一种基于密度的自适应最优 LDA 模型选择方法

曹 娟[1),2),3)] 张勇东[1),2)] 李锦涛[1),2)] 唐 胜[1),2)]

[1)](中国科学院计算技术研究所虚拟现实技术实验室 北京 100190)

[2)](中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

[3)](中国科学院研究生院 北京 100049)

摘 要 主题模型(Topic models)被广泛应用在信息分类和检索领域.这些模型通过参数估计从文本集合中提取一个低维的多项式分布集合,用于捕获词之间的相关信息,称为主题(Topic).针对模型参数学习过程对主题数目的指定和主题分布初始值非常敏感的问题,我们用图的形式阐述了 LDA(Latent Dirichlet Allocation)模型中主题产生的过程,提出并证明当主题之间的相似度最小时,模型最优的理论.并基于该理论,提出了一种基于密度的自适应最优 LDA 模型选择方法.实验证明该方法可以在不需要人工调试主题数目的情况下,用相对少的迭代,自动找到最优的主题结构.

关键词 主题模型;主题;LDA;密度

中图法分类号

A Method of Adaptively Selecting Best LDA Model Based on Density

CAO Juan[1),2),3)] ZHANG Yong-Dong[1),2)] LI Jin-Tao[1),2)] TANG Sheng[1),2)]

[1)](*Virtual Reality Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[2)](*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[3)](*Graduate University of the Chinese Academy of Sciences, Beijing* 100049)

Abstract Topic models have been successfully used to information classification and retrieval. These models can capture word correlations in a collection of textual documents with a low-dimensional set of multinomial distribution, called "topics". It is important but difficult to select an appropriate number of topics for a specific dataset. In this paper, we propose a theorem that the model reaches optimum as the average similarity among topics reaches minimum, and based on this theorem, we propose a method of adaptively selecting the best LDA model based on density. Experiments show that the proposed method can achieve performance matching the best of LDA without manually tuning the number of topics.

Keywords topic model; topic; LDA; density

1 引 言

统计主题模型(Statistical Topic Models)近年来得到了非常广泛的应用,包括在文本分类、信息检索等领域取得了非常好的应用效果[1-6].给定一个文档集合,主题模型通过参数估计寻找一个低维的多项式分布集合,每个多项式分布称为一个主题(Topic),

收稿日期:2007- - ;最终修改稿收到日期:2007- - .本课题得到国家"九七三"重点基础研究发展规划项目基金(2007CB311100)、国家"八六三"高技术研究发展计划项目基金(2007AA01Z416)、国家自然科学基金(60773056)和北京市科技新星计划(2007B071)资助.
曹 娟,女,1980 年生,博士研究生,主要研究方向为多媒体检索、机器学习. E-mail: caojuan@ict.ac.cn. **张勇东**,男,1973 年生,副研究员,主要研究方向为图像和视频处理技术. **李锦涛**,男,1962 年生,研究员,博士生导师,主要研究领域为多媒体技术、虚拟现实技术. **唐 胜**,男,1972 年生,博士,助理研究员,主要方向为多媒体技术.

用来捕获词之间的相关信息. 主题模型可以在不需要计算机真正理解自然语言的情况下,提取可以被人理解的,相对稳定的隐含语义结构,为大规模数据集中的文档寻找一个相对短的描述.

统计主题模型的思想最早来源于隐含语义索引(Latent Semantic Indexing,LSI)[7]. 其工作原理是利用矩阵理论中的"奇异值分解(SVD)"技术,将词频矩阵转化为奇异矩阵,通过去除较小的奇异值向量,只保留前 K 个最大的值,将文档向量和查询向量从词空间映射到一个 K 维的语义空间(主题). 在该空间中,来自词项-文档矩阵的语义关系被保留,同时词项用法的变异(如同义性、多义性)被抑制. 主题模型的第二次重大突破是 Hofmann 提出的 PLSI (Probabilistic Latent Semantic Indexing)模型[8]. PLSI 通过概率模型来模拟文档中词的产生过程,将 LSI 扩展到概率统计的框架下. 它将文档 d 表示为一个主题混合,文档中每个词作为主题混合中的一个抽样. 但是 PLSI 并没有用一个概率模型来模拟文档的产生,只是通过对训练集中的有限文档进行拟合,得到特定文档的主题混合比例. 这个过程导致 PLSI 模型参数随着训练集中文档数目线性增加,出现过度拟合现象;而且,对于训练集以外的文档,很难分配合适的概率. 针对这些问题,Blei 等在 2003 年提出了 LDA(Latent Dirichlet Allocation)[1],在 PLSI 的基础上,用一个服从 Dirichlet 分布的 K 维隐含随机变量表示文档的主题混合比例,模拟文档的产生过程. 在文本的产生过程中,LDA 首先从 Dirichlet 分布中抽样产生一个文本特定的主题多项式分布;然后对这些主题反复抽样产生文本中的每个词. 在 LDA 的基础上,很多研究人员根据不同的应用需求,开发了如基于无向图模型理论的 Harmonium 模型[9]、支持多模态特征的 dual-wing harmonium 模型[10]、GM-LDA 模型[11]等.

