewed Dirichlet priors and a beta distribution is used to capture the temporal properties of the user searches. Our experimental results show that each user's web search trail has unique characteristics, and that in the case of there being a large amount of real query log data, in comparison to the latent Dirichlet allocation (LDA) and topic over time (TOT) models, our proposed models have advantages with respect to generalized performance and effectively describes the temporal change process of user search queries.
Keywords: web search; search engine; natural language processing; topic model; data mining; burstiness; temporal analysis; parameter estimate

1931 年,Zipf[1] 发现在自然语言中,词的频率与它在词汇表中的排名成反比,服从幂律分布,他把这种现象称为上下文语言模型中词的突发性。后来发现,在金融、基因表达、计算机视觉等方面的数据也存在这种突发现象。网络搜索已成为人们日常生活中必不可少的一部分,用户提交的搜索查询词是人类智慧的结晶,并在搜索查询和微博等网络

收稿日期:2017-07-01.　**网络出版日期**:2017-10-21.
基金项目:国家自然科学基金重点项目(U1201258) 教育部人文社会科学研究项目(15YJAZH042).
通信作者:张晨. E-mail: zhangchen.sdufe@gmail.com.

信息中显现出与传统的自然语言不同的特点,网络搜索中每一个用户的搜索条目都包括查询词和 URL 两项。已经提出的 Dirichlet Compound Multinomial (DCM)模型[2]和 Dirichlet Compound Multinomial Latent Dirichlet Allocation (DCMLDA)模型[3]可以对文章中词的突发性现象建模,但如果直接应用于网络搜索建模却不是很理想。虽然大多数的点击图模型[4]及其变体[5-6]可以对网络搜索建模,但都是针对用户群体进行研究而忽略了用户个人特点。

本文通过分析用户查询日志来获取网络搜索突发现象,并提出了两个模型:SBM (search burstiness model)和 CS-SBM (coupling-sensitive search burstiness model)。SBM 是一个单极模型,假设查询词和 URL 之间主题独立,突发性的相关信息存储在偏 Dirichlet 先验里。CS-SBM 充分考虑查询词和 URL 之间的关联。本文还用 Beta 分布刻画了用户搜索的时间特性,使前面提出的模型能够用来捕获时间上的突发性。

1　相关工作

Madsen 指出,多项分布经常用于文本建模。然而,多项分布能获取到文档中词汇的突发性现象,即一个词如果出现过一次,那么它很有可能再次出现[7]。因此 Madsen 提出了 Dirichlet 多项分布(DCM)来代替传统的多项分布。DCM 拥有一级自由度,能获取到词的突发性,但是没有涉及文档中词汇的主题。文献[2]将 DCM 模型扩展成为了混合 DCM 分布,该模型能够训练表示一组文档,其中每一个文档都来自不同的高级主题。但是,该模型还是不能建模一个文档包含多个主题单词的情景。

上述工作中,之所以不能很好地刻画文档主题,其主要原因是 DCM 更关注突发性现象而非获取文档主题。2003 年 Blei[8]提出的 Latent Dirichlet Allocation (LDA)是非监督的贝叶斯生成模型,它可以将文档集中每篇文档的主题按照概率分布的形式给出。LDA 包括词汇、主题和文档 3 层。LDA 引入了 Dirichlet 先验分布,成为了一个完备的贝叶斯模型。LDA 文档生成过程为,从 Dirichlet 分布中采样文档与主题、主题与词汇分布,再重复从文档-主题多项式分布中采样主题以及由主题-词汇多项式分布生成词汇的过程,逐步生成整个文档。LDA 已经在学术和工业界得到广泛应用。但是,LDA 模型并不能预测词汇突发性出现的趋势。

为了能够在获取主题的同时预测词汇突发性现象,G.Doyle[3]提出了 DCMLDA 主题模型,该模型结合了 DCM 和 LDA 的优势,直接将 DCM 扩展合并到主题模型里面形成了一个比 LDA 更加复杂的模型。在 DCMLDA 中对于每个主题 k 和每个文档 d 服从新的多项式分布 θ_{kd},每个主题 k 都有不同的、非均匀的 $\boldsymbol{\beta}_k$ 向量。对于每一个文档 d,φ_{kd} 根据 Dirichlet($\boldsymbol{\beta}_k$)的变化而变化,因此每个主题实例在文档之间是相互联系的。文档中的主题实例允许在同一主题不同文档中每一个词汇的概率不同,这也就是突发现象。

随着带有时间标记的文本集合(例如,数字化的报纸、杂志、博客等)数量和体积的增加,如何有效地搜索这些数据变得更加重要。上述模型都难以发现主题的演化趋势。在这个背景下,文献[9]等提出了 Topic Over Time(TOT)模型。TOT 将文档的时间信息作为服从 Beta 分布的变量,将每个主题通过 Beta 分布与时间信息相关联。TOT 假设每个生成的词汇对应的时间信息也是通过它所属主题相关的 Beta 分布采样生成,这样主题与时间信息也有关系。TOT 不依赖马尔可夫假说,这样能够避免在离散化过程中遇到时间粒度选取的问题。

另外,文献[10]系统地总结了自然语言处理中主题模型的发展,对 LDA 模型进行了详细的分析,并对主题模型的发展趋势进行了预测。根据微博的特殊形式,在 LDA 的基础上进行了改进,分别提出了(MicroBlogs-LDA) MB-LDA 模型[11]和(MicroBlogs-HDP)MB-HDP 模型[12],同时证明了提出模型能够很好地对微博进行主题挖掘。

以前的工作都是集中在对自然语言文本中的同质项目进行分析,即对一个文档中的同质词汇进行建模。然而在网络搜索分析中,文档是由查询词和 URL 两个异构项目组成,并且带有时间信息。因此,本文结合网络搜索查询的文本特点,提出并研究了将主题模型运用到网络搜索分析中,对查询词和 URL 这两个异构项目和它们之间的关系进行建模。

2　以用户为中心的概率主题模型

本节主要介绍了 SBM 和 CS-SBM 主题模型,以及获取主题时间突发性的策略。

提出的模型应具备以下条件:

1)查询词和 URL 的突发性现象研究要分开建模[13]。

2)建模时,网络搜索特点,包括查询词、URL 和 session 3 个维度都要考虑在内[14-15]。

session 是指在短时间内提交的满足相同信息需求的一系列查询。为了避免同一会话中包含不相干的查询而导致的性能降低,本文优先考虑同一会话中查询词之间的语义一致性。通过对比分析,本文采用文献[16]提出的一系列规则来划分搜索 session。这些规则用于评估查询之间的词汇相似性,并在检测相关搜索查询时表现出很高精度。

2.1 查询词和 URL 独立的主题模型(SBM)

