Journal of Computer Applications
计算机应用,2016,36(8):2103 -2108
ISSN 1001-9081
CODEN JYIIDU
2016-08-10
http://www.joca.cn

文章编号:1001-9081(2016)08-2103-06
doi:10.11772/j.issn.1001-9081.2016.08.2103

基于概率主题模型的景点知识挖掘及其可视化

徐 洁,范玉顺*,白 冰
(清华大学 自动化系,北京 100084)
(* 通信作者电子邮箱 fanyus@tsinghua.edu.cn)

摘 要:针对旅游文本噪声多、景点多且展示不直观的问题,提出一种基于概率主题模型的景点-主题模型。模型假设同一篇文档涉及多个具有相关关系的景点,引入"全局景点"过滤噪声语义,并利用 Gibbs 采样算法估计最大似然函数的参数,获取目的地景点的主题分布。实验通过对景点主题特征进行聚类,评估聚类效果从而间接评价模型训练效果,并定性分析"全局景点"对模型的作用。实验结果表明,该模型对旅游文本的建模效果优于基准算法 TF-IDF 与隐含狄利克雷分布(LDA),且"全局景点"的引入对建模效果有明显的改善作用。最后通过景点关联图的方式对实验结果进行可视化展示。

关键词:概率主题模型;旅游文本;噪声;Gibbs 采样;可视化
中图分类号:TP391 **文献标志码**:A

Knowledge mining and visualizing for scenic spots with probabilistic topic model

XU Jie, FAN Yushun*, BAI Bing
(*Department of Automation, Tsinghua University, Beijing* 100084, *China*)

Abstract: Since the tourism text for destinations contains semantic noise and different scenic spots, which can not be displayed intuitively, a new scenic spots-topic model based on the probabilistic topic model was proposed. The model assumed that one document included several scenic spots with correlation, and a special scenic spot named "global scenic spot" was introduced to filter the semantic noise. Then Gibbs sampling algorithm was employed to learn the maximum a posteriori estimates of the model and get a topic distribution vector for each scenic spot. A clustering experiment was conducted to indirectly evaluate the effects of the model and analyze the impact of "global scenic spot" on the model. The result shows that the proposed model has better effect than baseline model such as TF-IDF (Term Frequency-Inverse Document Frequency) and Latent Dirichlet Allocation (LDA), and the "global scenic spot" can improve the modeling effect significantly. Finally, scenic spots association graph was employed to display the result visually.

Key words: probabilistic topic model; tourism text; noise; Gibbs sampling; visualization

0 引言

Web 2.0 技术及在线旅游代理(Online Travel Agent,OTA)的飞速发展导致旅游数据爆炸性增长。如何有效地从海量旅游数据中挖掘出有用的信息并以直观方式进行展示成为当前的迫切需求。

近年来,对旅游数据的挖掘工作多集中于对旅游照片及相应元数据、标签的研究,如文献[1-2]等利用 Flickr 网站用户上传的海量旅游照片及标签信息对景点进行聚类分析;文献[3]从 Panoramio[4] 网站采集照片聚成地标,并为每个地标找到最具代表性的照片与标签等。随着文本数据挖掘的快速发展,旅游文本数据相关的研究工作方兴未艾,相关研究工作通常可分为两类,即词频分析法和主题挖掘法。词频分析法利用词频统计结果进行文本分析,如文献[5]采用词频分析法刻画目的地旅游感知形象,文献[6]利用内容分析法(Content Analysis,CA)获取目的地语义网络分析图等。该类方法将单词视为单纯的文本符号,无法识别其中的语义信息。主题挖掘法采用或扩展隐含狄利克雷分布(Latent Dirichlet Allocation,LDA)[7],利用潜在主题识别语义信息,从而提高文本数据挖掘的效果,如文献[8-9]提出一种地点-主题(Location-Topic,LT)模型用于挖掘目的地的主题分布信息,以文本标签形式生成目的地概述。然而旅游目的地由景点组成,目的地特征由景点的类型与特征构成,同一文本可能涉及不同景点,这些景点间具有地理位置、主题等关联关系(如图 1 方框标注),上述方法对地点划分粒度较大且没有考虑景点关联关系。另外,旅游文本中常包含时间、门票、电话等与景点主题特征相关性不大的信息,即"噪声语义"(如图 1 椭圆标注),多数主题挖掘方法没有考虑噪声语义消除问题,LT 模型虽可利用"全局主题"过滤噪声语义,但模型复杂度较高。为充分利用景点间的关联关系,有效消除噪声语义,本文提出一种简单的基于概率主题模型的景点-主题模型(Scenic spots-Topic Model with Global Scenic spot,GS-STM)以无监督地从旅游文本中挖掘景点主题分布信息,并以景点关联图的形式展示旅游目的地的景点类型与主题特征。

收稿日期:2016-03-01;修回日期:2016-05-11。 基金项目:高校博士学科点专项科研基金资助项目(20120002110034)。
作者简介:徐洁(1990—),女,山东烟台人,硕士研究生,主要研究方向:业务流程管理、服务推荐、大数据; 范玉顺(1962—),男,江苏扬州人,教授,博士生导师,博士,主要研究方向:企业建模与优化分析、企业经营过程重组、工作流管理; 白冰(1990—),男,北京人,博士研究生,主要研究方向:服务计算、服务推荐、大数据。

