软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(5):1114−1126 [doi: 10.13328/j.cnki.jos.004955]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于层次狄利克雷过程的交互式主题建模*

严宇宇, 陶煜波, 林 海

(CAD&CG 国家重点实验室(浙江大学),浙江 杭州 310058)

通讯作者: 林海, E-mail: lin@cad.zju.edu.cn

摘 要: 随着信息技术的快速发展,大量的文本数据产生、被收集和存储.主题模型是文本分析的重要工具之一,被广泛地应用于分析大规模文本集.然而,主题模型通常无法直观而有效地结合用户的领域专业知识对模型结果进行修正.针对这一问题,提出了一个交互式可视分析系统,帮助用户对主题模型进行交互修正.首先对层次狄利克雷过程进行了改进,使其支持单词约束;然后,使用矩阵视图对主题模型进行展示,并使用语义相关的词云布局帮助用户寻找单词约束,用户通过添加单词约束迭代优化主题模型;最后,通过案例分析及用户研究来评价该系统的可用性.

关键词: 文本可视化;主题模型;文本分析;层次狄利克雷过程

中图法分类号: TP391

中文引用格式: 严宇宇,陶煜波,林海.基于层次狄利克雷过程的交互式主题建模.软件学报,2016,27(5):1114−1126. http://www.jos.org.cn/1000-9825/4955.htm

英文引用格式: Yan YY, Tao YB, Lin H. Interactive topic modeling based on hierarchical Dirichlet process. Ruan Jian Xue Bao/Journal of Software, 2016,27(5):1114−1126 (in Chinese). http://www.jos.org.cn/1000-9825/4955.htm

Interactive Topic Modeling Based on Hierarchical Dirichlet Process

YAN Yu-Yu, TAO Yu-Bo, LIN Hai

(State Key Laboratory of CAD & CG (Zhejiang University), Hangzhou 310058, China)

Abstract: With the rapid development of information technology, large amounts of text data have been produced, collected and stored. Topic modeling is one of the important tools in text analysis, and is widely used for large text collection analysis. However, the topic model usually cannot be combined with users' domain knowledge intuitively and effectively during the topic modeling process. In order to solve this problem, this paper proposes an interactive visual analysis system to help users refine generated topic models. First, the hierarchical Dirichlet process is modified to support the word constraints. Then, the generated topic models is displayed via a matrix view to visually reveal the underlying relationship between words and topics, and semantic-preserving word clouds is used to help users find word constraints effectively. User can interactively refine the topic models by adding word constraints. Finally, the applicability of this new system is demonstrated via case studies and user studies.

Key words: text visualization; topic model; text analysis; hierarchical Dirichlet process

随着技术的发展,大量的文本数据产生、被收集和存储.特别是近些年社交媒体的发展,每天都会产生大量的文本数据.由于文本数据具有数据量巨大和非结构化的特点,从文本数据中挖掘出对用户有价值的信息变得越来越具有挑战性.在众多的文本分析方法中,主题模型(topic model)可以从文本中挖掘出隐含的语义维度(主题),它提供了一个快速方式来得到一个未标注、有噪声,甚至动态增长的文本集的概要.在 Blei 等人[1]提出引入

* 基金项目: 国家自然科学基金(61472354); 国家高技术研究发展计划(863)(2012AA12A404)

Foundation item: National Natural Science Foundation of China (61472354); National High-Tech R&D Program of China (863) (2012AA12A404)

收稿时间: 2015-07-24; 修改时间: 2015-09-19, 2015-11-09; 采用时间: 2015-12-05

狄利克雷先验的潜在狄利克雷分布(latent Dirichlet allocation,简称 LDA)主题模型之后,主题模型不仅可以从训练集中提取主题,还可以预测测试集的主题,这使得我们可以用一个训练好的模型解释任何一段文本中的语义.目前,主题模型已被广泛地应用于信息检索、文本语义理解、个性化推荐等领域.

主题模型是一种无监督学习模型,其结果的好坏取决于所选的模型参数和训练集,并且具有很高的不确定性.用户通常无法在训练过程中对模型的结果进行修正,特别是一些专业领域,用户无法向主题模型提供一些领域知识来提高主题建模的质量.已有一些研究提出,通过对主题模型添加约束的形式来解决这一问题.Hu 等人[2]提出了交互式主题模型,然而主题建模的结果十分不直观,从中寻找不恰当的结果并添加合适的约束非常耗时、费力.因此,Choo 等人[3]提出了 UTOPIAN,通过可视化帮助用户从文本集中寻找合适的约束.但是,他们的方法是基于非负矩阵分解(nonnegative matrix factorization,简称 NMF),用户无法通过修正后的模型解释新的文本的语义,并且通过散点图的形式对主题关键词进行展示,用户无法直观、快速地查看每个点所代表的单词.为了解决 UTOPIAN 存在的这些问题,本文提出了一个基于层次狄利克雷过程(hierarchical Dirichlet process,简称 HDP)的交互式主题建模可视分析系统.用户可以通过词云向主题模型添加单词约束(规定一组单词必须出现在同一个主题当中),迭代优化主题模型,最终的结果也可以用于解释新的文本的语义.

本文系统首先使用层次狄利克雷过程主题模型训练文本集获得初始主题模型,基于 HDP 的优点是用户无需指定主题的数目.然后,将主题建模的结果通过以列表形式布局的 Termite 视图进行展示.与其他可视化方式相比,Termite 可以直观、简洁地展示主题与主题、主题与单词之间的关系.为了方便用户寻找单词约束,本文以词云的形式对关键单词进行展示.单词根据其语义信息进行布局.与其他可视化方式相比,语义相关的词云可以直观地展示单词与单词之间的语义关系,并且可以展示更多的单词.词云的布局我们使用了 ProjCloud[4]词云,相对于其他根据力引导布局的词云,它的布局更加简单,布局所需的时间相对较少.用户根据可视化界面展示的信息,交互地添加单词约束.在用户指定完所有单词约束后,系统自动基于当前主题模型和单词约束,使用本文改进的交互式层次狄利克雷过程对主题模型进行迭代修正,在修正后,主题模型中反映出用户的领域需求.最终的主题模型可以用来解释新的文本的语义.

本文的主要贡献如下:

- 本文提出了一个交互式主题建模可视分析系统,帮助用户修正主题模型.
- 本文对层次狄利克雷过程进行了改进,提出了交互式层次狄利克雷过程.
- 本文使用维诺图(Voronoi diagram)对 ProjCloud 词云进行了改进,充分利用了词云中间的空白区域.
- 最后,本文通过案例分析,用户研究对本文系统的可用性进行了评估.

本文第 1 节介绍相关的工作.第 2 节介绍层次狄利克雷过程以及如何向层次狄利克雷过程添加单词约束.第 3 节介绍系统的可视编码及交互设计.在第 4 节和第 5 节中,本文通过案例分析和用户研究对系统的可用性进行评测.第 6 节总结本文系统的优缺点及未来工作.