SBM 的生成过程是基于查询词和 URL 相互独立的假设。与 LDA 和 DCMLDA 等传统主题模型不同,SBM 中的文档有查询词、被点击 URL 和查询时间三项。图 1 是 SBM 的搜索主题模型概率图。首先,从超参数为 α 的 Dirichlet 分布中抽样生成文档与主题之间的关系矩阵 $\boldsymbol{\theta}$,$\boldsymbol{\theta}$ 是一个 $D\times K$ 的矩阵,其中 D 代表文档数量,K 代表主题数。对于每一个主题 k,从超参为 β_k 和 δ_k 的 Dirichlet 分布中分别取样生成查询词与主题之间的关系矩阵 $\boldsymbol{\theta}$ 和 URL 与主题之间的关系矩阵 $\boldsymbol{\Omega}$。$\boldsymbol{\theta}$ 和 $\boldsymbol{\Omega}$ 是 $D\times K\times V$ 的三维矩阵,V 代表训练语料库中出现的所有词的词表。对于文档中的每一个 session,从参数为 θ_d 的多项分布中选择一个主题 z,从参数为 φ_{zd} 的查询词的多项分布中,采样生成查询词 w。然后,生成点击事件的二项分布,如果 URL 被点击了,从参数为 $\boldsymbol{\Omega}_{zd}$ 的 URL 的多项分布中,采样生成 URL u。变量 $\boldsymbol{\theta}$ 和 δ 是基于单个特定文档的,因此 SBM 可以为每一个用户查询词和 URL 的突发性建模。

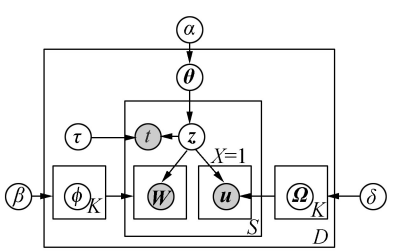


图 1 SBM 网络搜索分析主题模型

Fig.1 SBM web search analysis topic model

SBM 中的 Gibbs 采样方法借鉴了 LDA 和 DCMLDA 中 Gibbs 采样的方法,并进行了推导。模型的完全似然函数为

$$P(\boldsymbol{w},\boldsymbol{u},\boldsymbol{z}\mid\alpha,\beta,\delta)=P(\boldsymbol{u}\mid\boldsymbol{z},\delta)P(\boldsymbol{w}\mid\boldsymbol{z},\beta)P(\boldsymbol{z}\mid\alpha)$$

展开上式中的多项分布和 Dirichlet 分布,利用多项分布和 Dirichlet 分布的共轭性质,分别积分掉参数 $\boldsymbol{\theta}$ 和 φ 以后,通过借鉴 LDA 和 DCMLDA 中 Gibbs 采样的方法,在 SBM 中概率 $P(\boldsymbol{z}\mid\alpha)$ 为

$$P(z\mid\alpha)=\left(\frac{\Gamma(\sum_{z=1}^{K}\alpha_z)}{\prod_{z=1}^{K}\Gamma(\alpha_z)}\right)^{D}\prod_{d=1}^{D}\frac{\prod_{z=1}^{T}\Gamma(n_{dz}+\alpha_z)}{\Gamma(\sum_{z=1}^{Z}(n_{dz}+\alpha_z))}$$

式中 n_{dz} 为第 d 个文档中主题 z 的数量。

概率 $\boldsymbol{P}(\boldsymbol{w}\mid\boldsymbol{z},\beta)$ 为

$$\boldsymbol{P}(\boldsymbol{w}\mid\boldsymbol{z},\beta)=\prod_{d=1}^{D}\prod_{k=1}^{K}\left(\frac{\Gamma(\sum_{w=1}^{W}\beta_{kw})}{\prod_{w=1}^{W}\Gamma(\beta_{kw})}\frac{\prod_{w=1}^{W}\Gamma(n_{dkw}+\beta_{kw})}{\Gamma(\sum_{w=1}^{W}(n_{dkw}+\beta_{kw}))}\right)$$

式中 n_{dkw} 为第 d 个文档中第 k 个主题下查询词 w 的个数。

当该 session 中没有 URL 被点击时的条件概率为

$$P(z_i=k\mid X_i=0,z_{-i},\boldsymbol{w},\boldsymbol{u},\alpha,\beta,\delta)\propto\frac{C_{dk}^{DK}+\alpha_k}{\sum_{k'=1}^{K}(C_{dk'}^{DK}+\alpha_k')}\cdot\frac{\Gamma(\sum_{t=1}^{W}(C_{kwd}^{KWD}+\beta_{wk}))}{\Gamma(\sum_{t=1}^{W}(C_{kwd}^{KWD}+\beta_{wk}+N_{iw}))}\prod_{w=1}^{W}\frac{\Gamma(C_{kwd}^{KWD}+\beta_{wk}+N_{iw})}{\Gamma(C_{kwd}^{KWD}+\beta_w)}$$

式中:C_{dk}^{DK} 表示文档 d 中分配主题为 k 的 session 的数量,C_{kwd}^{KWD} 表示文档 d 中查询词 w 被分配主题 k 的次数,N_{iw} 表示第 i 个 session 中查询词 w 的个数。

当一个 session 中有 URL 被点击时的条件概率为

$$P(z_i=k\mid X_i=1,z_{-i},\boldsymbol{w},\boldsymbol{u},\alpha,\beta,\delta)\propto\frac{C_{dk}^{DK}+\alpha_k}{\sum_{k'=1}^{K}(C_{dk'}^{DK}+\alpha_k')}\cdot\frac{\Gamma(\sum_{w=1}^{W}(C_{kwd}^{KWD}+\beta_{wk}))}{\Gamma(\sum_{w=1}^{W}(C_{kwd}^{KWD}+\beta_w+N_{iw}))}\prod_{w=1}^{W}\frac{\Gamma(C_{kwd}^{KWD}+\beta_{wk}+N_{iw})}{\Gamma(C_{kwd}^{KWD}+\beta_{wk})}\cdot\frac{\Gamma(\sum_{u=1}^{U}(C_{kud}^{KUD}+\delta_{uk}))}{\Gamma(\sum_{u=1}^{U}(C_{kud}^{KUD}+\delta_{uk}+N_{iu}))}\prod_{u=1}^{U}\frac{\Gamma(C_{kud}^{KUD}+\delta_{uk}+N_{iu})}{\Gamma(C_{kud}^{KUD}+\delta_{uk})}$$

式中,C_{kud}^{KUD} 表示文档 d 中 URL u 被分配主题 k 的次数,N_{iu} 表示第 i 个 session 中 URL u 的数量。

2.2 查询词和 URL 相关联的主题模型(CS-SBM)

查询词和 URL 通过搜索引擎紧密地结合在一起,这使得本文研究的问题变得更加复杂。被点击的 URL 是由对应的查询词经过搜索得出的。在网络搜索的情境中,URL 是提交查询词给搜索引擎后