1 相关工作

1.1 概率主题模型

概率主题模型是针对文本中隐含主题的一种建模方法。由于不需要对文档进行人工标注及可自动分析主题的特点，概率主题模型已被成功运用到多种文本挖掘问题中。它的主要思想是认为文档是若干主题的混合分布，而每个主题又是一个关于单词的概率分布。

自提出以来，概率主题模型经历了潜在语义分析(Latent Sematic Analysis，LSA)[11]、概率潜在语义分析(probabilistic Latent Sematic Analysis，pLSA)[12]、LDA、分层狄利克雷过程(Hierarchical Dirichlet Process，HDP)[13]等阶段的发展，目前以 LDA 应用最为广泛。LDA 是一种生成模型：对于新文档中的每个单词，通过主题的分布随机得到文档的某个主题，然后通过该主题中单词的分布随机得到一个单词。

如图 2 所示，LDA 是典型的有向概率图模型[14]，超参数 α 反映了文档集合中隐含主题间的相对强弱，超参数 β 刻画所有隐含主题自身的概率分布。

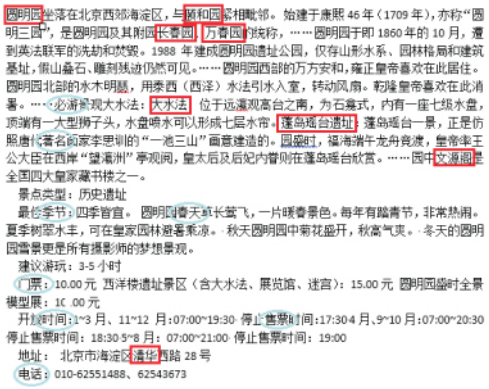
圆明园坐落在北京西郊海淀区，与颐和园紧相毗邻。始建于康熙 46 年（1709 年），亦称“圆明三园”，是圆明园及其附园长春园、万春园的统称，……圆明园于即 1860 年的 10 月，遭到英法联军的洗劫和焚毁。1988 年建成圆明园遗址公园，仅存山形水系、园林格局和建筑基址，假山蟠石、雕刻残迹仍然可见。……圆明园西部的万方安和，雍正皇帝喜欢在此居住。圆明园北部的水木明瑟，用泰西（西洋）水法引水入室，转动风扇。乾隆皇帝喜欢在此消暑。……必游景观大水法：大水法，位于远瀛观高台之南，为石幕式，内有一座七级水盘，顶端有一大型狮子头，水盘喷水可以形成七层水帘。蓬岛瑶台遗址：蓬岛瑶台一景，正是仿照唐代著名画家李思训的“一池三山”画意建造的。园盛时，福海湍午龙舟竞渡，皇帝率王公大臣在西岸“望瀛洲”亭观阅，皇太后及后妃内眷则在蓬岛瑶台欣赏。……园中文源阁是全国四大皇家藏书楼之一。
景点类型：历史遗址
最佳季节：四季皆宜。圆明园春天草长莺飞，一片暖春景色。每年有踏青节，非常热闹。夏季树荫水丰，可在皇家园林避暑乘凉。秋天圆明园中菊花盛开，秋高气爽。冬天的圆明园雪景更是所有摄影师的梦想景观。
建议游玩：3-5 小时
门票：10.00 元 西洋楼遗址景区（含大水法、展览馆、迷宫）：15.00 元 圆明园盛时全景模型展：10.00 元
开放时间：1~3 月、11~12 月：07:00~19:30 停止售票时间：17:30 4 月、9~10 月：07:00~20:30 停止售票时间：18:30 5~8 月：07:00~21:00 停止售票时间：19:00
地址：北京市海淀区清华西路 28 号
电话：010-62551488、62543673

图 1 景点描述文本示例

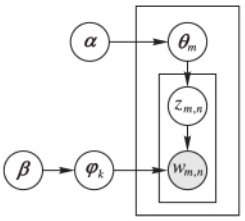

图 2 LDA 概率图模型

LDA 主题模型生成文本的过程如下：

1) 对于每个主题 k，根据 Dirichlet 分布 Dirichlet(β) 得到主题上单词的多项式分布 φ_k。

2) 对于文档 m，根据 Dirichlet 分布 Dirichlet(α) 得到文档中主题的多项式分布 θ_m。

3) 对于文档 m 中的每个单词：

a) 从多项式分布 Multinomial(θ_m) 中随机选择第 n 个单词的主题 $z_{m,n}$；

b) 从多项式分布 Multinomial($\varphi_{z_{m,n}}$) 中随机选择一个单词 $w_{m,n}$。

1.2 可视化模型

可视化技术因具备直观、易理解的特点被广泛应用于各个领域，它用二维或三维图像的方式展现数据，便于发现数据的分布特征及其中蕴含的模式特征[15]。图是一种典型的数据结构，很多数据均可通过图来表达。

力导向模型是一种基于物理方法的可视化模型。该模型将图类比为一个虚拟的物理系统，图的各个节点看作系统中的质点，节点之间的边看作节点间的相互作用力(同时包括引力和斥力)。模型将胡克定律作为基本算法，每次迭代，节点向所受合力的方向移动，经足够的迭代后，系统达到平衡，此时系统中的能量达到最小，图的可视化显示最为美观。