1 相关工作

无监督主题模型已广泛研究.Deerwester 等人[5]提出基于 SVD 矩阵分解的隐性语义分析(latent semantic analysis,简称 LSA)主题模型,其不足之处是 LSA 缺乏严谨的数理统计基础,并且 SVD 分解非常耗时.Hofmann 等人[6]提出了基于概率统计的概率隐性语义分析(probabilistic latent semantic indexing,简称 PLSI).Blei 等人在 PLSI 的基础上添加了狄利克雷先验,提出了 LDA[1]主题模型.Blei 等人又进一步提出了具有层次结构的层次潜在狄利克雷分布(hierarchical latent Dirichlet allocation,简称 hLDA)[7]模型以及主题不互相独立的关联主题模型(correlated topic model,简称 CTM)[8].之前的主题模型都需要预先设定主题数目,Teh 等人[9]基于狄利克雷过程(Dirichlet process,简称 DP)提出了 HDP 模型,无需事先指定主题数目.Hu 等人[2]对 LDA 进行了改进,使其支持单词约束.HDP 相比于 LDA 模型,它无需指定主题的数目,并且能够适应来自多个语料库的文本.但是 HDP 尚不支持用户的领域专业知识约束.本文扩展了 HDP,使其支持单词约束.

近年来,在文本可视分析领域提出了一系列基于主题模型的可视化技术.Havre 等人[10]提出了主题河流图

(ThemeRiver),采用河流这种隐喻来表达主题随时间的演变.河流从左向右流动,将文本按照主题进行分割.给不同的主题河流赋予不同的颜色,并使用堆叠的方式进行展示,同时使用河流的宽度随时间的变化来表达主题大小随时间的变化.Liu 等人[11]将河流这种隐喻和标签云进行结合,提出了基于 LDA 主题分析技术的 TIARA,以更好地展示每条河流所表示的主题含义.为了更好地展示主题的层次性,Dou 等人[12]使用贝叶斯玫瑰树(Bayesian rose tree,简称 BRT)把一系列主题组织成层次结构,并用树的形式对主题进行展示.他们还拓展了主题河流图,提出 Hierarchical ThemeRiver 来展示主题随时间的变化.TextFlow[13]使用递进式的 HDP 计算每个时间段产生的新主题及消失的主题,利用河流隐喻展示主题随着时间如何产生、分裂和合并.RoseRiver[14]对 TextFlow 进一步改进,将文本主题组织成树状结构,利用 tree cut 技术将主题树进行简化,实现可视化层次主题随时间变化的目的.Xu 等人[15]为了研究话题在社交媒体上的传播情况建立了话题竞争模型,并通过河流隐喻进行展示.Sun 等人[16]在 Xu 等人工作的基础上改进了话题竞争模型,引入了话题的合作关系.Liu 等人[17]针对信息在社交媒体上传播这一问题,基于 LBM(lattice boltzmann method)模型分析社交媒体信息传播情况,并使用流的可视化形式进行展示.Cao[18]等人研究社交媒体上用户观点分歧的发生及演化,并借鉴 DNA 双螺旋结构的形式对结果进行展示.

主题模型自身的可视化是展示抽象的主题模型内部信息.Chaney 等人[19]采用单词列表的形式来展示主题模型发现的隐含的语义结构,可以帮助用户理解大部分的主题以及主题在文本中分布.但是这种方式不够直观,并且没有展示主题与主题之间的相关性.Chuang 等人[20]提出了 Termite,它使用表格的形式展示了单词与主题、主题与主题之间的关系.本文使用 Termite 展示主题模型的结果.

如何结合用户领域专业知识提高主题建模的质量,是当前另一热门研究方向.Nguyen 等人[21]提出了一个交互式的框架以帮助用户分析主题模型结果,并对结果进行一定程度的修正.但是他们使用单词列表来展示主题模型,用户需要从大量的单词中寻找重要单词和需要忽略的单词.这种方式不够直观、高效.Choo 等人[3]提出了 UTOPIAN,使用半监督非负矩阵分解来得到潜在的主题.用户通过对降维后得到的主题关键词的散点图进行交互来修正主题模型的结果,这与本文的工作比较类似.但是 UTOPIAN 使用的是矩阵分解,很难拓展到常用的主题模型.通过矩阵分解虽然可以得到潜在的主题,但是无法通过修正后的模型解释新的文本的语义,这极大地降低了交互式修正主题模型的意义.同时,使用散点图的形式对主题关键词进行展示,用户无法快速地查看每个点表示的单词.本文是基于通用的 HDP 模型,学习到的模型可以应用于新的文档集.同时,本文基于单词的相似性,辅助用户在词云中快速、准确地选择单词约束.

2 交互式层次狄利克雷过程

本文使用层次狄利克雷过程,层次是指使用了多层狄利克雷过程(DP).由于它是非参数模型,不需要指定主题数目,并且能够很好地训练来自多个语料库的文本.基于层次狄利克雷过程,本文提出了交互式层次狄利克雷过程,即在原来的层次狄利克雷过程的基础上引入了单词约束.

2.1 层次狄利克雷过程

狄利克雷过程是概率度量(probability measure)的度量,它有两个参数:基础概率度量(base probability measure)和聚焦参数(concentration parameter).狄利克雷过程可以很好地对数据进行聚类,并且不需要指定聚类数目.层次狄利克雷过程(HDP)[9]是一个层次版本的狄利克雷过程,在 DP 之上再引入一层 DP.我们可以使用一个 DP 来描述一个文本,但是主题在文本之间是共享的,因此需要再引入一层 DP.先将整个文本集通过 DP 过程分成一个个主题;然后,每个文本的 DP 将文本分成一个个主题.这里,每个主题来自上一层 DP,即一个全局的度量 G_0 从一个基础概率度量为 H、聚焦参数为 γ 的 DP 中产生;之后,特定组的随机度量 G_j 从一个基本度量为 G_0 的 DP 中产生:

$$\left.\begin{aligned} G_0 &\sim DP(\gamma, H) \\ G_j / G_0 &\sim DP(\alpha_0, G_0) \end{aligned}\right\} \tag{1}$$

这里,α_0 是度量 G_j 的聚焦参数.

从 stick-breaking 表示的角度来看 HDP:

$$\left.\begin{aligned}&\beta\,|\,\gamma\sim Stick(\gamma)\\&\pi_{j|\alpha_0,\beta}\sim DP(\alpha_0,\beta)\\&\theta_k\,|\,H\sim H\\&z_{ji}\,|\,\pi_j\sim\pi_j\\&x_{ji}\,|\,z_{ji},(\theta_k)_{k=1}^{\infty}\sim F(\theta_{z_{ji}})\end{aligned}\right\}\tag{2}$$

这里,z_{ji}表示文档 j 中第 i 个单词的主题,x_{ji}表示文档 j 中第 i 个位置表示的单词,θ_k表示第 k 个主题的单词分布.