返回的结果。因此,URL 和查询词是相关的。为了获取它们两者之间的关系,这里引入变量 Δqku 来代表“查询词-URL”的多项分布,其先验由 δ 表示。“查询词-URL”多项式通过目前已经被广泛采用的点击图来获得,点击图由搜索查询和被点击的 URL 两部分组成。用于表示全局部分的 CS-SBM 定义为 CS-SBM-G。因为本文重点研究以单个用户为核心的信息能否增强主题模型的表现,所以基于单个用户“查询词-URL”多项分布的 CS-SBM 定义为 CS-SBM-U。

图 2 给出了 CS-SBM 的生成过程。与 SBM 相比,查询词的生成过程不变,两者最主要的区别在于,CS-SBM 在 URL 生成的时候要考虑查询词的影响,对于 session 中每一个查询项 q,首先生成点击事件的二项分布,如果 URL 被点击了,则从参数为 $\boldsymbol{\Omega}_{qz}$ 的 URL 多项分布中采样生成 URL u。

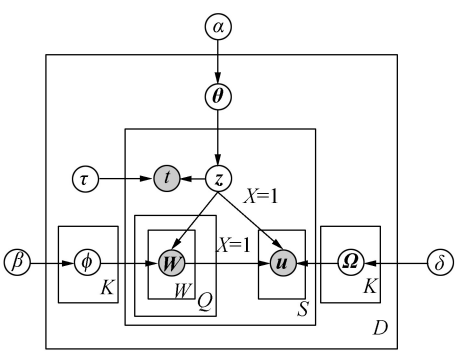


图 2　CS-SBM 网络搜索分析主题模型

Fig.2　CS-SBM web search analysis topic model

模型的完全似然函数为

$$P(\boldsymbol{w},\boldsymbol{u},\boldsymbol{z}\mid\alpha,\beta,\delta)=P(\boldsymbol{u}\mid\boldsymbol{z},\boldsymbol{w},\delta)P(\boldsymbol{w}\mid\boldsymbol{z},\beta)P(\boldsymbol{z}\mid\alpha)$$

在 CS-SBM 中,同样采用了与 LDA 类似的 Gibbs 采样方法。$P(\boldsymbol{z}\mid\alpha)$ 和 $P(\boldsymbol{w}\mid\boldsymbol{z},\beta)$ 都与 SBM 中的相同。两者主要的区别就是给定的主题中的 $\boldsymbol{w}$ 和 $\boldsymbol{u}$ 不再相互独立。$\boldsymbol{u}$ 的生成过程可能受到主题 $\boldsymbol{z}$ 和相关搜索查询 q 的影响。

$$P(\boldsymbol{u}\mid\boldsymbol{z},\boldsymbol{w},\delta)=$$

$$\int\prod_{d=1}^{D}\prod_{i=1}^{N_d}P(\boldsymbol{u}_{di}\mid\Delta_{q_{di}z_{di}})\prod_{q=1}^{Q}\prod_{z=1}^{K}p(\Delta_{qz}\mid\delta)\mathrm{d}\Delta=$$

$$\int\prod_{z=1}^{K}\prod_{q=1}^{Q}\prod_{u=1}^{U}\Delta_{qzu}^{n_{qzu}}\prod_{q=1}^{Q}\prod_{z=1}^{K}\left(\frac{\Gamma(\sum_{u=1}^{U}\delta_{qu})}{\prod_{u=1}^{U}\Gamma(\delta_{qu})}\prod_{u=1}^{U}\Delta_{qzu}^{\delta_{qu}-1}\right)\mathrm{d}\Delta=$$

$$\prod_{q=1}^{Q}\left(\frac{\Gamma(\sum_{u=1}^{U}\delta_{qu})^{K}}{\prod_{u=1}^{U}\Gamma(\delta_{qu})}\right)\times\prod_{z=1}^{K}\prod_{q=1}^{Q}\frac{\prod_{u=1}^{U}\Gamma(n_{qzu}+\delta_{qu})}{\Gamma(\sum_{u=1}^{U}(n_{qzu}+\delta_{qu}))}$$

式中,n_{qzu} 是第 z 个主题的第 q 个查询中的 URL u 的数量。

当 session 中没有 URL 被点击时的条件概率为

$$P(z_i=k\mid X_i=0,\boldsymbol{z}_{-i},\boldsymbol{w},\boldsymbol{u},\alpha,\beta,\delta,\Psi)\propto$$

$$\frac{C_{dk}^{DK}+\alpha_k}{\sum_{k'=1}^{K}(C_{dk'}^{DK}+\alpha_k')}\cdot\frac{\Gamma(\sum_{t=1}^{W}(C_{kwd}^{KWD}+\beta_{wk}))}{\Gamma(\sum_{t=1}^{W}(C_{kwd}^{KWD}+\beta_{wk}+N_{iw}))}\cdot\prod_{w=1}^{W}\frac{\Gamma(C_{kwd}^{KWD}+\beta_{wk}+N_{iw})}{\Gamma(C_{kwd}^{KWD}+\beta_w)}$$

当 session 中有 URL 被点击时的条件概率为

$$P(z_i=k\mid X_i=1,\boldsymbol{z}_{-i},\boldsymbol{w},\boldsymbol{t},\boldsymbol{u},\alpha,\beta,\delta,\Psi)\propto$$

$$\frac{C_{dk}^{DK}+\alpha_k}{\sum_{k'=1}^{K}(C_{dk'}^{DK}+\alpha_k')}\cdot\frac{\Gamma(\sum_{t=1}^{W}(C_{kwd}^{KWD}+\beta_{wk}))}{\Gamma(\sum_{t=1}^{W}(C_{kwd}^{KWD}+\beta_{wk}+N_{iw}))}\cdot\prod_{w=1}^{W}\frac{\Gamma(C_{kwd}^{KWD}+\beta_{wk}+N_{iw})}{\Gamma(C_{kwd}^{KWD}+\beta_w)}\prod_{q\in s_i}\frac{\Gamma(\sum_{u=1}^{U}(C_{qzu}^{QZU}+\delta_{qu}))}{\Gamma(\sum_{u=1}^{U}(C_{qzu}^{QZU}+\delta_{qu}+N_{iu}))}\cdot\prod_{u\leftarrow q}\frac{\Gamma(C_{qzu}^{QZU}+\delta_{qu}+N_{iu})}{\Gamma(C_{qzu}^{QZU}+\delta_u)}$$

式中:C_{qzu}^{QZU} 表示在一个 session 里查询项 q 中的 URL u 被分配主题为 z 的次数;N_{iw} 表示 session i 中 URL 的数量。