力导向算法基本过程如下：

1) 随机分布初始节点位置；

2) 分别计算局部区域内边的引力和斥力所产生的两端节点的单位位移；

3) 累加步骤 2) 得到的所有节点的单位位移；

4) 重复步骤 2)、3) 直到达到理想效果。

2 景点-主题模型

本章介绍 GS-STM，并采用 Gibbs 采样[16]算法对模型进行求解，从而获得景点与主题、主题与单词之间的概率分布。

2.1 模型介绍

假设数据集中的文档以词袋模型(bag of words model)表示。GS-STM 中，每篇文档的单词均与潜在变量——景点 s 和主题 z 相联系。模型中，已知文档 d 中景点集合 s_d 和词典 v。如图 3 概率图模型所示，对于文档 d 中单词的生成，首先从景点集合 $s_d \cup \{gs\}$ 中随机选取一个景点 s，然后根据主题在该景点上的分布 θ_s 选择一个主题 z，最后根据单词在该主题上的分布 φ_z 生成文本 d 中的一个单词 w。

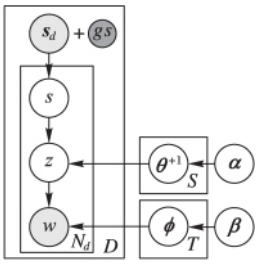

图 3 景点-主题模型的概率图模型

模型中，gs 表示“全局景点(Global Scenic Spot)”，定义为文档集中所有文档添加的共同景点。若全局景点与某主题相关性远大于其他景点，则该主题存在于多数文档中，并不能代表景点特征，设定该主题为“无效主题(invalid topic)”；否则为“有效主题(valid topic)”。因而，全局景点可用于过滤文档中与景点特征无关的“噪声语义”。

GS-STM 具体生成过程如下：

1) 对于每一个景点 $s \in \{s_1, s_2, \cdots, s_S, gs\}$，选择 T 维随机变量 $\theta_s \sim \text{Dirichlet}(\alpha)$。

2) 对于每一个主题 $t \in \{1, 2, \cdots, T\}$，选择 V 维随机变量 $\varphi_t \sim \text{Dirichlet}(\beta)$。

3) 对于每一个文档 $d \in \{1, 2, \cdots, D\}$，给定景点集合 s_d，对于其中的每一个单词 $w_i (i \in \{1, 2, \cdots, N_d\})$：

a) 选择一个景点 $s_i \sim \text{Uniform}(s_d \cup \{gs\})$；

b) 根据 s_i，选择主题 $z_i \sim \text{Multinomial}(\theta_{s_i})$；

c) 根据 z_i，选择单词 $w_i \sim \text{Multinomial}(\varphi_{z_i})$。

2.2 参数估计

模型构建过程中，需要对模型参数进行最大似然估计(Maximum Likelihood Estimation，MLE)。本文采用 Gibbs 采样算法进行参数估计，算法中符号定义如表 1 所示。

GS-STM 的生成过程分为三个部分：

1) $s_d \cup \{gs\} \rightarrow s$。文档 d 第 i 个单词所属景点 $s_i \sim \text{Uniform}(s_d \cup \{gs\})$，且每篇文档中景点的生成过程相互独

立,故:

$$p(\boldsymbol{s}) = \prod_{d=1}^{D}\left[\frac{1}{\|\boldsymbol{s}_d\|+1}\right]^{N_d} \tag{1}$$

其中 $\boldsymbol{s}_{1\times N}$ 为每个单词所属景点。

2) $\boldsymbol{\alpha}\rightarrow\boldsymbol{\theta}_s\rightarrow\boldsymbol{z}_s$。由于 $z_i \sim \text{Multinomial}(\boldsymbol{\theta}_s)$, $\boldsymbol{\theta}_s \sim \text{Dirichlet}(\boldsymbol{\alpha})$, 且不同景点中主题生成过程相互独立,故在已知景点的生成概率条件下,主题的生成概率:

$$p(\boldsymbol{z}\mid\boldsymbol{s},\boldsymbol{\alpha}) = \prod_{k=1}^{S+1}p(\boldsymbol{z}_k\mid\boldsymbol{s}_k,\boldsymbol{\alpha}) = \prod_{k=1}^{S+1}\frac{\Delta(\boldsymbol{n}_k+\boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})} \tag{2}$$