LDA 模型中发现的主题可以捕获词之间的相关性,但 LDA 不能表示主题之间的相关性(基于 Dirichlet 分布的抽样假设主题之间相互独立). 然而,主题的相关性在真实的数据集合中普遍存在,忽略这些相关性将限制 LDA 模型对大规模数据集合的表示能力以及对新数据的预测能力. 因此,近年来很多学者开始研究更丰富的结构来描述主题之间的相关性. Blei 在 2006 年提出了 CTM(Correlated Topic Model)[2]. 与 LDA 类似,CTM 将每个文档表示成一个主题混合,但主题混合比例从对数正态分布(Logistic Normal)中抽样获得. 先验参数包括一个协方差矩阵,用每个主题对之间的协方差描述它们之间的相关性. 基于 CTM 只能描述两两之间相关性的局限性,Li 等进一步提出了 PAM(Pachinko Allocation Model)[3],用一个有向无环图 DAG (Directed Acyclic Graph)表示语义结构. 在 PAM 对应的 DAG 中,每个叶子节点为词表中的一个词,每个中间节点为一个主题,每个主题是基于它的孩子节点的一个多项式分布. PAM 对主题的含义进行了扩展,不仅可以是基于词空间的多项式分布,而且可以是基于其它主题的多项式分布,称为超主题(Super Topic). 所以,PAM 不仅可以描述词之间的相关性,而且可以灵活的描述主题之间的相关性.

虽然 PAM 可以模拟主题之间的相关性,但它和其它主题模型一样,都需要人工确定主题的数目(本文又称参数 K). 参数 K 的设置直接影响到模型提取的主题结构. 一种解决办法是通过 HDP(Hierarchical Dirichlet Process)[12]自动学习主题的数目. HDP 对分组数据建模,该数据具有预先定义的多层结构,每个预先定义的组用一个 DP(Dirichlet Process)表示,该 DP 对应的概率测度从更高层的 DP 中抽样获取. 考虑到 HDP 与 LDA 的结构相似性,Teh 等在文献[12]中用一个 DP 取代 LDA 中有限的主题混合,根据不同的混合比例为每个文档建立一个新的 DP,提取的主题被所有 DP 所共享. Teh 通过分析 HDP 混合模型中混合成分数目抽样直方图,预测主题数目的后验值正好与最优的 LDA 参数 K 一致,从而解决了最优 K 值的选择问题. 在此基础上,Li 等基于 HDP 的思想提出了一个非参数的 PAM[13].

HDP 通过 DP 的非参数特性解决了 LDA 中主题数目选择问题,但这种方法需要为同一个集合分别建立一个 HDP 模型和一个 LDA 模型. 本文通过理论证明和实验分析,得到了最优主题数与主题相似度之间的关系. 以此为约束条件,将最优 K 值选择与 LDA 模型参数估计统一在一个框架里,提出了一种新的基于密度的最优主题数目选择算法. 通过图的形式模拟 LDA 中主题的产生过程,发现新的主题通常由造成主题之间相关性的词(主题分布的重叠区域)产生. 进一步通过实验证明,最优 K 不仅跟文本集合的大小有关,而且跟集合中文本之间的相关程度有关. 通过计算每个主题的密度,寻找已知结构下最不稳定的主题,反复迭代,直到模型稳定. 从而解决最优 K 的选择问题.

2　相关工作

2.1　LDA(Latent Dirichlet Allocation)

LDA 是一个多层的产生式概率模型,包含词、主题和文档三层结构. LDA 将每个文档表示为一个主题混合,每个主题是固定词表上的一个多项式分布. LDA 假设词由一个主题混合产生,同时每个主题是在固定词表上的一个多项式分布;这些主题被集合中的所有文档所共享;每个文档有一个特定的主题比例,从 Dirichlet 分布中抽样产生. 作为一种产生式模型,用 LDA 提取隐含语义结构和表示文档已经成功的应用到很多文本相关的领域[1,4,6].