2.3　主题在时间上的突发性

网络搜索中另一个常见现象就是时间上的突发性。一个用户更倾向于在一个很短的时间内集中查询一些内容。因此,本文假设每一个用户的查询轨迹都有一个时间上的突发性,这种突发性由与查询相关的时间戳来体现。因为时间粒度是很难设定的,所以本文采用 TOT 中提出的方法,使用连续的 Beta 分布来捕获主题时间上的突发性。通过引入 Beta 分布,使一个主题能够更容易在一个短的时间周期内出现。在这种情况下,本文从整个语料库级别(定义为 X-TG)和基于单一用户级别(定义为 X-TU)出发,刻画一个主题在时间上的变化趋势。

因为主题-周期的多项分布是固定的(对每一个用户而言),所以主题中存在的时间周期将会证明突发性。每天或者每小时都可以观察到突发性现象,这表明去离散化是更合适的。

下面是 SBM 和 CS-SBM 在添加了时间信息以后的模型算法。

the same as the original model

for each session s in d do

choose a topic z:Multinomial(θ_d);

generate timestamps $t \sim \text{Beta}(\tau_z)$ (X-TG) or $t:\text{Beta}(\tau_{zd})$ (X-TU);

the same as the original model

end for

它主要的变化在于,对文档中的每一个 session,从参数为 θ_d 的多项分布中采样一个主题 z,然后根据数据集的不同,从 Beta 分布 $\text{Beta}(\tau_z)$ 和 $\text{Beta}(\tau_{zd})$ 分别生成基于全局的时间戳和基于特定用户的时间戳。

TS-SBM 的 Gibbs 采样与 LDA 方法类似。本文给出了一些简单的推导。首先,模型的完全似然函数为

$$P(\boldsymbol{w},\boldsymbol{u},\boldsymbol{t},\boldsymbol{z}\mid\alpha,\beta,\delta,\tau)=P(\boldsymbol{t}\mid\boldsymbol{z},\tau)P(\boldsymbol{u}\mid\boldsymbol{z},\delta)\cdot P(\boldsymbol{w}\mid\boldsymbol{z},\beta)P(\boldsymbol{z}\mid\alpha)$$

如果 session 中没有 URL 被点击,那么此时的条件概率是:

$$P(z_i=k\mid X_i=0,\boldsymbol{z}_{-i},\boldsymbol{w},\boldsymbol{u},\alpha,\beta,\delta,\tau)\propto \prod_{j=1}^{T}\frac{(1-t_j)^{\tau_{dk1}-1}t_j^{\tau_{dk2}-1}}{B(\tau_{dk1},\tau_{dk2})}\frac{C_{dk}^{DK}+\alpha_k}{\sum_{k'=1}^{K}(C_{dk'}^{DK}+\alpha_{k'})}\cdot \frac{\Gamma(\sum_{t=1}^{W}(C_{kwd}^{KWD}+\beta_{wk}))}{\Gamma(\sum_{t=1}^{W}(C_{kwd}^{KWD}+\beta_{wk}+N_{iw}))}\prod_{w=1}^{W}\frac{\Gamma(C_{kwd}^{KWD}+\beta_{wk}+N_{iw})}{\Gamma(C_{kwd}^{KWD}+\beta_w)}$$

式中:τ_{dk1} 和 τ_{dk2} 是 Beta 分布的超参数。

对于 SBM 模型,如果 session 中有 URL 被点击,此时的条件概率为

$$P(z_i=k\mid X_i=1,\boldsymbol{z}_{-i},\boldsymbol{w},\boldsymbol{u},\alpha,\beta,\delta)\propto \prod_{j=1}^{T}\frac{(1-t_j)^{\tau_{dk1}-1}t_j^{\tau_{dk2}-1}}{B(\tau_{dk1},\tau_{dk2})}\frac{C_{dk}^{DK}+\alpha_k}{\sum_{k'=1}^{K}(C_{dk'}^{DK}+\alpha_{k'})}\cdot \frac{\Gamma(\sum_{w=1}^{W}(C_{kwd}^{KWD}+\beta_{wk}))}{\Gamma(\sum_{w=1}^{W}(C_{kw}^{KW}+\beta_w+N_{iw}))}\prod_{w=1}^{W}\frac{\Gamma(C_{kwd}^{KWD}+\beta_{wk}+N_{iw})}{\Gamma(C_{kwd}^{KWD}+\beta_{wk})}\cdot \frac{\Gamma(\sum_{u=1}^{U}(C_{kud}^{KUD}+\delta_{uk}))}{\Gamma(\sum_{u=1}^{U}(C_{kud}^{KUD}+\delta_{uk}+N_{iu}))}\prod_{u=1}^{U}\frac{\Gamma(C_{kud}^{KUD}+\delta_{uk}+N_{iu})}{\Gamma(C_{kud}^{KUD}+\delta_{uk})}$$

对于 CS-SBM 模型,当 session 中有 URL 被点击时的条件概率为

$$P(z_i=k\mid X_i=1,\boldsymbol{z}_{-i},\boldsymbol{w},\boldsymbol{t},\boldsymbol{u},\alpha,\beta,\delta,\Psi)\propto \prod_{j=1}^{T}\frac{(1-t_j)^{\tau_{dk1}-1}t_j^{\tau_{dk2}-1}}{B(\tau_{dk1},\tau_{dk2})}\frac{C_{dk}^{DK}+\alpha_k}{\sum_{k'=1}^{K}(C_{dk'}^{DK}+\alpha_{k'})}\cdot \frac{\Gamma(\sum_{t=1}^{W}(C_{kwd}^{KWD}+\beta_{wk}))}{\Gamma(\sum_{t=1}^{W}(C_{kwd}^{KWD}+\beta_{wk}+N_{iw}))}\prod_{w=1}^{W}\frac{\Gamma(C_{kwd}^{KWD}+\beta_{wk}+N_{iw})}{\Gamma(C_{kwd}^{KWD}+\beta_w)}\cdot \prod_{q\in s_i}\frac{\Gamma(\sum_{u=1}^{U}(C_{qzu}^{QZU}+\delta_{qu}))}{\Gamma(\sum_{u=1}^{U}(C_{qzu}^{QZU}+\delta_{qu}+N_{iu}))}\prod_{u\leftarrow q}\frac{\Gamma(C_{qzu}^{QZU}+\delta_{qu}+N_{iu})}{\Gamma(C_{qzu}^{QZU}+\delta_u)}$$

时间的参数按照如下方法更新:

$$\tau_{kd1}=\bar{t}_{kd}\left[\frac{\bar{t}_{kd}(1-\bar{t}_{kd})}{s_{kd}^2}-1\right]$$

$$\tau_{kd2}=(1-\bar{t}_{kd})\left[\frac{\bar{t}_{kd}(1-\bar{t}_{kd})}{s_{kd}^2}-1\right]$$