其中: $\boldsymbol{n}_k = (N_k^{(1)}, N_k^{(2)}, \cdots, N_k^{(T)})$, $N_k^{(t)}$ 表示第 k 个景点中第 t 个主题产生的单词个数; $\boldsymbol{z}_{1\times N}$ 表示每个单词所属主题; $\Delta(\boldsymbol{\alpha})$ ($\boldsymbol{\alpha}=\{\alpha_1,\alpha_2,\cdots,\alpha_n\}$) 为归一化因子 $\text{Dirichlet}(\boldsymbol{\alpha})$, 即:

$$\Delta(\boldsymbol{\alpha}) = \int\prod_{i=1}^{n}p_i^{\alpha_i-1}\mathrm{d}p$$

3) $\boldsymbol{\beta}\rightarrow\boldsymbol{\varphi}_z\rightarrow\boldsymbol{w}_z$。因 $w_i \sim \text{Multinomial}(\boldsymbol{\varphi}_z)$, $\boldsymbol{\varphi}_z \sim \text{Dirichlet}(\boldsymbol{\beta})$, 且 T 个主题的单词生成过程相互独立,故语料库中单词的生成概率如下:

$$p(\boldsymbol{w}\mid\boldsymbol{z},\boldsymbol{\beta}) = \prod_{t=1}^{T}p(\boldsymbol{w}_t\mid\boldsymbol{z}_t,\boldsymbol{\beta}) = \prod_{t=1}^{T}\frac{\Delta(\boldsymbol{n}_t+\boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})} \tag{3}$$

其中: $\boldsymbol{n}_t = (N_t^{(1)}, N_t^{(2)}, \cdots, N_t^{(V)})$, $N_t^{(v)}$ 表示第 t 个主题中单词 v 的个数; $\boldsymbol{w}_{1\times N}$ 表示语料库中所有单词。

结合式(1)~(3)可得:

$$p(\boldsymbol{w},\boldsymbol{z},\boldsymbol{s}\mid\boldsymbol{\alpha},\boldsymbol{\beta})(1) = p(\boldsymbol{w}\mid\boldsymbol{z},\boldsymbol{\beta})\,p(\boldsymbol{z}\mid\boldsymbol{s},\boldsymbol{\alpha})\,p(\boldsymbol{s}) = \prod_{d=1}^{D}\left[\frac{1}{\|\boldsymbol{s}_d\|+1}\right]^{N_d}\prod_{k=1}^{S+1}\frac{\Delta(\boldsymbol{n}_k+\boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})}\prod_{t=1}^{T}\frac{\Delta(\boldsymbol{n}_t+\boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})}$$

因此,Gibbs 循环采样中,更新公式为:

$$P(z_i=t,s_i=k\mid\boldsymbol{z}_{-i},\boldsymbol{s}_{-i},\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta}) \propto P(z_i=t,s_i=k,w_i=v\mid\boldsymbol{z}_{-i},\boldsymbol{s}_{-i},\boldsymbol{w}_{-i},\boldsymbol{\alpha},\boldsymbol{\beta}) \propto \frac{N_{k,-i}^{(t)}+\alpha_t}{\sum_{t'}N_{k,-i}^{(t')}+T\alpha_t}\cdot\frac{N_{t,-i}^{(v)}+\beta_v}{\sum_{v'}N_{t,-i}^{(v')}+V\beta_v}$$

取参数在后验分布下的期望值作为参数估计值,则有:

$$\theta_{kt} = \frac{N_{k,-i}^{(t)}+\alpha_t}{\sum_{t'}N_{k,-i}^{(t')}+T\alpha_t} \tag{4}$$

$$\varphi_{tv} = \frac{N_{t,-i}^{(v)}+\beta_v}{\sum_{v'}N_{t,-i}^{(v')}+V\beta_v} \tag{5}$$

表 1　Gibbs 采样公式中的字符定义

参数	说明
$z_i=j$	第 i 个单词属于主题 j
$s_i=k$	第 i 个单词属于景点 k
$w_i=v$	第 i 个单词为词典中第 v 个单词
$\boldsymbol{z}_{-i}$	除第 i 个单词外,文档中其余单词与主题的对应关系
$\boldsymbol{s}_{-i}$	除第 i 个单词外,文档中其余单词与景点的对应关系
$\boldsymbol{w}_{-i}$	除第 i 个单词外,文档中其余单词与词典中单词的对应关系
$N_{k,-i}^{(t)}$	除第 i 个单词外,第 k 个景点中第 t 个主题产生的单词个数
$N_{t,-i}^{(v)}$	除第 i 个单词外,第 t 个主题中单词 v 的个数

Gibbs 采样算法的基本过程为:

1) 随机初始化,对语料库中每篇文档中的词 w, 随机为景点 s 和主题 z 赋值;

2) 重新扫描语料库,对每个词 w, 按照 Gibbs 采样更新公式重新采样其所属主题和景点,并更新语料库;

3) 重复以上采样过程直到 Gibbs 采样收敛;

4) 按照式(4)、(5)计算景点-主题关系矩阵 $\boldsymbol{\theta}$ 和主题-单词关系矩阵 $\boldsymbol{\varphi}$。

3　实验设计与结果分析

3.1　数据说明

本文使用百度旅游网站[17]关于北京地区的 1 943 篇景点描述文本作为语料库,其中含有有效景点数 1 932 个。数据集经过分词、去除停用词、TF-IDF (Term Frequency-Inverse Document Frequency) 权值[18]过滤处理后,得到单词总数为 89 804,经去重后得到含有 6 404 个单词的词典。

3.2　评价指标

对于无分类标签的数据集,本文利用 K-means 算法对训练结果进行聚类,并采用 DBI (Davies-Bouldin Index) 指数[19]度量聚类性能。DBI 指数计算公式如下:

$$R_{\text{DBI}} = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{avg(C_i)+avg(C_j)}{d_{\text{cen}}(C_i,C_j)}\right)$$