LDA 的图模型表示如图 1 所示. 给定一个文档集合 D,包含 M 个文档和 V 个不同的词. 每个文档 d 包含一个词序列$\{w_1,w_2,\cdots,w_N\}$. 在集合 D 对应的 LDA 模型中,假设主题数目固定为 K,则一个文档 d 的产生可以表示为以下两个过程:

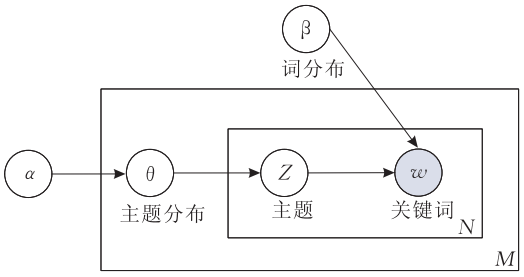


图 1　LDA 的图模型表示(图中空心点表示隐含变量,实心点表示可观察值;矩形表示重复过程. 大矩形表示从 Dirichlet 分布中为文档集中的每个文档 d 反复抽取主题分布 $\boldsymbol{\theta}_d$;小矩形表示从主题分布中反复抽样产生文档 d 的词$\{w_1,w_2,\cdots,w_N\}$)

(1)从 Dirichlet 分布 $p(\theta|\alpha)$中随机选择一个 k 维的向量 $\boldsymbol{\theta}_d$,表示文档 d 中的主题混合比例;

(2)根据特定的主题比例对文档 d 中的每个词进行反复抽样,得到 $p(w_n|\boldsymbol{\theta}_d,\boldsymbol{\beta})$.

其中,α 是一个 K 维的 Dirichlet 参数:

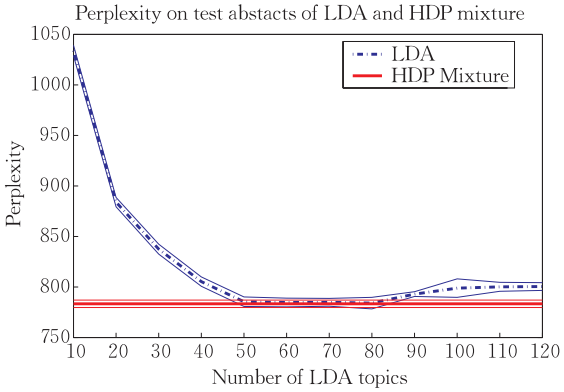
$$p(\theta|\alpha)=\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1} \tag{1}$$

$\boldsymbol{\beta}$ 是一个 $K\times V$ 的矩阵,$\beta_{ij}=p(w_j=1|z_i=1)$,$i=1,2,\cdots,K$;$j=1,2,\cdots,V$.

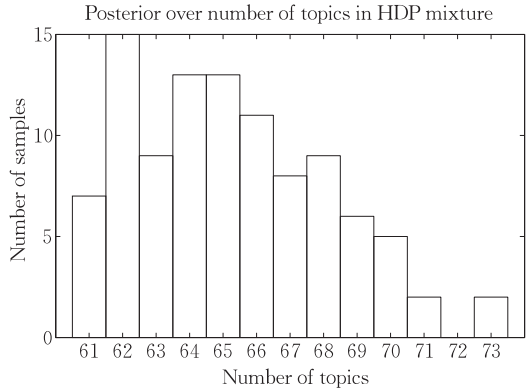
LDA 需要人工指定最优的主题数目 K. 而且,在 LDA 中,主题服从 Dirichlet 分布,该分布假设一个主题的出现与其它主题的出现无关. 在真实数据中,很多主题之间是存在关联的,如“NBA”出现的情况下,“sports”出现的概率比较高,而不太可能出现“disease”. 这种独立假设与真实数据的矛盾使得 LDA 对于主题数目 K 的变化非常敏感. 模型只能预测由同一个主题产生的词,而不能预测跟该主题相关的其它主题产生的词.