式中,$\bar{t}_{kd}$ 和 s_{kd}^2 表示每一个文档中主题 z 时间上的样本均值和样本偏方差。

3 参数估计

关于超参数 α 和 β 设置问题,一些 LDA 应用采用默认相同值的方法获得了成功,例如由 Griffiths 和 Steyers[17] 提出的 $\alpha=50/k$,$\beta=0.01$,K 是主题的数量。因此,在 LDA 中没有必要去研究超参数。然而,在本文提出的模型中,超参数的设置是至关重要的。因为 LDA 中的 φ 和 Ω 值,也被包括在 SBM 和 CS-SBM 的 β 和 δ 中。

SBM 完全似然 $P(\boldsymbol{w},\boldsymbol{u},\boldsymbol{z}\mid\alpha,\beta,\delta)$ 的计算如下所示:

$$P(\boldsymbol{w},\boldsymbol{u},\boldsymbol{z}\mid\alpha,\beta,\delta)=\prod_d\left(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right)\frac{\prod_{k=1}^{T}\Gamma(m_{dk}+\alpha_k)}{\Gamma(\sum_{k=1}^{K}(m_{dk}+\alpha_k))}\cdot \prod_{d,k}\left(\left(\frac{\Gamma(\sum_{w=1}^{W}\beta_{wk})}{\prod_{w=1}^{W}\Gamma(\beta_{wk})}\right)\frac{\prod_{w=1}^{W}\Gamma(n_{kwd}+\beta_{wk})}{\Gamma(\sum_{w=1}^{W}(n_{kwd}+\beta_{wk}))}\right)\cdot \left(\frac{\Gamma(\sum_{u=1}^{U}\delta_{uk})}{\prod_{u=1}^{U}\Gamma(\delta_{uk})}\right)\frac{\prod_{u=1}^{U}\Gamma(n_{kud}+\delta_{uk})}{\Gamma(\sum_{u=1}^{U}(n_{kud}+\delta_{uk}))}\Bigg)$$

CS-SBM 完全似然 $P(\boldsymbol{w},\boldsymbol{u},\boldsymbol{z}\mid\alpha,\beta,\delta)$ 的计算如下所示:

$$P(\boldsymbol{w},\boldsymbol{u},\boldsymbol{z}\mid\alpha,\beta,\delta)=\prod_d\left(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right)\frac{\prod_{k=1}^{T}\Gamma(m_{dk}+\alpha_k)}{\Gamma(\sum_{k=1}^{K}(m_{dk}+\alpha_k))}\cdot$$

$$\prod_{d,k}\left(\left(\frac{\Gamma\left(\sum_{w=1}^{W}\beta_{wk}\right)}{\prod_{w=1}^{W}\Gamma(\beta_{wk})}\right)\frac{\prod_{w=1}^{W}\Gamma(n_{kwd}+\beta_{wk})}{\Gamma\left(\sum_{w=1}^{W}(n_{kwd}+\beta_{wk})\right)}\right)\cdot$$

$$\prod_{k=1}^{K}\prod_{q=1h}^{Q}\left(\left(\frac{\Gamma\left(\sum_{u=1}^{U}\delta_{qu}\right)}{\prod_{u=1}^{U}\Gamma(\delta_{qu})}\right)\frac{\prod_{u=1}^{U}\Gamma(n_{qku}+\delta_{qu})}{\Gamma\left(\sum_{u=1}^{U}(n_{qku}+\delta_{qu})\right)}\right)$$

SBM-T 完全似然 $P(\boldsymbol{w},\boldsymbol{u},\boldsymbol{z}|\alpha,\beta,\delta)$ 的计算如下所示:

$$P(\boldsymbol{w},\boldsymbol{u},\boldsymbol{z}|\alpha,\beta,\delta)=$$

$$\prod_{d}\left(\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right)\frac{\prod_{k=1}^{T}\Gamma(m_{dk}+\alpha_k)}{\Gamma\left(\sum_{k=1}^{K}(m_{dk}+\alpha_k)\right)}\cdot$$

$$\prod_{d,k}\left(\left(\frac{\Gamma\left(\sum_{w=1}^{W}\beta_{wk}\right)}{\prod_{w=1}^{W}\Gamma(\beta_{wk})}\right)\frac{\prod_{w=1}^{W}\Gamma(n_{kwd}+\beta_{wk})}{\Gamma\left(\sum_{w=1}^{W}(n_{kwd}+\beta_{wk})\right)}\cdot\right.$$

$$\left.\left(\frac{\Gamma\left(\sum_{u=1}^{U}\delta_{uk}\right)}{\prod_{u=1}^{U}\Gamma(\delta_{uk})}\right)\frac{\prod_{u=1}^{U}\Gamma(n_{kud}+\delta_{uk})}{\Gamma\left(\sum_{u=1}^{U}(n_{kud}+\delta_{uk})\right)}\right)\prod_{d,s,i}p(t_{dsi}\mid\tau_{dk_s})$$

CS-SBM-T 完全似然 $P(\boldsymbol{w},\boldsymbol{u},\boldsymbol{z}|\alpha,\beta,\delta,\tau)$ 的计算如下所示:

$$P(\boldsymbol{w},\boldsymbol{u},\boldsymbol{z}|\alpha,\beta,\delta,\tau)=$$

$$\prod_{d}\left(\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right)\frac{\prod_{k=1}^{T}\Gamma(m_{dk}+\alpha_k)}{\Gamma\left(\sum_{k=1}^{K}(m_{dk}+\alpha_k)\right)}\cdot$$

$$\prod_{d,k}\left(\left(\frac{\Gamma\left(\sum_{w=1}^{W}\beta_{wk}\right)}{\prod_{w=1}^{W}\Gamma(\beta_{wk})}\right)\frac{\prod_{w=1}^{W}\Gamma(n_{kwd}+\beta_{wk})}{\Gamma\left(\sum_{w=1}^{W}(n_{kwd}+\beta_{wk})\right)}\right)\cdot$$

$$\prod_{z=1}^{K}\prod_{q=1}^{Q}\left(\left(\frac{\Gamma\left(\sum_{u=1}^{U}\delta_{qu}\right)}{\prod_{u=1}^{U}\Gamma(\delta_{qu})}\right)\frac{\prod_{u=1}^{U}\Gamma(n_{qku}+\delta_{qu})}{\Gamma\left(\sum_{u=1}^{U}(n_{qku}+\delta_{qu})\right)}\right)\cdot$$

$$\prod_{d,s,i}p(t_{dsi}\mid\tau_{dk_s})$$

进行对数似然转换:

$$\alpha_{\cdot}'=\sum_{d,k}(\ln\Gamma(n_{\cdot kd}+\alpha_k)-\ln\Gamma(\alpha_k))+\sum_{d}(\ln\Gamma(\sum_{k}\alpha_k)-\ln\Gamma(\sum_{k}n_{\cdot kd}+\alpha_k))$$

$$\beta_{\cdot k}'=\sum_{d,k,w}(\ln\Gamma(n_{wkd}+\beta_{wk})-\ln\Gamma(\beta_{wk}))+\sum_{d,k}(\ln\Gamma(\sum_{w}\beta_{wk})-\ln\Gamma(\sum_{w}n_{wkd}+\beta_{wk}))$$