其中:

1) $avg(C) = \dfrac{2}{|C|(|C|-1)}\sum_{1\leqslant i<j\leqslant|C|}dist(x_i,x_j)$, 表示簇 C 内样本间的平均距离;

2) $d_{\text{cen}}(C_i,C_j) = dist(\mu_i,\mu_j)$, 表示簇中心点 μ_i 与 μ_j 间的距离。

显然,相同聚类个数下,DBI 值越小,聚类效果越好,即簇内相似度越大,簇间相似度越小。本文中,DBI 值反映了模型以主题分布表征景点(LDA 中假设一篇文档为一个景点)的有效性——DBI 值越小,模型越能捕捉景点在主题上的相关性与差异性。

3.3　实验结果

本文实验分为两部分: 一是定量分析,选择词频统计法(以 TF-IDF 模型为例)、主题挖掘法(以 LDA 为例)作为基准算法,分别对加入全局景点的 GS-STM、不加入全局景点的 STM、LDA 和 TF-IDF 四种模型进行聚类实验,并评测 DBI 指数; 二是定性分析,以直观形式展示训练结果及全局景点对减少噪声语义、识别"无效主题"的作用。实验设定超参数 $\alpha = 50/T$, $\beta = 0.01$, Gibbs 采样迭代3 000次。

3.3.1　定量分析

利用 K-Means 算法进行聚类实验,比较 GS-STM、STM、LDA、TF-IDF 四种模型在不同主题数下的聚类效果,结果如图 4 所示。

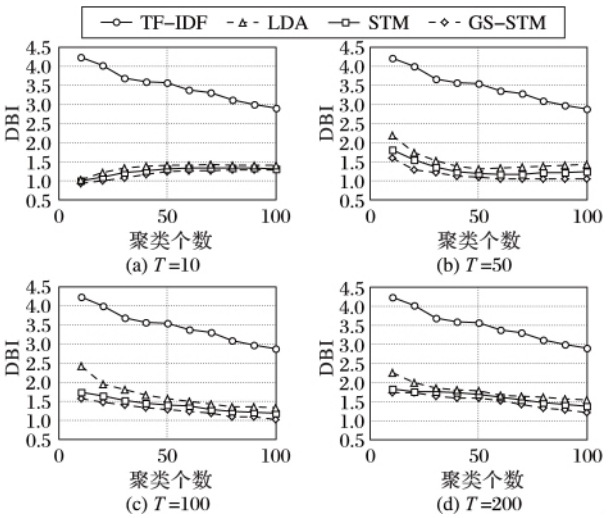

图 4　四种方法聚类效果比较

从图4可以看出，三种基于概率主题模型的方法——GS-STM、STM、LDA的DBI值均低于TF-IDF，说明基于概率主题模型的方法能够有效利用文档中的语义信息；不同主题数下，GS-STM、STM的DBI值均高于LDA，说明考虑文档中多个景点对提升模型建模效果是有效的；而GS-STM的DBI值总是高于STM，说明全局景点的引入能明显改善模型建模效果。

3.3.2 定性分析

分别采用GS-STM、STM对旅游文本进行训练，结果显示当主题数为80时，训练效果最好。设定主题数为80，STM得到80个主题，而GS-STM方法得到68个有效主题、12个无效主题。

表2～4分别列出了GS-STM训练得到的5个“有效主题”“无效主题”及STM得到的5个主题，每个主题显示5个最相关单词和5个最相关景点。

表2中，“有效主题”对应特定景点类型，如“运动”“购物”“电影”等主题。具有地理相关或主题相关关系的景点被列入同一主题，如Topic#38中，“鸟巢”“奥林匹克体育中心”等体育场馆被列入同一主题，同时与之地理邻近且主题相关的“奥林匹克森林公园”等也被列入同一主题。

表3中，从主题最相关单词角度看，各主题中单词多为“噪声语义”，如Topic#32中，“门票”“电话”“世界”等在多数景点介绍文档中均有出现；从主题最相关景点角度看，各主题中全局景点概率最大，且远高于其他景点，因而利用全局景点将该类主题设为“无效景点”是合理有效的。

表4中，Topic#8Ⅱ和Topic#19Ⅱ分别对应表5中的“购物”主题和“电影”主题，即Topic#4和Topic#75，对比主题相关单词构成可见，Topic#8Ⅱ和Topic#19Ⅱ中的“电话”“核心”等单词并不能准确描述并区分主题，GS-STM通过全局景点将这些词归属到“无效主题”(Topic#17，Topic#32)中，从而有效减少主题描述单词中的噪声语义，使得主题描述单词更准确有效；Topic#55Ⅱ、Topic#67Ⅱ、Topic#78Ⅱ所示主题中的单词并不能准确描述相关景点，实为“无效主题”，STM不能识别。