在给定 z^{-ji},β,H 和α_0的情况下,文档 j 中第 i 个单词 w 的主题为

$$\begin{aligned}p(z_{ji}=k\,|\,z^{-ji},\beta,H,\alpha_0)&=\frac{(n_{kw}+H)}{(n_k+HV)}\frac{(\alpha_0\beta_k+n_{jk})}{(\alpha_0+n_j)}\\&\sim\frac{(n_{kw}+H)(\alpha_0\beta_k+n_{jk})}{(n_k+HV)}\end{aligned}\tag{3}$$

这里都不包含第 j 篇文档中第 i 个单词.n_{kw}表示属于第 k 个主题的单词 w 出现的次数,n_k表示属于第 k 个主题的单词出现的次数,n_{jk}表示第 j 篇文档属于第 k 个主题的单词出现的次数.

2.2 交互式层次狄利克雷过程

Hu 等人[2]提出了交互式主题模型,他们拓展了潜在狄利克雷分布,使 LDA 支持一定程度的单词间的约束.本文将这一思想拓展到了 HDP.用户可以向 HDP 添加约束,规定一组单词必须出现在同一个主题当中.如图 1 所示,“hardware”和“cpu”这两个词很有可能属于同一个主题,所以我们添加了{hardware,cpu}这样一个约束.这意味着,这两个词在一个主题中出现的可能性可能同时高或者同时低.这里,约束是可以传递的:“hardware”和“cpu”之间存在约束,“cpu”和“disk”之间也存在约束,则“hardware”和“disk”之间也存在约束.

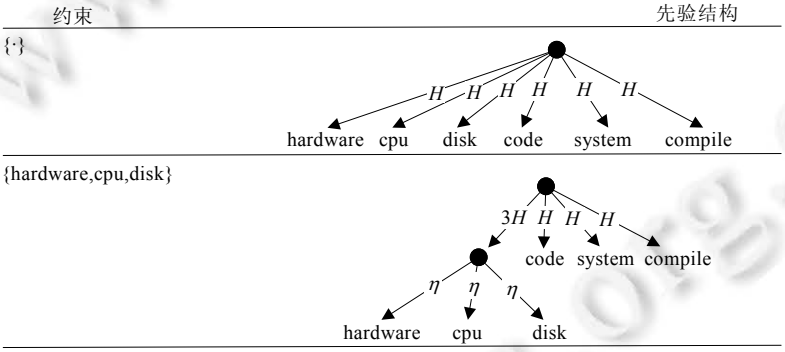


Fig.1 Add word constraints to create new topic priors

图 1 添加约束生成新的先验主题

从图 1 中可以看出,对于每个主题分布θ_k生成的分支,可能是单词或者是一个约束.如果它是一个约束,我们选择约束索引 l_{ji},然后从这个约束的单词分布中产生属于该约束的一个单词 w_{ji}.同时,我们需要引入约束的单词分布的超参数η.

在给定 z^{-ji},β,H,α_0 和η的情况下,文档 j 中第 i 个单词 w 的主题为

$$p(z_{ji}=k\,|\,z^{-ji},\beta,H,\alpha_0,\eta)\sim\begin{cases}\dfrac{(n_{kw}+H)(\alpha_0\beta_k+n_{jk})}{(n_k+HV)}, & \text{if }\forall l,w_{j,i}\notin\Omega_l\\[2ex]\dfrac{(n_{kl}+C_lH)(\alpha_0\beta_k+\varepsilon n_{jk})}{(n_k+HV)}\dfrac{n_{k,l,w_{ji}}+\eta}{n_{kl}+C_l\eta}, & w_{j,i}\in\Omega_l\end{cases}\tag{4}$$

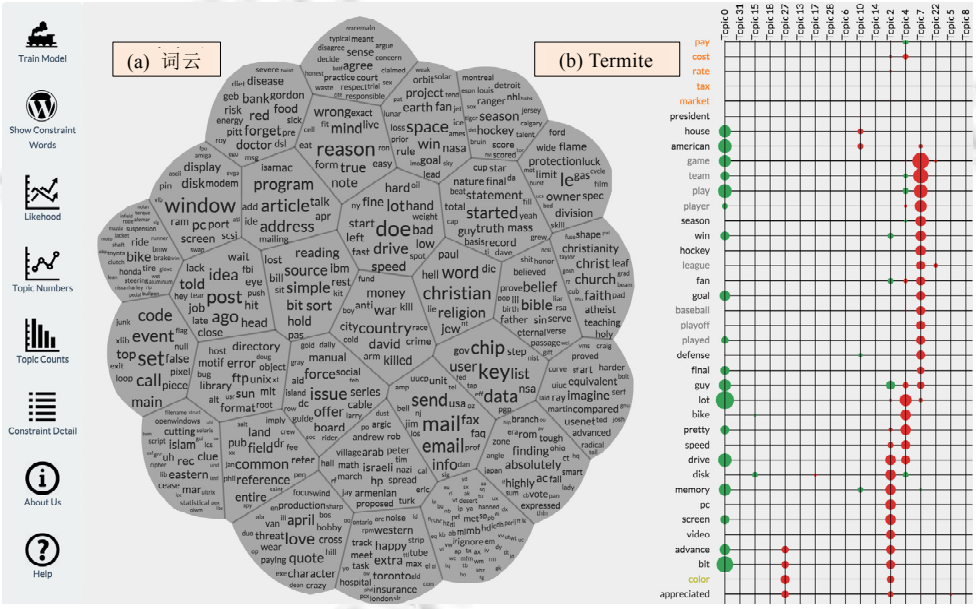
公式(4)与公式(3)相同,不包含第 j 篇文档中第 i 个单词.Ω_l 表示第 l 个单词约束,n_{kl} 表示属于第 k 个主题并且属于约束 l 的单词出现的次数,C_l 表示约束 l 包含的单词数目,$n_{k,l,w_{ji}}$ 表示出现约束 l 中的单词 w 且属于主题 k 的次数,n_{kl} 表示出现约束 l 中的单词且属于主题 k 的次数.w_{ji} 表示第 j 篇文档 i 位置上的单词 w.这里,我们对 n_{jk} 添加了一个权值 ε.由于我们发现当训练的文本较短时,n_{jk} 对单词的主题分配会起到很大的影响,添加约束不能达到很好的效果.因此,我们给 n_{jk} 添加了一个权值,希望减少它对约束单词主题分配的影响.

当一个单词不存在约束时,它的迭代过程和 HDP 是相同的.当单词属于一个约束时,把约束中的单词看成一个整体.使用层次狄利克雷过程计算得到的是约束 l 在给定 z^{-ji},β,H 和 α_0 的情况下,出现主题 k 的概率.对于约束中的某个单词,还需要乘上该单词在其约束中出现的概率.

本文系统是对已构建的主题模型进行修正,所以本文使用当前的主题分配的结果作为一个新的马尔可夫链的初始值进行迭代.由于在原来模型的基础上添加了新的单词约束,这些单词已有主题的分配就显得不大合理.所以,我们把这些单词的主题标记为未分配,然后把这个主题分配的结果作为一个新的马尔可夫链的初始值进行迭代.训练之后,约束单词的主题分布会变得相似,同时,这一结果会影响到与其相关的单词,如在同一文本中或同一主题中的单词.