对于 SBM:

$$\delta_{\cdot k}'=\sum_{d,k,u}(\ln\Gamma(n_{ukd}+\delta_{uk})-\ln\Gamma(\delta_{uk}))+\sum_{d,k}(\ln\Gamma(\sum_{u}\delta_{uk})-\ln\Gamma(\sum_{u}n_{ukd}+\delta_{uk}))$$

对于 CS-SBM:

$$\delta_{\cdot q}'=\sum_{q,k,u}(\ln\Gamma(n_{qku}+\delta_{qu})-\ln\Gamma(\delta_{qu}))+\sum_{q,k}(\ln\Gamma(\sum_{u}\delta_{qu})-\ln\Gamma(\sum_{u}n_{qku}+\delta_{qu}))$$

上面的每一个公式无论 $\alpha_{\cdot}'$、$\beta_{\cdot k}'$还是 $\delta_{\cdot k}'$、$\delta_{\cdot q}'$都定义了一个向量。本文采用文献[18]中提出的有限空间的 BFGS 方法使它最大化。运行 Gibbs 采样,然后选择 α、β 和 δ 使 $P(\boldsymbol{w},\boldsymbol{u},\boldsymbol{z}|\alpha,\beta,\delta,\tau)$ 完全似然最大化,直到达到稳定状态。重复上述过程,直到 α、β 和 δ 收敛。

4　实验结果分析

4.1　数据集

本文选择的实验数据是搜狗搜索发布的匿名查询日志。它是搜狗在网上公开发布的用户查询日志。该日志包括了用户 2008 年 6 月整月的网络查询记录。日志主要包括 5 部分,即用户匿名 ID、查询词、查询时间、点击的 URL 的排名、点击的 URL。这些数据是按照匿名用户的 ID 顺序依次排列的。本文选取了在一个月内查询日志条目大于 500 条的用户进行建模。首先,将同一用户的搜索查询日志放到一个文档中。然后,用文献[16]提出的方法将搜索查询日志切分成了647 164个 session,用于下一步搜索主题的发现。接下来,根据文献[16]提出的停用词列表过滤掉那些没有意义的查询词。同时,例如 www.sougou.com、www.baidu.com 等主要的搜索引擎和门户网站也要过滤掉[19],因为它们没有提供有用信息。每一个文档的时间戳是由搜索日志上提供的查询时间决定的,并且根据文献[20]提出的 SSTM 模型中用到的方法,将时间按照先后顺序归一化到(0,1)。实验数据中,每一个文档都包括了一些 session,每一个 session 都包括一些查询词、URL(如果有点击事件)和时间戳。

本文选用了两个衡量标准。第 1 个衡量标准是用部分 Held-Out 数据评估模型预测未知数据的能力。第 2 个衡量标准,本文参照了文献[21]提出的方法,即在观察部分用户搜索记录以后,预测剩余查询项的能力。两个衡量标准都选择了困惑度作为评估模型泛化能力的衡量指标。一般而言,模型的困惑度越低,表明泛化能力越强,对模型的拟合程度越高。由于很少有概率模型做有关获取网络搜索查询突发性和时间上的主题突发性研究,很难

找到提出模型的直接竞争对手。所以本文选取了 3 个常用的主题模型作为比较基线,即 LDA、DCMLDA 和 TOT。

4.2 模型困惑度分析

对于第一个衡量标准,困惑度义如下:

$$\text{Perplexity}_{\text{held-out}}(M) = \left(\prod_{d=1}^{D}\prod_{i=1}^{N_d} p(w_i|M)\right)^{\frac{-1}{\sum_{d=1}^{D}(N_d)}}$$

式中,M 是模型通过训练过程学习到的,N_d 是指第 d 个文档中词汇数量。图 3 展示了困惑度的比较结果,从中可以发现这两个提出的模型与 3 个基线模型相比表现出了更好的预测未知数据的能力。因此,把搜索主题数设置为 1 000 时,SBM、CS-SBM、LDA、DCMLDA、TOT 的困惑度分别为 430.347、400.16、1 080.41、995.76、830.23。SBM 和 CS-SBM 自身的困惑度低,并且随着主题数增加困惑度还会进一步降低。实验结果表明,SBM 和 CS-SBM 更适合于从给予的用户搜索历史中预测用户未来网络搜索查询。

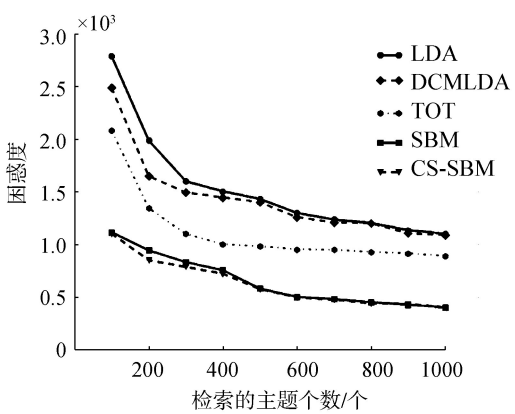


图 3 Held-out 数据的困惑度

Fig.3 Perplexity of held-out data

第 2 种衡量标准的困惑度为

$$\text{Perplexity}_{\text{portion}}(M) = \left(\prod_{d=1}^{D}\prod_{i=P+1}^{N_d} p(w_i|M, W_{a:P})\right)^{\frac{-1}{\sum_{d=1}^{D}(N_d)}}$$

第 2 种衡量标准可以评估提出的模型在观察一部分用户搜索历史记录以后,预测剩余查询项的能力。例如,从用户的查询日志中得到已经观察的查询词 $w_{1:P}$,那么剩余查询项的预测分布为 $P(w|W_{1:P})$。测试数据的困惑度是根据上面的困惑度公式进行计算。举例来说,选择数据集中前 80% 的搜索查询词作为观察训练数据,剩余的 20%作为测试数据。图 4 呈现了部分观察数据的预测困惑度,LDA、DCMLDA、TOT 的困惑度分别是 684.83、671.09、561.26。本文中提出的模型再次取得了显著的优势,其中 SBM 的困惑度是 206.76,而 CS-SBM 达到了 204.87。实验结果表明,SBM 和 CS-SBM 有着更好的预测剩余查询项的能力。