表2 GS-STM模型，部分有效主题单词分布及相关景点

Topic #4		Topic#15		Topic#47		Topic#38		Topic#75	
单词	概率/%	单词	概率/%	单词	概率/%	单词	概率/%	单词	概率/%
购物	6.230	长城	30.830	皇家	5.820	奥林匹克	8.684	电影	10.380
商业	5.340	八达岭	10.975	乾隆	5.335	奥运会	4.581	影视	8.136
美食	4.450	居庸关	3.322	皮影	3.590	奥运	4.295	拍摄	3.648
大厦	4.228	古北口	2.528	遗址	2.911	体育	3.913	影院	2.806
商业街	3.783	龙山	2.239	清华	2.232	比赛	3.818	电影院	2.526
Topic #4		Topic#15		Topic#47		Topic#38		Topic#75	
景点	概率/%	景点	概率/%	景点	概率/%	景点	概率/%	景点	概率/%
王府井	6.932	八达岭长城	13.626	圆明园	13.177	鸟巢	10.977	八一电影制片厂	2.618
西单	1.352	司马台长城	4.849	颐和园	9.591	奥林匹克森林公园	4.180	飞腾影视城	2.403
瑞蚨祥	1.352	金山岭长城	2.555	万寿山	2.236	北京奥林匹克公园	2.401	中影基地旅游城	1.864
簋街	1.303	慕田峪长城	2.516	清华园	0.718	奥林匹克体育中心	1.944	中国电影博物馆	1.541
峨嵋酒家	1.009	居庸关长城	2.397	七孔桥	0.488	北京科技大学体育馆	1.671	明皇蜡像宫	1.541

表3 GS-STM模型，部分无效主题单词分布及相关景点

Topic #17		Topic#26		Topic#32		Topic#40		Topic#77	
单词	概率/%	单词	概率/%	单词	概率/%	单词	概率/%	单词	概率/%
京城	6.643	文化	8.283	门票	14.843	历史	12.526	场所	5.943
建筑面积	5.558	大型	3.931	电话	10.916	四季	5.044	项目	4.458
观光	3.389	活动	3.851	世界	5.350	印象	4.456	专业	2.839
北京地区	3.164	主题	3.329	免费	4.145	展示	4.120	保存	2.653
典型	1.853	设计	3.069	风格	3.229	全天	3.825	生活	2.627
Topic #17		Topic#26		Topic#32		Topic#40		Topic#77	
景点	概率/%	景点	概率/%	景点	概率/%	景点	概率/%	景点	概率/%
全局	49.585	全局	72.766	全局	66.501	全局	43.701	全局	65.531
潘家园	0.436	朝阳公园	0.206	天安门	0.465	长安街	3.254	牡丹园	0.481

4 景点可视化

利用Gephi复杂网络分析软件，绘制北京地区景点关联图，如图5所示。景点关联图由节点和节点之间的连线构成。节点代表景点，其中：

1) 节点大小表示景点与主题的相关度。景点与主题相关性越大，$Size(k)$越大，公式定义如下：

$$Size(k) \propto \max_{t\in\{1,2,\cdots,T\}}(\theta_{kt})$$

2) 节点颜色表示景点所属主题。实验中，由于同一景点可能属于不同主题，为使可视化结果更为清晰，本文选择其中景点分布概率最大的主题作为其代表主题，景点所属主题$Topic(k)$定义如下：

$$Topic(k) \propto \arg\max_{t} \max_{t\in\{1,2,\cdots,T\}}(\theta_{kt})$$

3) 节点间距表示景点相似度。景点相似度$Sim(i,j)$越大，节点间连接强度越大，节点间距越小；反之，节点间距越大。景点相似度$Sim(i,j)$以余弦相似度来衡量，计算公式如下：

$$Sim(i,j) = \frac{\boldsymbol{\theta}_i \cdot \boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_i\| \cdot \|\boldsymbol{\theta}_j\|}$$

图6和图7分别为“长城”主题和“运动”主题景点关联图。从图6可以看出，“长城”主题包含“八达岭长城”“居庸关长城”等各段长城景点及“詹天佑铜像”“清华园火车站遗址”等位置相关景点，且“八达岭长城”节点为该主题的热门

景点,与经验判断相符; 从图 7 可以看出, "运动"主题,不仅包含"鸟巢" "首都体育馆"等运动场馆,还包含"奥林匹克森林公园"等体育公园、"央视总部大楼"等赛事转播单位,且"鸟巢"节点为该主题的热门节点,与经验判断相符。

表 4　STM 模型,部分主题单词分布及相关景点

Topic #8 Ⅱ		Topic#19 Ⅱ		Topic#55 Ⅱ		Topic#67 Ⅱ		Topic#78 Ⅱ	
单词	概率/%	单词	概率/%	单词	概率/%	单词	概率/%	单词	概率/%
时尚	6.159	基地	13.044	大型	6.339	门票	29.428	门票	8.277
朝阳	5.791	电影	7.848	昌平	5.015	电话	9.973	雨水	6.109
品牌	4.596	影视	6.151	十三陵	4.920	免费	8.648	高速	6.109
购物中心	3.677	学生	3.606	风景	3.880	四季	7.044	春意盎然	6.011
商务	2.758	第一	2.864	独特	3.501	印象	6.416	复苏	5.715
Topic #8 Ⅱ		Topic#19 Ⅱ		Topic#55 Ⅱ		Topic#67 Ⅱ		Topic#78 Ⅱ	
景点	概率/%	景点	概率/%	景点	概率/%	景点	概率/%	景点	概率/%
燕莎奥特莱斯购物中心	1.910	飞腾影视城	2.666	十三陵水库	5.754	西山	1.382	崔永平皮影艺术博物馆	1.372
津乐汇	1.417	中影基地旅游城	1.851	密云水库	1.301	天安门	1.305	中国坦克博物馆	0.863
世贸天阶	1.238	中国电影博物馆	1.420	圣恩禅寺	1.119	香山	1.266	上地樱桃观光园	0.817
金源新燕莎 MALL	1.148	怀柔影视基地	1.180	小汤山	1.119	八达岭长城	1.227	御生堂中医药博物馆	0.817
新中关购物中心	1.059	中国影视大乐园	1.036	西山	0.983	天安门广场	0.839	北京通信电信博物馆	0.817