2.2　基于 HDP 的最优 LDA 模型选择算法

Teh 在文献[12]中提出用 HDP 寻找 LDA 中的最优 K 值. HDP 是在随机概率测度(probability measure)集上的一种分布,可以对分组数据建模. HDP 包含一个全局的概率测度 G_0,每个组对应一个 DP_i 以及一个概率测度 G_i. G_0 中的成员被所有的 DP 所共享,但不同的 DP 具有不同的混合比例,通过从上层的 DP 中抽样获取. Teh 考虑到 HDP 与 LDA 在结构上的相似性,运用 HDP 的非参数特性来解决 LDA 中主题数目的选取问题. 用一个趋向



(a) LDA与HDP性能比较结果: perplexity随着主题数目变化的曲线(评测指标perplexity的计算公式如式(12)所示, 值越低, 说明性能越好)



(b) HDP中使用主题数目的直方图, 横坐标表示主题数目, 纵坐标表示对应主题数被抽样的次数(抽样大于100次)[12]

图　2

无限的概率测度 G_0 取代 LDA 中有限的主题混合，根据不同的混合比例为每个文档建立一个新的 DP_i 和 G_i，所有文档共享 G_0 中的混合成分. Teh 将两个模型应用在同一个数据集上，结果如图 2 所示：LDA 在主题数为 50～80 时性能最好(图 2(a)). 通过分析该 HDP 中混合成分抽样直方图(图 2(b))发现，最佳的混合成分数正好与 LDA 的最优主题数一致，从而解决 LDA 中最优 K 值的选择问题.

HDP 利用 DP 的非参数特性解决了 LDA 中主题数目的选取问题. 但这种方法需要为同一个集合分别建立一个 HDP 模型和一个 LDA 模型. 本文将最优 K 值选择与 LDA 模型参数估计统一在一个框架里，提出了一种新的基于密度的最优主题数目选择算法. 在第 3 节，我们从 LDA 中主题产生的过程，深入分析了 LDA 中最优主题数与主题相似度之间的关系. 基于这种关系，我们在第 4 节提出了一种基于密度的最优主题数目选择算法，并在第 5 节通过实验验证了该算法的有效性.

3 最优模型与主题相似度的关系

主题模型通过分析大量的统计数据，提取数据集中隐含的主题结构. 每个数据集都有一个最优的结构. 在本节中，我们提出，当主题之间平均相似度最小时，模型最优. 并分别从理论和实验两方面进行了证明.

3.1 理论证明

我们用 $\boldsymbol{\beta}$ 矩阵中主题在 V 维词空间的分布 $p(w_v|Z_i)$ 来表示主题向量，并通过标准的向量余弦距离度量主题向量之间的相关性：

$$corre(Z_i,Z_j)=corre(\beta_i,\beta_j)=\frac{\sum_{v=0}^{V}\beta_{iv}\times\beta_{jv}}{\sqrt{\sum_{v=0}^{V}(\beta_{iv})^2\sum_{v=0}^{V}(\beta_{jv})^2}} \tag{2}$$

$corre(Z_i,Z_j)$ 越小，主题之间越独立；

我们用所有主题之间的平均相似度来度量该主题结构的稳定性：

$$avg\_corre(structure)=\frac{\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}corre(Z_i,Z_j)}{K\times(K-1)/2} \tag{3}$$

定理 1. 当主题结构的平均相似度最小时，对应的模型最优.

证明.

$$\arg\min(avg\_corre(structure))=\arg\min\left(\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{\sum_{v=0}^{V}\beta_{iv}\cdot\beta_{jv}}{\sqrt{\sum_{v=0}^{V}(\beta_{iv})^2\sum_{v=0}^{V}(\beta_{jv})^2}}\right) \tag{4}$$

根据贝叶斯定理，有

$$p(w_n|Z_n)\propto p(Z_n|w_n)p(w_n),$$

所以式(4)可以转换为关于 $p(Z_i|w_v)$ 的函数. 结合 LDA 模型中的约束条件 $\sum_{i=1}^{K}p(Z_i|w_v)=1$，当所有 w_v 都在某个 Z 上取得明显的峰值时，上式达到满足条件. 该条件可以进一步表示为

$$\max\sum_{i=1}^{K}p(Z_i|w_v)^2 \tag{5}$$

同时，在 EM 算法迭代计算模型参数 α 和 $\boldsymbol{\beta}$ 的过程中，优化函数为

$$(\alpha^*,\beta^*)=\arg\max\ p(D|\alpha,\boldsymbol{\beta})=\arg\max\prod_{d_i\in D}p(d_i|\alpha,\boldsymbol{\beta}) \tag{6}$$

其中，

$$p(d|\alpha,\boldsymbol{\beta})=\prod_{n=1}^{N_d}\sum_{i=1}^{K}P(\boldsymbol{\theta}_d)P(Z_{di}|\boldsymbol{\theta}_d)P(w_n|Z_{di}) \tag{7}$$