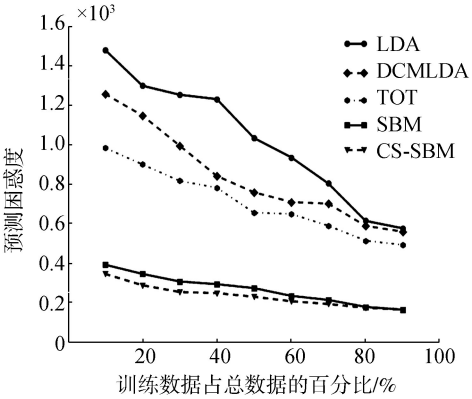


图 4 Observed 数据的困惑度

Fig.4 Perplexity of observed data

4.3 搜索主题发现及分析

发现搜索主题的合理性是判断模型是否成功的一个重要指标。本文中的实验结果证明了提出的模型在发现搜索主题方面表现突出,同时还准确地预测了查询主题在时间上的演化趋势。

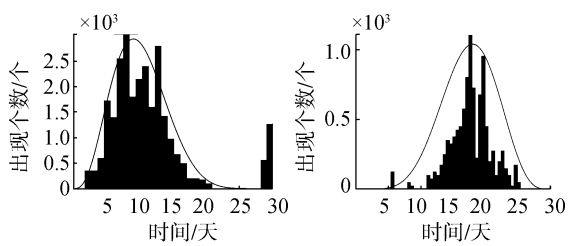
实验中设置主题的数目为 $K=50$。搜索主题是从 Gibbs 采样的 1000 次迭代中单次采样提取的。在表 1~4 中,展示了 4 个由 SBM 和 CS-SBM 分别在语料库级上和单个用户级上发现的搜索主题,并列出了各主题概率值最大的前 5 个词汇及其概率。主题名称是根据该主题下词汇具体的语义信息定义的。图 5~8 中,直方图显示了主题在时间轴上的概率分布,光滑曲线为拟合 Beta 分布的概率密度函数曲线。下面将选取图中的两个搜索主题进行具体分析。

表 1~2 中的第一个主题是“地震”,通过图 5~6 可以发现,本文提出的 SBM 模型在语料库级上和单个用户级上都成功获取了主题时间上的突发性。根据其时间分布来看,由于刚刚发生过汶川地震,因此在 6 月份的前半个月,人们对地震的搜索比较频繁,但随着时间的推移,搜索数量逐渐减少。从语料库级的分析结果看,“汶川、地震、救灾”等高频词汇都与地震相关。对于单个用户,本文从实验结果中选择了一位有大量查询日志并且有地震主题的用户做具体分析描述。结果发现,在这个主题下的词,例如地震、汶川、唐山、四川等词汇出现的概率较高。总体来看,汶川地震引起了人们广泛的关注,对于全局来说,用户更关注地震救援工作;对于

表 2 中的个人用户来说,他们只是关注地震本身,而没有救援相关的查询。

表 3~4 中,最后一个主题是“欧洲杯”,图 7~8 中,CS-SBM 的实验结果显示关于“欧洲杯”这个话题的查询从第十天到第三十天越来越多,这也符合随着欧洲杯的进行人们关注的热情越来越高的现象。从整个语料库得到的结果来看,搜索主题“欧洲杯”主要包括欧洲杯、瑞士、奥地利等词。其中“瑞士”和“奥地利”是本届欧洲杯的举办地,这些词汇都与欧洲杯的主题紧密相关。而对于表 4 中的个人用户,我们仍然从实验结果中选取了与“欧洲杯”主题相关的用户进行分析说明。它包括欧洲杯、西班牙、冠军等关键词,这证明了该用户更关注于欧洲杯的冠军归属问题。

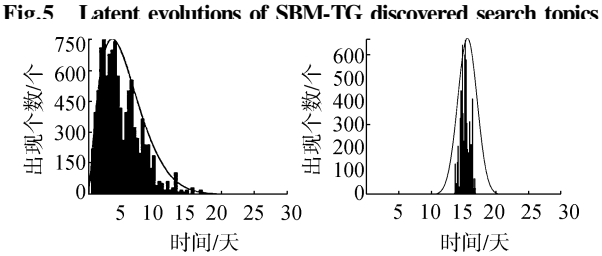
总体而言,发现的搜索主题与实际的情况大体相吻合,而且能较好地反应主题变化的趋势。



(a)主题“地震”随着时间的变化趋势　(b)主题“NBA”随着时间的变化趋势

图 5　CS-SBM 网络搜索分析主题模型

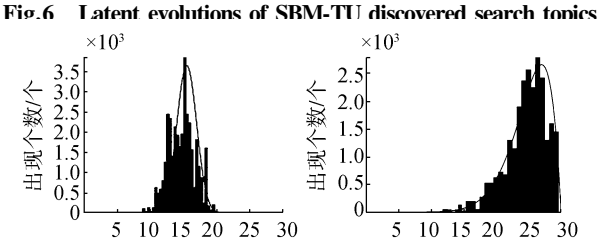
Fig.5　Latent evolutions of SBM-TG discovered search topics



(a)主题“地震”随着时间的变化趋势　(b)主题“NBA”随着时间的变化趋势

图 6　SBM-TU 发现的搜索主题在时间上的演化趋势

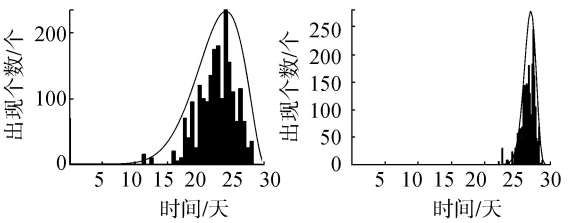
Fig.6　Latent evolutions of SBM-TU discovered search topics



(a)主题“电子产品”随着时间的变化趋势　(b)主题“欧洲杯”随着时间的变化趋势

图 7　CS-SBM-TG 发现的搜索主题在时间上的演化趋势

Fig. 7　Latent evolutions of CS-SBM-TG discovered search topics



(a)主题“电子产品”随着时间的变化趋势　(b)主题“欧洲杯”随着时间的变化趋势

图 8　CS-SBM-TU 发现的搜索主题在时间上的演化趋势

Fig. 8　Latent evolutions of CS-SBM-TU discovered search topics

表 1　SBM-TG 发现的搜索主题分布情况

Table 1　SBM-TG discovered search topics distribution

主题 1		主题 2	
地震		NBA	
搜索词	概率	搜索词	概率
汶川	0.036 28	NBA	0.031 28
地震	0.033 42	季后赛	0.028 83
救灾	0.032 17	西部	0.028 32
武警	0.027 74	湖人	0.023 36
哄抢	0.027 74	冠军	0.020 92

表 2　SBM-TU 发现的搜索主题分布情况

Table 2　SBM-TU discovered search topics distribution

主题 1		主题 2	
地震		NBA	
搜索词	概率	搜索词	概率
地震	0.047 62	总决赛	0.042 93
汶川	0.043 58	NBA	0.039 12
唐山	0.043 02	湖人	0.036 41
四川	0.032 75	凯尔特人	0.033 79
地壳	0.030 28	斯台普斯	0.023 11