3 可视编码与交互

本文系统主要由两部分组成:词云和 Termite[20],如图 2 所示.词云展示单词的主题分布的相似程度:分布近似的单词互相靠近,差异大的单词相互远离.当用户看到两个相似的单词彼此远离时,用户可以对其添加约束,使它们相互靠近.Termite 展示主题建模的结果(主题的单词分布以及主题与主题之间的关系),让用户可以直观地判断主题建模结果的好坏.用户可以根据词云视图选择需要添加约束的单词,也可以通过 Termite 分析主题分布,添加相应的约束.



(a) 词云根据单词的语义信息对单词进行布局,语义相近的单词会出现在同一个词云区域或者附近的词云区域;
(b) Termite 是一个矩阵视图,每一行表示单词分布,每一列表示主题分布

Fig.2 System overview

图 2 系统概览

3.1 可视编码

3.1.1 单词显著性度量

由于文本集中可能包含成千上万的单词,显示所有单词既不可行也没有必要.本文通过定义单词的显著性程度,筛选出重要的单词进行显示.单词的显著性度量方法和 Termite 中的方法类似.

$$saliency(w)=P(w)\times distinctiveness(w) \tag{5}$$

$P(w)$为单词的频率,$distinctiveness(w)$作为单词的权重,因为直接使用频率并不能很好地表达单词的显著性.有的单词比较常用,出现的频率自然高.我们希望那些在每个主题中都有可能出现的单词得到一个较低的权重.而那些只出现在特定主题中、与主题含义紧密相连的单词得到一个较高的权重.所以,我们需要定义一个单词在各个主题中分布的差异性程度.

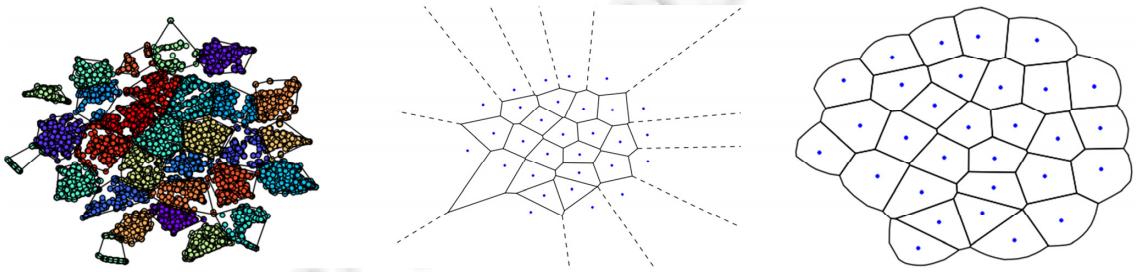
给定一个单词 w,我们计算它的条件概率,即 $P(T|w)$(单词 w 是由主题 T 产生的可能性),同时我们计算边缘概率,即 $P(T)$(任何一个随机的单词由主题 T 产生的可能性).我们使用 $P(T|w)$和 $P(T)$的 Kullback-Leibler(KL)散度定义单词 w 的差异性程度:

$$distinctiveness(w)=\sum_{T}P(T\mid w)\log\frac{P(T\mid w)}{P(T)} \tag{6}$$

$distinctiveness(w)$反映单词 w 的主题分布与整个文本集合的主题分布的差异程度:如果单词 w 的主题分布是随机的,它会接近整个文本集合的主题分布,$distinctiveness(w)$的值会越小.所以,$distinctiveness(w)$值越大,则该单词越是主题相关的.公式(5)中,我们把 $distinctiveness(w)$作为一个单词的权重乘以单词出现的概率,并把计算得到的 $saliency(w)$值作为单词显著性的度量来进行单词的筛选.

3.1.2 词云布局

本文词云的布局使用了 ProjCloud[4],由于 ProjCloud 使用凸包(convex hull)来确定词云边界,会造成每个区域的词云之间存在大量空隙,不仅浪费空间而且十分不美观.因此,本文使用维诺图对 ProjCloud 进行了一定的修改.凸包是指包含聚类点集的最小凸多边形,因此凸包之间往往存在很大的空隙;而维诺图构建多边形对平面进行划分,多边形之间不存在多余的空隙.根据主题模型的结果,我们构造这样一个矩阵 W,W_{ik}表示单词 i 由主题 k 产生的条件概率 $P(T|w)$.这个矩阵有 V 行 T 列,V 表示词汇的数量,T 表示主题的数量.我们使用 t-SNE 将这个矩阵降到二维,每个单词被映射为二维空间中的一个点.然后,使用聚类算法对这些点进行聚类,这里我们使用了 k-means 算法,由于其快速、简单的特点.通过聚类算法,单词映射到平面上的点被分成了几类.我们需要根据聚类后的结果对平面进行分割.ProjCloud 中直接根据聚类后的点计算凸包,然后在每个凸包中填入单词,这样造成了大量显示空间的浪费.我们借鉴了 GMap[22]的思想,使用维诺图代替凸包.我们在相应的凸包边缘加入随机点,然后根据随机点和聚类中心求解维诺图.然后,对每个类聚类中心和随机点对应的多边形进行合并.但是,这样得到的结果不够美观,有很多尖锐的角,并且有的多边形不封闭,如图 3(b)所示.这里,我们根据所有点产生一个凸包,然后在凸包外面添加随机点.然后产生相应的维诺图,取合并后的多边形作为我们需要的词云边界,如图 3(c)所示.



(a) 根据凸包求得多边形的边界 (b) 不加随机点的维诺图结果 (c) 在边界外添加随机点后的维诺图结果

Fig.3 Comparison between convex hull layout and Voronoi diagram layout

图 3 凸包布局与维诺图布局的比较

因此,多边形所包含的面积与每个类所占的面积有关.然后在多边形区域内按 $saliency(w)$值的大小顺序填入单词,这样可以保证聚类后属于同一类别的单词会出现在同一个词云区域内.单词的字体大小与 $saliency(w)$值相关.单词的颜色用来区分不同的约束,其中,黑色表示不在任何约束中.

最后词云的结果如图 2 所示,从图中我们可以看出:单词的位置基本表达了它的语义信息,语义相近的单词基本会出现在同一个词云区域或者附近的词云区域.当用户发现两个语义相近的单词离得很远时,说明当前的主题模型可能将这两个语义相近的单词分配到了不同主题中,因此需要添加单词约束增加这两个单词出现在同一个主题中的概率,从而达到修正主题模型结果的目的.