Blei 采用变分法进行近似推演[1]，引入了两个隐含变元 γ 和 ϕ. 其中 ϕ_{ni} 表示第 n 个词由主题 Z_i 产生的概率 $p(Z_i|w_n)$. 并得到如下结论：

$$\gamma_i=\alpha_i+\sum_{n=1}^{N}\phi_{ni} \tag{8}$$

$$\beta_{ij}\propto\sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{ni}w_{dn}^{j} \tag{9}$$

将式(8)和式(9)代入式(7)，可推出式(7)满足如下关系：

$$p(d|\alpha,\boldsymbol{\beta})\propto\prod_{n=1}^{N_d}\sum_{i=1}^{K}\phi_{ni}^2 \tag{10}$$

显然，当式(5)满足时，$p(d|\alpha,\boldsymbol{\beta})$ 即可达到最大值，满足最优模型条件. 定理 1 得证. 证毕.

3.2 实验分析

下面我们用一组真实数据来观察 LDA 在指定不同参数 K 得到的模型中，主题之间相似度的变化情况. 我们在一个数据集上建立了 3 个 LDA 模型，K 分别等于 2，3，4. 该文本集合包含 4 个文档，共 7 个不同的词.

Doc1：drug clinical patients

Doc2：drug disease

Doc3：hiv virus aids

Doc4：aids hiv disease

当 $K=2$ 时,主题在词空间的分布如图 3 所示.两个主题在词“disease”上重叠,而且分布比例非常接近,导致两个主题之间相关性强,是该主题结构中一个不稳定的元素.

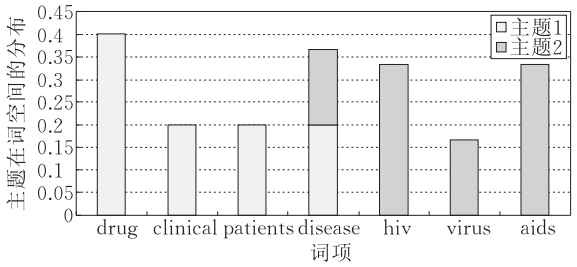


图 3　$K=2$ 时主题在词空间的分布图

表 1 为该模型下所有词在主题之间的分配(W_n 属于主题 $I=\arg\max_i(p(W_n|Z_i))$.

表 1　词在主题中的分配($K=2$)

主题	词
1	Drug clinical patients disease
2	Aids hiv virus

当 $K=3$ 时,主题在词空间的分布如图 4 所示.图 3 中的不稳定因素“disease”从主题 1 中分离出去,形成一个新的主题,如表 2 所示.在新的主题结构中,主题在词表中的分布相对 $K=2$ 的主题结构比较稳定(主题基本没有重叠).

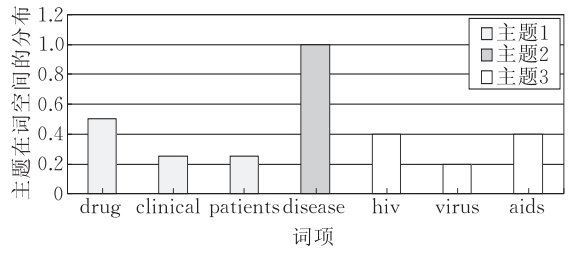


图 4　$K=3$ 时主题在词空间的分布图

表 2　词在主题中的分配($K=3$)

主题	词
1	Drug clinical patients
2	disease
3	Aids hiv virus

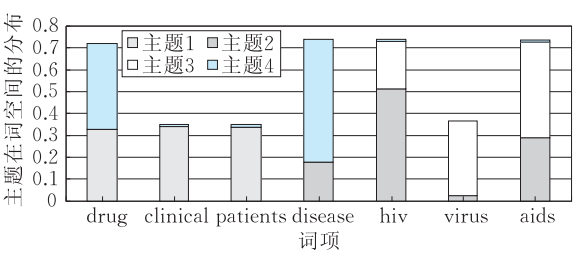


图 5　$K=4$ 时主题在词空间的分布图

从图 5 可以看出,当 K=4 时,主题分布的重叠现象比较多.而且,从表 3 中的词分配可以看出,该结构下的主题之间互异性降低,每个主题表示的意义不完整(如主题 2 和主题 3 之间相关性太大).