图 5　北京地区景点关联图

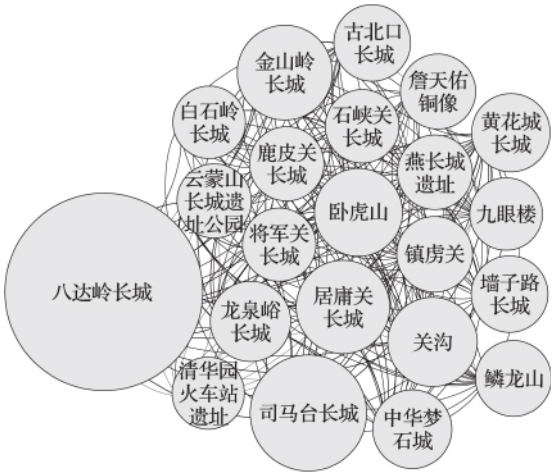

图 6　"长城"主题景点关联图

5　结语

本文基于概率主题模型提出了一种景点-主题模型,用以无监督地从海量的旅游文本中挖掘景点类型与主题特征。模型中引入"全局景点"以过滤噪声语义及无效主题。聚类实验表明,该模型可利用旅游文本中多景点关联关系更准确地捕捉景点主题特征,且"全局景点"的引入能明显改善模型训练效果。另外,本文利用复杂网络图对模型训练结果进行可视化展示,形成旅游目的地景点关联图。

由于概率主题模型发展迅速,本文后续研究拟基于 HDP 改进景点-主题模型,自动计算主题变量个数,以期进一步提高模型效果。

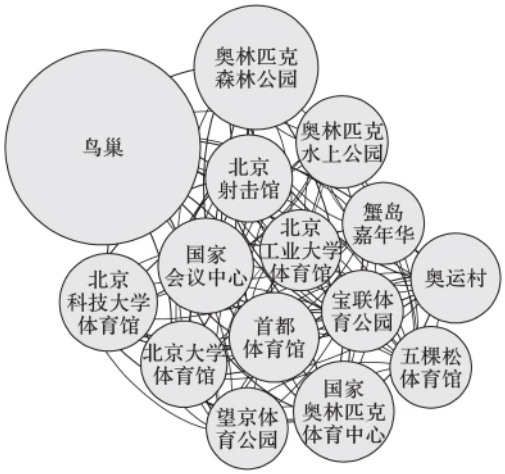

图 7　"运动"主题景点关联图

参考文献:

[1] KOFLER C, CABALLERO L, MENENDEZ M, et al. Near2me: an authentic and personalized social media-based recommender for travel destinations [C]// WSM '11: Proceedings of the 2011 3rd ACM SIGMM International Workshop on Social Media. New York: ACM, 2011: 47 - 52.

[2] CAO L, LUO J, GALLAGHER A, et al. A worldwide tourism recommendation system based on geotagged Web photos[C]// Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2010: 2274 - 2277.

[3] JIANG K, WANG P, YU N. ContextRank: personalized tourism recommendation by exploiting context information of geotagged Web photos [C]// ICIG '11: Proceedings of the 2011 Sixth International Conference on Image and Graphics. Washington, DC: IEEE Computer Society, 2011: 931 - 937.

[4] Panoramio [EB/OL]. [2015-12-10]. http://www.panoramio.com/.

[5] 王媛,许鑫,冯学钢,等. 基于文本挖掘的古镇旅游形象感知研究——以朱家角为例[J]. 旅游科学, 2013, 27(5): 86 - 95. (WANG Y, XU X, FENG X G, et al. Research on tourists' percie-

ved image of ancient town using Web text mining methods: a case study of Zhujiajiao [J]. Tourism Science, 2013, 27(5): 86 - 95.)

[6] 方雅贤,宋文琴. 基于网络文本分析旅游目的地形象——以大连为例[J]. 旅游世界·旅游发展研究,2014(4):24 - 31. (FANG Y X, SONG W Q. Research of tourism destination image based on Web text analysis: a case study of Dalian[J]. Journal of Tourism Development, 2014(4):24 - 31.)

[7] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993 - 1022.

[8] MA W-Y, WANG C, WANG J, et al. Mining geographic knowledge using a location aware topic model: US, US7853596[P]. 2010-12-14.

[9] HAO Q, CAI R, WANG X-J, et al. Generating location overviews with images and tags by mining user-generated travelogues [C]// MM '09: Proceedings of the 2009 17th ACM International Conference on Multimedia. New York: ACM, 2009: 801 - 804.