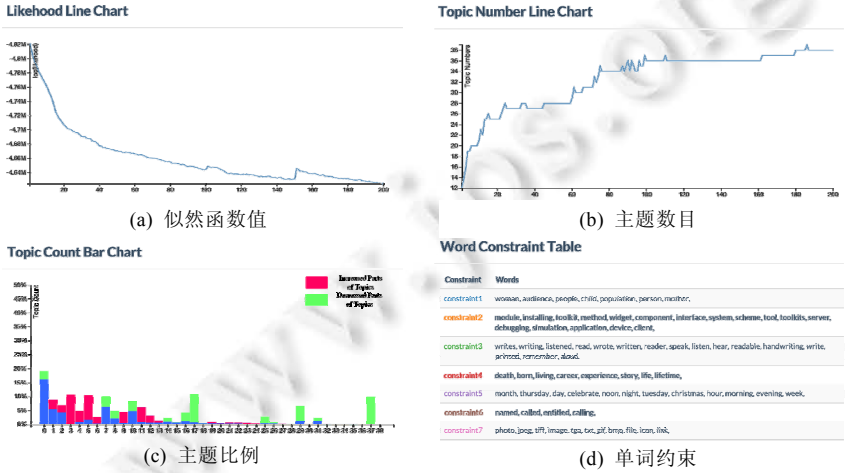
3.1.3 Termite

Termite 的可视化编码比较简单.如图 2 中的 Termite 所示.Termite 是一个矩阵视图,每一行表示一个单词的主题分布,每一列表示主题的单词分布.这里,我们并没有使用单词 w 在主题 T 下出现的概率 $P(w|T)$作为矩阵中元素的值,而是使用 $P(T|w)\log\frac{P(T|w)}{P(T)}$ 作为矩阵中元素的值.$P(w|T)$的大小并不能很好地反映单词的主题相关性,当单词比较常见时,$P(w|T)$更有可能取到较大的值.而 $P(T|w)\log\frac{P(T|w)}{P(T)}$ 像第 3.1.1 节中描述的那样,可以很好地描述单词 w 与主题 T 的相关性.在 Termite 视图中,本文使用圆的面积表示 $P(T|w)\log\frac{P(T|w)}{P(T)}$ 值的大小,使用圆的颜色表示 $P(T|w)\log\frac{P(T|w)}{P(T)}$ 值的正负(红色表示正值,绿色表示负值).因此,红色圆的面积表示该单词出现在该主题中的概率大于随机概率的程度,绿色圆的面积表示该单词出现在该主题中的概率小于随机概率的程度.

单词之间是不存在顺序关系的,若是按照随机排序,结果就会显得很混乱.我们使用 Bond Energy 算法[23]根据行列的相似程度进行排序,希望相似的行列能够相互靠近.从图 2(b)中我们可以看到,每一个主题相关的单词都被聚在一起.从中我们可以快速地理解每个主题的含义,如,主题 7 很有可能与运动比赛有关.

3.1.4 其他视图

除了上面描述的两个基本视图之外,本文还提供了一些重要信息的统计视图,如似然函数值、主题数目的变化、当前各个主题的比例以及单词约束.如图 4 所示,使用折线图表示似然函数值,主题数目在迭代过程中的变化情况.当前各个主题所占的比例使用柱状图进行表示,红色表示迭代后主题比例增加的部分,绿色表示迭代后降低的部分.除此之外,本文使用列表的形式列举每条约束所包含的单词.



(a) 似然函数值 (b) 主题数目

(c) 主题比例 (d) 单词约束

Fig.4 Other views

图 4 其他视图

3.2 交互式单词约束

当用户浏览主题信息和词云后,需要增加单词约束以修正主题模型结果的不恰当部分.用户通过连续点击一批单词的方式将单词加入约束当中.但是我们发现,从词云显示的上百个单词中寻找单词约束仍然是一件困难的事情.我们希望当用户点击某个单词之后,系统可以帮助用户过滤掉大部分不相似的单词,只显示小部分可能相关的单词.因此,我们需要通过额外的语料库获得单词的相似性程度信息.本文使用 wikipedia 上的英文语料库,它包含数据量大,能够得到更加精确的结果.我们通过维基百科训练得到单词向量,通过计算向量的相似性可以得到单词的语义相似度.

用户交互的流程如图 5 所示,当用户点击 Termite 或者词云中的某个单词 w_0 时,词云中会显示固定数目的语义相近的单词,如图 5(b)所示.然后,用户点击其他单词 w_1,则把该单词加入约束中,如图 5(c)所示.如果单词 w_1 属于其他约束了,则对这两个约束进行合并.当用户点击已经在约束中的单词时,可以将该单词从约束中剔除.当选择完单词后,用户点击空白区域表示完成本次单词约束定义,词云就会显示所有的单词,如图 5(d)所示.用户也直接通过点击 Termite 中的单词来添加单词约束.

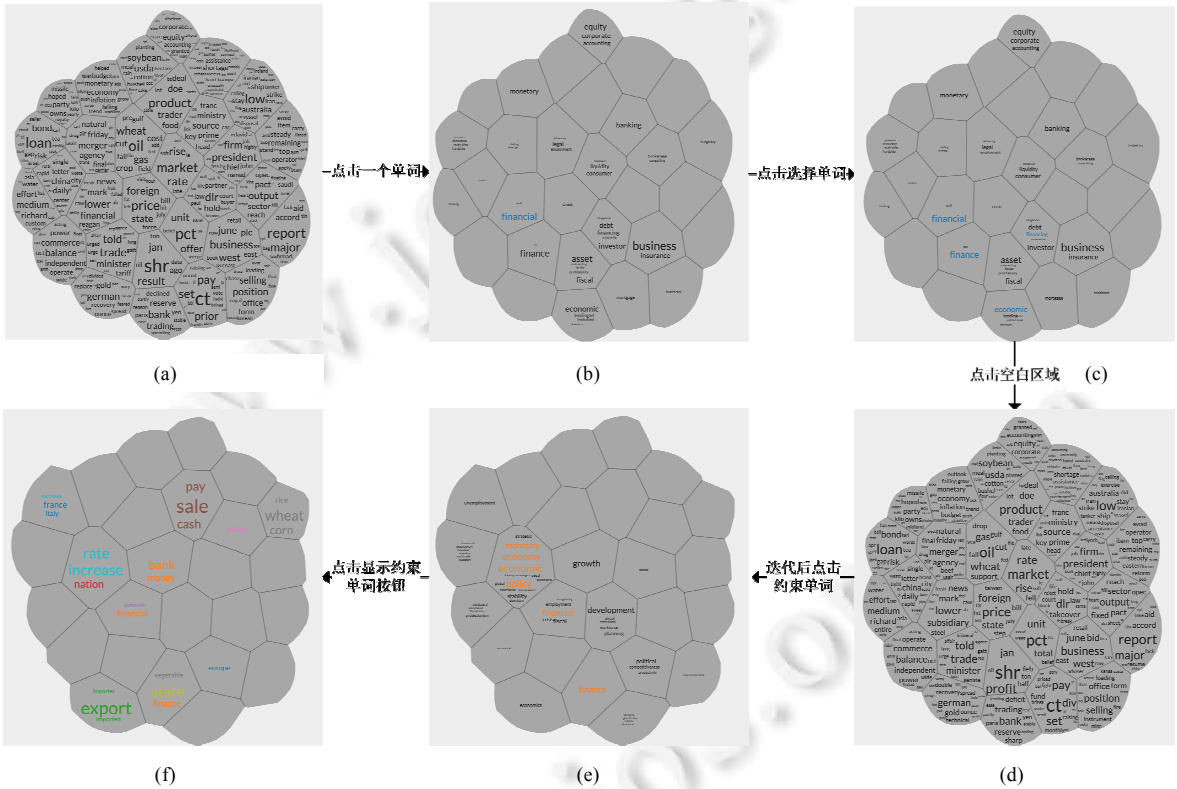


Fig.5 Process of adding word constraints

图 5 添加单词约束流程

由于词云的显示空间有限,只能显示部分单词.部分 *saliency*(*w*)高、本应该使用较大字体着重显示的单词,由于显示空间的限制而无法显示.而用户对某个单词添加完约束后,该单词的重要性程度已经大为降低.如果继续显示该单词,则会造成显示空间的浪费.所以当用户添加完约束后,我们从词云中删除包含在约束中的单词,腾出显示空间来显示其他单词.用户可以通过如图 4(d)所示的单词约束列表,查看当前已构建的单词约束.