表 3　词在主题中的分配($K=4$)

主题	词
1	clinical patients
2	hiv
3	Aids virus
4	Disease Drug

通过上面的主题产生过程发现,新的主题通常由造成主题之间相似的词(主题重叠区域)产生.LDA 模型的产生过程,就是在已知主题数目的情况下,通过调节主题在词空间的比例,不断去除主题之间相关性的过程.

计算以上 3 个主题结构的平均相似度分别为 0.1195,0.00014 和 0.2791.显然 $K=3$ 时结构最稳定.再次验证了定理 1.

4　基于密度的最优 LDA 模型选择方法

LDA 可以在主题数目固定的条件下,找到最优的主题分布.我们的工作是在此基础上,根据第 3 节提出的定理 1,自动寻找最优主题数,达到全局最优.在第 3 节,我们已经证明了最优 LDA 模型与主题相关性之间的关系.本节,我们引入基于密度的聚类算法 DBSCAN[14] 中计算样本密度的思想来度量主题之间相关性,提出了基于密度的最优 LDA 模型选择算法.

定义 1(主题密度).　对给定的主题 Z 和距离 $Radius$,以 Z 为中心,半径为 $Radius$ 画一个圆,分别计算 Z 与其它主题之间的相似度(式(2)所示),相似度的值落在圆内的主题数目称为 Z 基于 $Radius$ 的密度,记为 $Density(Z,Radius)$.

定义 2(模型基数).　给定一个主题模型 M 和正整数 n,模型中密度小于或等于 n 的主题数目称为该模型的基数,记为 $Cardinality(M,n)$.

定义 3(参考样本).　对于主题分布中的一个点 Z、距离半径 r 和阈值 n,如果满足 $Density(Z,r)\leq n$,则称 Z 代表的词空间向量为主题 Z 的一个参考样本.

参考样本不是实际数据集中的一个文档向量,而是词空间分布上的一个虚拟点.

在以上定义的基础上,基于密度的最优模型选择算法可描述为以下过程:

1. 根据任意给定的初始 K 值,以随机抽样方式对 Dirichlet 分布的完全统计矩阵进行初始化,得到一个初始模型 $LDA(\alpha,\boldsymbol{\beta})$;

2. 将初始模型的主题分布矩阵 $\boldsymbol{\beta}$ 作为一个初始聚类结果,计算所有主题之间的相似度矩阵和平均相似度 $r=avg\_corre(\boldsymbol{\beta})$,基于 r 得到所有主题的密度 $Density(Z,r)$. 最后,设 $n=0$,计算该模型 M 的基数 $C=Cardinality(M,0)$;

3. 根据第 2 步的参考 K 值重新估计 LDA 模型参数. K 的更新函数为

$$K_{n+1}=K_n+f(r)\times(K_n-C_n) \tag{11}$$

其中 $f(r)$ 表示 r 的变化方向. 当 r 的变化方向为负时(与前一次相反),$f_{n+1}(r)=-1\times f_n(r)$;当 r 的变化方向为正时(与前一次相同),$f_{n+1}(r)=f_n(r)$; $f_0(r)=-1$.

当 $f(r)=-1$ 时,将主题从小到大按密度排序,将前 C 个主题视为参考样本,对下一次 LDA 模型参数估计的 Dirichlet 完全统计矩阵进行初始化;反之采用从集合中抽样的方式对完全矩阵进行初始化;

反复执行步 2~步 3,直到平均相似度 r 和参数 K 同时收敛.

由式(11)可知,K 收敛的条件是 $\arg\min(K_n-C_n)$. 由模型基数 C 的定义可知,C 随着平均相似度 r 的减小而增加,且 $C_n\leq K_n$. 当 r 达到最小时,C 达到最大. 所以,可以保证 r 和参数 K 同时收敛.

5 实 验

5.1 实验数据

我们在 TRECVID2005 的所有英语新闻测试集上构造了 3 个有针对性的测试集来验证前面提出的理论和算法.

测试集 D_0 是整个 2005 的英语 ASR 文本集,包含 20932 个镜头文本,8410 个不同的词项;

测试集 D_1 由 search 任务中的 0168,0160,0169 3 个查询对应的标注集组成,包含 3754 个镜头文本,5535 个不同的词,按 10∶1 分成 D_1_train 集和 D_1_test 集;

测试集 D_2 由 search 任务中的 0168,0165,0172 三个查询对应的标注文档集合组成,包含 4129 个镜头文本,5681 个不同的词,按 10∶1 分成 D_2_train 集和 D_2_test 集;

其中 5 个 search 任务的查询分别为

0160="Find shots of something (e.g., vehicle, aircraft, building, etc) on fire with flames and smoke visible";

0165="Find shots of basketball players on the court";

0168="Find shots of a road with one or more cars";

0169="Find shots of one or more tanks or other military vehicles";

0172="Find shots of an office setting, i.e., one or more desks/tables and one or more computers and one or more people";

由上可知,数据集 D_1 和 D_2 的大小基本相同,但 D_1 集合中的文档之间具有更强的相关性. D_0 比 D_1 和 D_2 大,但集合中包含很多噪声.