表 3　CS-SBM-TG 发现的搜索主题分布情况

Table 3　CS-SBM-TG discovered search topics distribution

主题 3		电子产品	
主题 4		欧洲杯	
搜索词	概率	搜索词	概率
手机	0.050 04	欧洲杯	0.034 87
内存卡	0.048 18	瑞士	0.029 83
耳机	0.045 57	奥地利	0.026 49
笔记本	0.043 65	足球	0.025 37
MP3	0.038 76	决赛	0.024 30

表 4 CS-SBM-TG 发现的搜索主题分布情况

Table 4 CS-SBM-TG discovered search topics distribution

主题 3		电子产品	
主题 4		欧洲杯	
搜索词	概率	搜索词	概率
诺基亚	0.038 82	欧洲杯	0.032 03
手机	0.039 12	西班牙	0.030 64
内存	0.036 41	德国	0.029 91
相机	0.033 79	冠军	0.029 01
HTC	0.023 11	瑞士	0.025 06

5 结束语

本文提出了两个主题模型 SBM 和 CS-SBM,从全局和基于特定用户来建模网络搜索分析。SBM 主要是用于获取网络查询的突发现象。CS-SBM 主要添加了查询词和 URL 之间的关系,获取了主题突发性。为了使 SBM 和 CS-SBM 可以获取时间上的突发性,本文采用 Beta 分布拟合主题在时间上的变化的策略。本文还设计了一系列的实验验证了提出模型拥有较好的泛化能力、主题发现能力和反应主题时间上突发性的能力。本文的贡献主要有三个方面:第一,研究了搜索引擎用户行为分析中突发性现象;第二,提出了两种新型的模型用来捕获网络搜索中各个方面的突发性;第三,通过大量的实验验证了模型的有效性。下一步工作中,将把这些模型运用到团购广告投放。通过发现用户的搜索主题,然后将同一主题下的用户按照社团发现的规则进行分类,并进行广告投放。

参考文献:

[1] SUNEHAG P. Using two-stage conditional word frequency models to model word burstiness and motivating TF-IDF[J]. Journal of machine learning reasearch, 2017, 2: 8.

[2] ELKAN C. Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution [C]//Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh. Pennsylvania, USA, 2006: 289-296.

[3] DOYLE G, ELKAN C. Accounting for burstiness in topic models[C]//Proceedings of the 26th Annual International Conference on Machine Learning Montreal. QC, Canada, 2009: 281-288.

[4] XUE G R, ZENG H J, CHEN Z, et al. Optimizing web search using web click-through data [C]//Proceedings of the thirteenth ACM international conference on Information and Knowledge Management. Washington, USA, 2004: 118-126.

[5] GUO F, LIU C, WANG Y M. Efficient multiple-click models in web search[C]//Proceedings of the Second ACM International Conference on Web Search and Data Mining. Barcelona, Spain, 2009: 124-131.

[6] 张宇, 宋巍, 刘挺, 等. 基于 URL 主题的查询分类方法[J]. 计算机研究与发展, 2012, 49(6): 1298-1305.
ZHANG Yu, SONG Wei, LIU Ting, et al. Query classification based on url topic [J]. Journal of computer research and development, 2012, 49(6): 1298-1305.

[7] MADSEN R E, KAUCHAK D, ELKAN C. Modeling word burstiness using the dirichlet distribution[C]//Proceedings of the 22nd international conference on Machine iearning. Bonn, Germany, 2005: 545-552.

[8] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3(1): 993-1022.

[9] 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. Philadelphia, USA, 2006: 424-433.

[10] 徐戈, 王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报, 2011, 34(8): 1423-1436.
XU Ge, WANG Houfeng. The development of topic model in natural language processing [J]. Chinese journal of computers, 2011, 34(8): 1423-1436.

[11] 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.
ZHANG Chenyi, SUN Jianling, DING Yiqun. Topic mining for microblog based on mb-lda model[J]. Journal of computer research and development, 2011, 48 (10): 1795-1802.

[12] 刘少鹏, 印鉴, 欧阳佳, 等. 基于 MB-HDP 模型的微博主题挖掘[J]. 计算机学报, 2015, 38(7): 1408-1419.
LIU Shaopeng, YIN Jian, OUYANG Jia, et al. Topic mining from microblogs based on MB-HDP model [J]. Chinese Journal of Computers, 2015, 38(7): 1408-1419.

[13] JIANG D, TONG Y, SONG Y. Cross-lingual topic discovery from multilingual search engine query log[J]. ACM transactions on information systems (TOIS), 2016, 35(2): 9.

[14] JIANG D, LEUNG K W T, NG W. Query intent mining with multiple dimensions of web search data[J]. World wide web, 2016, 19(3): 475.

[15] JIANG D, YANG L. Query intent inference via search engine log [J]. Knowledge and information systems, 2016, 49(2): 661-685.

[16] HUANG J, EFTHIMIADIS E N. Analyzing and evaluating query reformulation strategies in web search logs[C]//

Proceedings of the 18th ACM Conference on Information and Knowledge Management. Hong Kong, China, 2009: 77-86.

[17] GRIFFITHS T L, STEYVERS M. Finding scientific topics [J]. Proceedings of the national academy of sciences, 2004, 101(1): 5228-5235.

[18] ZHU C, BYRD R H, LU P, et al. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization [J]. ACM transactions on mathematical software (TOMS), 1997, 23(4): 550-560.

[19] MANNING C D, RAGHAVAN P, SCHÜTZE H. Introduction to information retrieval [M]. Cambridge: Cambridge University Press, 2008: 1-16.

[20] JIANG D, NG W. Mining web search topics with diverse spatiotemporal patterns [C]//Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 2013: 881-884.

[21] LI W, MCCALLUM A. Pachinko allocation: DAG-structured mixture models of topic correlations [C]// Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA, 2006: 577-584.

作者简介:

张森,男,1992 年生,硕士研究生,主要研究方向为信息检索、自然语言处理。

张晨,男,1988 年生,副教授,博士,主要研究方向为众包、数据分析与数据挖掘、机器学习。在 TKD、VLDB、SIGMOD、ICDE 等国内外重要期刊和顶级学术会议上发表论文 10 余篇。

林培光,男,1978 年生,副教授,博士,主要研究方向为信息检索、海量数据处理和集成。主持教育部课题 2 项、山东省自然科学基金项目 1 项、济南市科技局自主创新计划 1 项和青年科技明星计划 1 项,另外参与国家自然科学基金以及省部级课题多项。发表学术论文 30 余篇,被 SCI 检索 3 篇,EI 检索 30 余篇。