当用户点击左侧训练模型按钮时,可以对主题模型添加约束并进行迭代.如图 5(e)是迭代后、在 Termite 视图点击单词“finance”后的结果.可以看出,约束单词被聚在一起.当用户需要查看约束单词在词云中的位置时,可

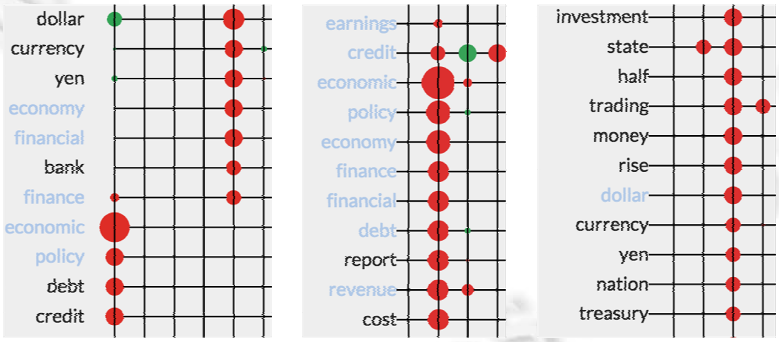
以点击左侧的显示约束单词按钮,则词云会隐藏非约束单词,只显示约束单词.再次点击该按钮则会恢复,如图5(f)所示.

4 案例分析

本文系统采用浏览器/服务器模式,使本文的系统更容易分享和访问.服务器端我们使用 python 的 Django 框架实现.主题模型和降维方法 T-SNE 因为速度的需要,我们使用 C++实现.前端我们使用 Bootstrap 前端框架,并使用 D3[24]实现可视化编码.

4.1 路透社新闻数据

路透社(Reuters)文本分类数据集(Reuters-21578,Distribution 1.0)(https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection)是比较常用的一个文本数据集,包含 13 328 篇文档.对该文本集加载并训练迭代 100 次后,我们在词云中选择了一个感兴趣的单词“financial”,如图 5(b)所示.然后发现,单词“financing”,“finance”,“economic”等相关单词并没有分布在一起.我们将这些单词添加到约束中,然后在 Termite 中查看这些单词在哪些主题中.我们发现,“finance”出现在一个无关的主题中,该主题排名靠前的几个单词为“west”, “dealer”,“german”,“spokesman”.这个主题的含义很混乱.我们添加完与“finance”单词相关的约束后,对模型进行迭代 50 次.迭代后我们发现:“finance”单词被转移到了另一个主题中,如图 6(a)所示.我们发现,约束单词并不在一个主题中,但是这两个主题的语义十分近似.我们希望对这两个主题进行合并,因此,我们将在这两个主题中与财政相关的单词“bank”,“dollar”,“debt”,“credit”也添加到约束当中,继续迭代了 50 次.最后,主题的结果如图 6(b)所示,约束单词被很好地聚在一个主题当中.但是单词“dollar”并不在该主题当中,如图 6(c)所示,它出现在另一个主题中,该主题包含“money”,“yen”,“trading”等单词.可以猜测,这个主题应该与货币交易有关.对于“yen”,“currency”这些非约束单词,他们与“dollar”一起发生了转移.从这个结果中可以看出,单词约束并不会把单词强行捆绑在一个主题中,只是使它们出现在一个主题中的概率变大.这样可以有效防止不恰当约束对结果产生太大的影响.



(a) (b) (c)

Fig.6 Reuters news data case study

图 6 路透社新闻数据案例分析

这次案例分析的似然函数如图 4(a)所示.添加单词约束使得似然函数减小,但是很快就收敛并比迭代前的结果更好.层次狄利克雷过程的主题是不固定的,图 4(b)显示这次案例分析中两次交互迭代过程中主题数目的变化情况.图 4(c)显示了最后一次迭代时每个主题所占比例的变化.在案例分析中,我们对两个主题进行了合并,合并后,主题 9 所占的比例明显增加.

4.2 20 Newsgroups数据集

20 Newsgroups(http://qwone.com/~jason/20Newsgroups/)数据集包含 18 846 篇文档,被分成 20 个不同的类.

我们移除了文本中的一些元信息,例如消息头、脚注、引用等.同时,我们删除了那些正文内容只包含几个单词的文档.

我们对这个数据加载并迭代 100 次,得到主题模型的一个初始结果.然后,我们对数据包含的主题进行了分析,我们发现,数据中包含不少与计算机相关的单词,但是这些单词分布在一些不相干的主题当中,如图 7(a)所示.我们将“pc”,“unix”,“chip”等单词加入到约束当中,然后迭代了 50 次,然而却没有得到我们所期望的结果,如图 7(b)所示,有 4 个主题的单词被混杂在约束单词所在的主题.“pc”,“ibm”,“software”等单词明显与计算机有关,“war”,“israel”,“army”这些单词与战争有关,“god”,“jesus”,“christian”这些单词与宗教信仰有关,“government”,“country”,“federal”这些单词与政府有关.原来这些单词所在的主题混杂着其他单词,同时,同一主题的单词分散在多个主题中.迭代后,单词被聚在一起,但是这 4 个不同的主题被分在一个主题中.所以我们分别对这 4 个主题的单词添加约束,使它们和语义相近的单词之间建立约束关系,然后继续迭代.最后,得到的结果如图 7(c)~图 7(f)所示,这 4 个主题被很好地分离开了.