我们采用语言模型中标准的评判准则困惑度(perplexity)评价各种 LDA 模型的性能[1]. 我们在训练集合上训练得到最优 LDA 模型;通过计算一个给定的测试集合的困惑度可以评价该模型产生文本的能力. 困惑度越低,说明模型具有更好的推广性. 对于一个具有 M 个文档的测试集 D_{test},困惑度计算公式如下

$$perplexity(D_{\text{test}})=\exp\left\{-\frac{\sum_{d=1}^{M}p(d_d)}{\sum_{d=1}^{M}N_d}\right\} \tag{12}$$

其中,N_d 为文本 d 的长度;$p(d_d)$ 是待测试模型产生文档 d_d 的概率.

同时,我们采用第 3 节介绍的平均相似度(式(3))表示主题结构的相关性.

5.2 主要实验结果

针对前面提出的观点,我们分别设计了 3 组实验.

实验 1. 我们针对不同数据集的最优 K 值选择问题进行了两组对比实验.

图 6 表示的是尺度相同,但集合内文档之间相关度不同的集合的最优 K 值比较;集合 D_1 在 $K=100$ 时得到最优性能,而 D_2 在 $K=80$ 时得到最优性能. 说明当集合中的文本相关性强时,LDA 模型需要更多的主题(主题越具体,可以减小主题之间的相关性);而且,D_1 的困惑度曲线整体高于 D_2(困惑度越大,说明模型推广能力越差),说明当集合内部相关程度高时,LDA 表示数据的能力下降. 再一次验证 LDA 不考虑主题之间相关性的假设限制了 LDA 精确表示数据的能力.

同时,我们对尺度相差很大的文档集合 D_0 也做了最优 K 值选择的实验,分别测试了 K 为 10,30,50,100,200,300,400 和 500 的情况,当 $K=30$ 时达到最优性能. 进一步证明,最优 K 值的选择不只是与集合尺度有关,而且对集合内的相关程度敏感.

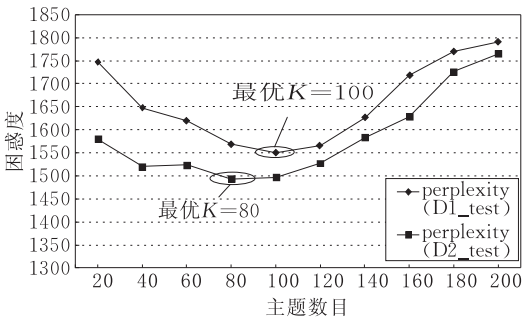


图 6　D_1 和 D_2 模型在不同主题数目下的 Perplexity 结果比较

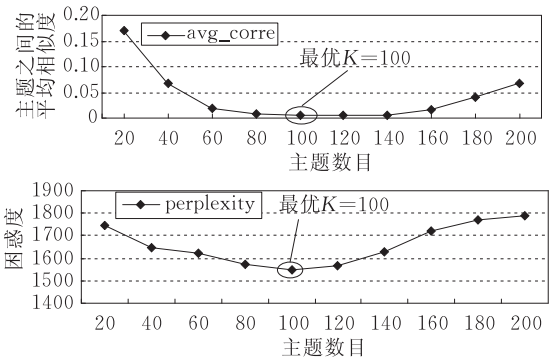


图 7　ave_corre曲线和 perplexity曲线在不同 K 值下的结果比较

实验 2.　我们以集合 D_1 为测试数据,分析了最优 K 值与主题之间平均余弦距离的关系.

通过比较 *avg_corre* 曲线和 *perplexity* 曲线在不同参数 K 下的变化趋势发现,两者的变化趋势基本相同,且当 *avg_corre* 达到最小值($K=100$)时,对应的模型达到最佳性能.

实验 3.　采用我们提出的基于密度的最优模型选择算法,在 D_1 上进行了 6 组实验.分别为 K 赋初值 10,50,100,200,300 和 500,经过若干次迭代后,基本都可以找到最优 K 值($K=100$).初值越接近最优 K 值,迭代次数越少.结果如表 4 所示.