[10] HAO Q, CAI R, WANG C, et al. Equip tourists with knowledge mined from travelogues [C]// WWW '10: Proceedings of the 2010 International Conference on World Wide Web. New York: ACM, 2010: 401 - 410.

[11] LANDAUER T K, DUMAIS S T. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge [J]. Psychological Review, 1997, 104 (2): 211 - 240.

[12] HOFMANN T. Probabilistic latent semantic analysis [C]// UAI '99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 1999: 289 - 296.

[13] TEH Y W, JORDAN M I, BEAL M J, et al. Hierarchical Dirichlet processes [J]. Journal of the American Statistical Association, 2006, 101(476): 1566 - 1581.

[14] KOLLER D, FRIEDMAN N. Probabilistic Graphical Models: Principles and Techniques — Adaptive Computation and Machine Learning[M]. Cambridge, MA: MIT Press, 2011: 45 - 93.

[15] 周宁,吴佳鑫,张少龙. 基于图的Web信息可视化探析[J]. 情报学报,2008,27(5):714 - 720. (ZHOU N, WU J X, ZHANG S L. Research on graph based Web information visualization [J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(5): 714 - 720.)

[16] CASELLA G, GEORGE E I. Explaining the Gibbs sampler [J]. American Statistician, 1992, 46(3): 167 - 174.

[17] 百度旅游[EB/OL]. [2015-11-10]. http://lvyou.baidu.com/. (Baidu Travel [EB/OL]. [2015-11-10]. http://lvyou.baidu.com/.)

[18] WU H C, LUK R W P, WONG K F, et al. Interpreting TF-IDF term weights as making relevance decisions [J]. ACM Transactions on Information Systems, 2008, 26(3): Article No. 13.

[19] 周志华. 机器学习[M]. 北京. 清华大学出版社,2016: 198 - 199. (ZHOU Z H. Machine Learning [M]. Beijing: Tsinghua University Press, 2016: 198 - 199)

[20] ROSEN-ZVI M, GRIFFITHS T, STEYVERS M, et al. The author-topic model for authors and documents [C]// UAI '04: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Arlington, Virginia, US: AUAI Press, 2010: 487 - 494.

[21] 文益民,史一帆,蔡国永,等. 个性化旅游推荐研究综述[C]// 2015中国旅游科学年会论文集. 北京: 中国旅游研究院,2015. (WEN Y M, SHI YF, CAI G Y, et al. A Survey of Personalized Travel Recommendation [C]//the Proceedings of 2015 China Tourism Scientific Annual Meeting. Beijing: China Tourism Academy, 2015.

Background

This work is partially supported by the Specific Research Fund for Doctoral Program of Higher Education, China (20120002110034).

XU Jie, born in 1990, M. S. candidate. Her research interests include business process management, service recommendation, big data.

FAN Yushun, born in 1962, Ph. D., professor. His research interests include enterprise modeling and optimization analysis, business process reengineering, workflow management.

BAI Bing, born in 1990, Ph. D. candidate. His research interests include service computing, service recommendation, big data.

(上接第2098页)

[18] COHEN W W, SCHAPIRE R E, SINGER Y. Learning to order things [J]. Journal of Artificial Intelligence Research, 1999, 10 (5): 243 - 270.

[19] JOACHIMS T. Optimizing search engines using clickthrough data [C]// KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2002: 133 - 142.

[20] HOLLAND S, ESTER M, KIEßLING W. Preference mining: a novel approach on mining user preferences for personalized applications [C]// PKDD 2003: Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, LNCS 2838. Berlin: Springer-Verlag, 2003: 204 - 216.

[21] JIANG B, PEI J, LIN X, et al. Mining preferences from superior and inferior examples [C]// KDD '08: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008: 390 - 398.

[22] KORICHE F, ZANUTTINI B. Learning conditional preference networks [J]. Artificial Intelligence, 2010, 174(11): 685 - 703.

[23] DE AMO S, DIALLO M S, DIOP C T, et al. Mining contextual preference rules for building user profiles [C]// Proceedings of the 14th International Conference on Data Warehousing and Knowledge Discovery, LNCS 7448. Berlin: Springer-Verlag, 2012: 229 - 242.

[24] DE AMO S, BUENO M L P, ALVES G, et al. Mining user contextual preferences [J]. Journal of Information & Data Management, 2013, 4(1): 37 - 46.

[25] DOYLE J, SHOHAM Y, WELLMAN M P. A logic of relative desire (preliminary report) [C]// ISMIS '91: Proceedings of the 6th International Symposium on Methodologies for Intelligent Systems. London, UK: Springer-Verlag, 1991: 16 - 31.

[26] MCGEACHIE M, DOYLE J. Utility functions for ceteris paribus preferences [J]. Computational Intelligence, 2004, 20(2): 158 - 217.

Background

This work is partially supported by the National Natural Science Foundation of China (61572419, 61403328, 61403329), the Natural Science Foundation of Shandong Province (ZR2013FM011, 2015GSF115009, ZR2014FQ016, ZR2014FQ026).

Xin Guanlin, born in 1991, M. S. candidate. Her research interests include graphical model reasoning and learning about CP-nets.

Liu Jinglei, born in 1970, M. S., associate professor. His research interests include artificial intelligence, theoretical computer science.