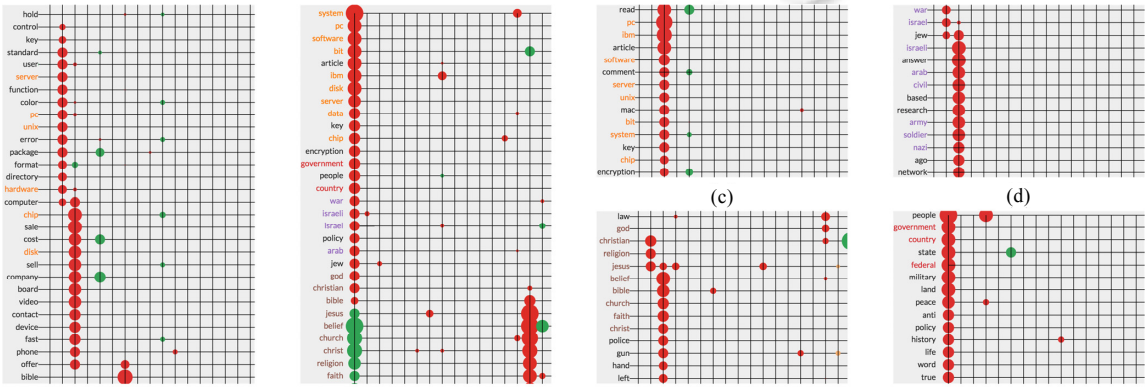


Fig.7 20 Newsgroups data set case study

图 7 20 Newsgroups 数据集案例分析

5 用户研究

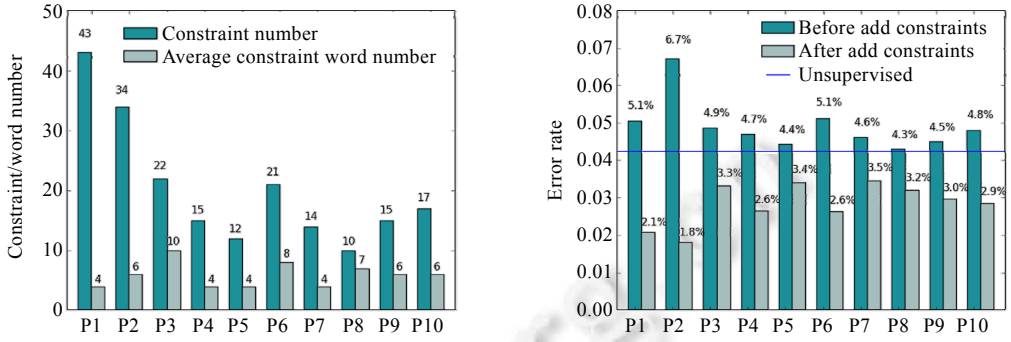
使用聚类算法检测主题建模结果的好坏,是一种常用的文本主题建模评价方法[21,25].根据每个文本包含各个主题的比例,对文本使用 *k*-means 聚类.聚类结果的错误率越低,说明主题模型得到的隐含语义维度越能表达文本集原有的结构.

5.1 过 程

为了验证本文系统的可用性,我们邀请了 10 名用户对本文的系统进行使用.这 10 名用户中,有 3 位博士研究生、6 位硕士研究生、1 位本科生,其中有 6 位从事可视化方面的研究工作,2 位对可视化有一定的了解,剩下 2 位对可视化不太熟悉.10 名用户中,有 3 位从事文本可视化方面的研究.我们要求他们根据我们呈现的可视化结果,对主题模型添加相应的约束.在我们讲解如何使用本文的系统之后,他们很快就掌握了如何使用本文的系统.这里,我们使用了 20 Newsgroups 这个数据集对主题模型进行训练,并把数据本身的分类结果作为真实值来检验我们的聚类结果.

为了保证我们计算聚类结果时主题模型已经充分迭代,而不是还未收敛的结果,我们对没加约束的主题模型进行了分析.我们发现,当主题模型迭代到 300 次之后,继续迭代,聚类的准确率基本保持不变.所以我们认为,主题模型迭代 300 次后,主题模型已经充分收敛.我们首先迭代 300 次,将这个结果通过本文的系统呈现给用户.用户根据这个结果添加相应的约束,然后点击训练按钮对模型添加约束并迭代 300 次.迭代完之后,本文的系统会显示相应的结果,用户可以继续添加约束进行迭代.在这次测试中,为了方便评估单次交互对主题模型的影

响,本文只取第 1 次交互后的结果.为了避免迭代次数对结果造成影响,我们将用户第 1 次交互后的结果与没有添加约束直接迭代 600 次后的结果进行比较.由于主题建模结果的不确定性,每次得到的结果存在波动.我们重复进行了 100 次,取平均值作为没有约束存在的情况下主题建模的结果,与用户交互修正后的主题模型建模结果进行比较,结果如图 8 所示.



(a) 每位用户添加的平均约束数目及每条约束的平均单词数目 (b) 添加约束与不添加约束的错误率比较

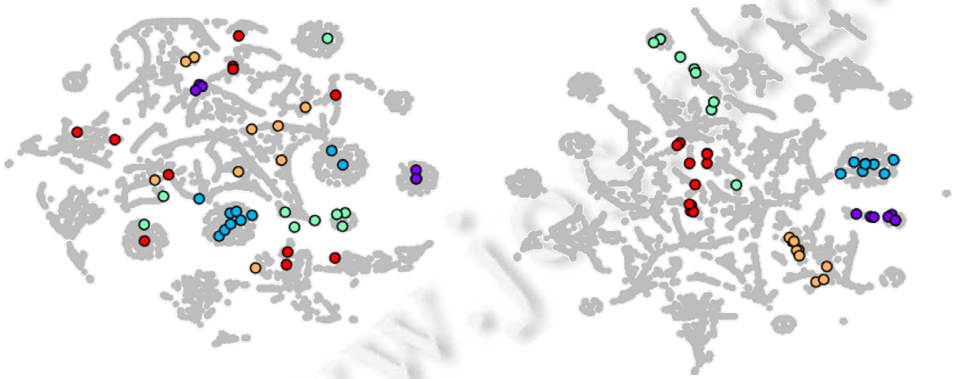
Fig.8 Result of user study

图 8 用户研究结果

5.2 结果和讨论

用户对主题模型进行修正后,得到的分类结果的准确率都比直接迭代 600 次的要高.而准确率的提高程度,与添加的约束数目及平均约束单词数目呈现正相关关系.聚类的结果在一定程度上可以表示主题模型提取的主题质量,因此我们认为,本文所提出的系统可以有效地帮助用户提高主题建模的质量.

为了说明我们的模型可以按照用户的意图,尽可能地使同一约束的单词出现在同一主题中,即在降维后的平面上,同一约束的单词的位置尽可能地接近.我们选择了一位用户的部分约束,如图 4(d)所示,查看它们在交互前后的位置变化.图 9(a)是迭代前每个单词的位置,每个点表示一个单词,非约束单词使用灰色表示,属于不同约束的单词使用不同的颜色进行表示;图 9(b)表示添加这些约束进行迭代后,单词在平面中的位置.从图中我们可以发现,用户添加的约束起到了很好的作用,属于同一约束的单词基本都聚到了一起.



(a) 迭代前约束单词原来的位置 (b) 迭代后约束单词的位置

Fig.9 Position change of constraint words

图 9 约束单词的位置变化

6 总 结

本文展示了一个基于层次狄利克雷过程的交互式主题建模可视分析系统.用户可以使用该系统对得到的主题模型进行修正.由于系统使用的是基于层次狄利克雷过程改进的主题模型,所以不用预先设置主题的数目.在不了解文本集的情况下,预先设定主题的数目是件困难的事情.本文的系统可以很好地拓展到其他主题模型.由于本文提取主题的方法是基于主题模型而非矩阵分解,用户可以使用修正后的主题模型去解释新的文本的语义.