表 4　迭代寻找最优 K 的算法结果

初始 K 值	最优 K 值	迭代次数
10	97	19
50	99	26
100	102	2
200	109	3
300	106	5
500	102	34

6　结　论

为了更准确的描述和分析大规模文本数据,各种复杂的主题模型不断出现,但是如何选择最优的主题数目仍然是这些模型共同面临的难题.在本论文里,我们对最优 K 值与主题之间的相关性进行了深入分析,提出当主题之间平均余弦距离最小时,模型达到最优性能.我们将最优 K 值选择与模型参数估计统一在一个框架里,提出了一种新的基于密度的最优 K 值选择方法.实验证明该方法可以在不需要人工指定主题数目的情况下,自动找到 LDA 模型中的最优主题结构.

参　考　文　献

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

[2] Blei D, Lafferty J. Correlated topic models//Weiss Y, Schölkopf B, Platt J eds. Advances in Neural Information Processing Systems 18. Cambridge, MA: MIT Press, 2006

[3] Li W, McCallum A. Pachinko allocation: DAG-structured mixture models of topic correlations//Proceedings of the International Conference on Machine Learning (ICML). 2006

[4] Xing E, Yan R, Hauptmann A. Mining associated text and images with dual-wing harmoniums//Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05). 2005

[5] Li F-F, Perona P. A bayesian hierarchical model for learning natural scene categories//Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). Washington, DC, USA, 2005: 524-531

[7] Deerwester S, Dumais S, Furnas G, Lanouauer T, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990, 41: 391-407

[6] Wei X, Croft W B. LDA-based document models for ad-hoc retrieval//Proceedings of the 29th SIGIR Conference. 2006: 178-185

[9] Yang J, Liu Y, Xing E P, Hauptmann A. Harmonium-based models for semantic video representation and classification//Proceedings of the 7th SIAM International Conference on Data Mining. 2007

[10] Xing E, Yan R, Hauptmann A. Mining associated text and images with dual-wing harmoniums//Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05). 2005

[11] Barnard K, Duygulu P, Forsyth D, de Freitas N, Blei D, Jordan M. Matching words and pictures. Journal of Machine Learning Research, 2003, 3:

[12] Teh Y, Jordan M, Beal M, Blei D. Hierarchical dirichlet processes. Journal of the American Statistical Association, 2007, 101(476): 1566-1581

[8] Hofmann T. Probabilistic latent semantic indexing//Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 1999: 50-57

[13] Li W, McCallum A. Nonparametric bayes pachinko allocation//Proceedings of the UAI. 2007

[14] Ester M, Kriegel H P, Sander J, Xu X. A density based algorithm for discovering clusters in large spatial databases with noise//Simoudis E, Han J W, Fayyad U M eds. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland: AAAI Press, 1996: 226-231

CAO Juan, born in 1980, Ph. D. candidate. Her research interests focus on multimedia retrieval and machine learning.

ZHANG Yong-Dong, born in 1973, Ph. D., associate professor. His major field includes image processing and video processing.

LI Jin-Tao, born in 1962, professor, Ph. D., supervisor. His major field includes multimedia processing and VR technology.

TANG Sheng, born in 1972, Ph. D., assistant researcher. His major field includes multimedia retrieval and video processing

Background

Statistical topic models such as Latent Dirichlet Allocation (LDA) have been successfully used to analyze large amounts of textual information in many tasks, including language modeling, document classification, information retrieval, document summarization and data mining. These models can capture word correlations in a collection of textual documents with a low-dimensional set of multinomial distribution, called "topics". To further model the inter-topic correlations, recent advances such as Correlated Topic Model (CTM) in this area have explored richer structures to discover large numbers of more accurate and fine-grained topics. But all these models have the same practical difficulty to determine the number of topics. Model selection methods such as cross-validation and Bayesian model testing are usually inefficient. Teh et al. propose the Hierarchical Dirichlet Process (HDP) to solve the problem. Dirichlet process does not require specifying the number of mixture components in advance, and the HDP can realize the share of the mixture components among a set of mixture models.

This paper is Supported by the National High Technology Research and Development Program (863 Program) of China under grant No. 2007AA01Z416; the National Basic Research Program (973 Program) of China under grant No. 2007CB311100; the National Natural Science Foundation of China under grant No. 60773056; the Beijing New Star Project on Science & Technology under grant No. 2007B071.