同时,本文的系统还存在一些不足.由于我们对主题模型的求解是基于吉布斯采样(Gibbs sampling),每次得到的主题模型结果存在不确定性;同时,求解时迭代的速度还未达到实时交互的要求.在单词可视化方面,词云所能显示的单词仍然太少,虽然通过交互进行了一定的弥补.同时,由于单词的歧义性,使得确定单词间的约束关系变得困难.用户可能会添加一些不恰当的约束:对一些主题明确的单词,这些约束不会产生太大的影响;而对于一些主题不是非常明确的单词,会使这些单词被分配到不正确的主题当中.

未来的工作在于应用一些更快、结果不确定性小的主题模型求解方法.由于显示空间的限制,不可能显示过多的单词,如果能粗步筛选出需要添加约束的单词,然后让用户做进一步筛选,这将会极大地提高用户的效率.除此之外,目前约束的类型还是过于单一,还需要添加新的功能,使系统支持主题的分裂合并以及文档层次的约束;另外,还需要添加相应的文本内容视图,以帮助用户更好地理解单词及主题的含义.

References:

[1] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003,3:993−1022.

[2] Hu Y, Boyd-Graber J, Satinoff B. Interactive topic modeling. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT 2011), Vol.1. 2011. 248−257.

[3] Choo J, Lee C, Reddy CK, Park H. UTOPIAN: User-Driven topic modeling based on interactive nonnegative matrix factorization. IEEE Trans. on Visualization and Computer Graphics, 2013,19(12):1992−2001. [doi: 10.1109/TVCG.2013.212]

[4] Paulovich FV, Toledo FMB, Telles GP, Minghim R, Nonato LG. Semantic wordification of document collections. Computer Graphics Forum, 2012,31(3pt3):1145−1153. [doi: 10.1111/j.1467-8659.2012.03107.x]

[5] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990,41(6):391−407. [doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9]

[6] Hofmann T. Probabilistic latent semantic indexing. In: Proc. of the 22nd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM Press, 1999. 50−57. [doi: 10.1145/312624.312649]

[7] Griffiths D, Tenenbaum M. Hierarchical topic models and the nested Chinese restaurant process. In: Advances in Neural Information Processing Systems 16: Proc. of the 2003 Conf. 2004.

[8] Blei D, Lafferty J. Correlated topic models. Advances in Neural Information Processing Systems, 2006,18:147.

[9] Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 2006, 101(476):1566−1581. [doi: 10.1198/016214506000000302]

[10] Havre S, Hetzler B, Nowell L. Themeriver: Visualizing theme changes over time. In: Proc. of the IEEE Symp. on Information Visualization (InfoVis 2000). IEEE, 2000. 115−123. [doi: 10.1109/INFVIS.2000.885098]

[11] Liu S, Zhou MX, Pan S, Song Y, Qian W, Cai W, Lian X. Tiara: Interactive, topic-based visual text summarization and analysis. ACM Trans. on Intelligent Systems and Technology, 2012,3(2):25. [doi: 10.1145/2089094.2089101]

[12] Dou W, Yu L, Wang X, Ma Z, Ribarsky W. Hierarchicaltopics: Visually exploring large text collections using topic hierarchies. IEEE Trans. on Visualization and Computer Graphics, 2013,19(12):2002−2011. [doi: 10.1109/TVCG.2013.162]

[13] Cui W, Liu S, Tan L, Shi C, Song Y, Gao Z, Qu H, Tong X. Textflow: Towards better understanding of evolving topics in text. IEEE Trans. on Visualization and Computer Graphics, 2011,17(12):2412−2421. [doi: 10.1109/TVCG.2011.239]

[14] Cui W, Liu S, Wu Z, Wei H. How hierarchical topics evolve in large text corpora. IEEE Trans. on Visualization and Computer Graphics, 2014,20(12):2281−2290. [doi: 10.1109/TVCG.2014.2346433]

[15] Xu P, Wu Y, Wei E, Peng TQ, Liu S, Zhu JJ, Qu H. Visual analysis of topic competition on social media. IEEE Trans. on Visualization and Computer Graphics, 2013,19(12):2012−2021. [doi: 10.1109/TVCG.2013.221]

[16] Sun G, Wu Y, Liu S, Peng TQ, Zhu JJH, Liang R. EvoRiver: Visual analysis of topic coopetition on social media. IEEE Trans. on Visualization and Computer Graphics, 2014,20(12):1753−1762. [doi: 10.1109/TVCG.2014.2346919]

[17] Liu YH, Wang CB, Ye P, Zhang K. Analysis of micro-blog diffusion using a dynamic fluid model. Journal of Visualization, 2015, 18(2):201−219. [doi: 10.1007/s12650-015-0277-y]

[18] Cao N, Lu L, Lin YR, Wang F, Wen Z. SocialHelix: Visual analysis of sentiment divergence in social media. Journal of Visualization, 2015,18(2):221−235. [doi: 10.1007/s12650-014-0246-x]

[19] Chaney AJ, Blei DM. Visualizing topic models. In: Proc. of the 6th Int'l Conf. on Weblogs and Social Media. 2012.

[20] Chuang J, Manning CD, Heer J. Termite: Visualization techniques for assessing textual topic models. In: Proc. of the Int'l Working Conf. on Advanced Visual Interfaces. ACM Press, 2012. 74−77. [doi: 10.1145/2254556.2254572]

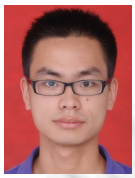
[21] Nguyen VA, Hu Y, Boyd-Graber JL, Resnik P. Argviz: Interactive visualization of topic dynamics in multi-party conversations. In: Proc. of the HLT-NAACL. 2013. 36−39.

[22] Gansner ER, Hu Y, Kobourov S. Gmap: Visualizing graphs and clusters as maps. In: Proc. of the Pacific Visualization Symp. (PacificVis 2010). IEEE, 2010. 201−208. [doi: 10.1109/PACIFICVIS.2010.5429590]

[23] McCormick Jr WT, Schweitzer PJ, White TW. Problem decomposition and data reorganization by a clustering technique. Operations Research, 1972,20(5):993−1009. [doi: 10.1287/opre.20.5.993]

[24] Bostock M, Ogievetsky V, Heer J. D^3 data-driven documents. IEEE Trans. on Visualization and Computer Graphics, 2011,17(12): 2301−2309. [doi: 10.1109/TVCG.2011.185]

[25] Mao XL, Ming ZY, Chua TS, Li S, Yan H, Li X. SSHLDA: A semi-supervised hierarchical topic model. In: Proc. of the 2012 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012. 800−809.



严宇宇(1991－),男,浙江温州人,博士生,CCF 学生会员,主要研究领域为信息可视化,可视分析.

陶煜波(1980－),男,博士,副教授,CCF 专业会员,主要研究领域为数据可视化,可视分析.

林海(1965－),男,博士,教授,博士生导师,主要研究领域为数据可视化,可视分析,电磁计算.