基于频繁词网络的 LDA 最优主题个数选取方法

李菲菲,王移芝
(北京交通大学 计算机与信息技术学院,北京 100044)

摘 要:LDA(latent Dirichlet allocation,隐含狄利克雷分布)主题模型被广泛应用于大规模文档处理,通常用于主题提取、情感分析和文本降维等。这些模型使用类似期望最大算法从文档集合中提取低维语义分布,并将每一维分布有效结合,形成主题。在模型构建过程中,初始主题数 K 对迭代过程与结果非常重要。针对这一问题,根据文档聚类簇数(即社区个数)与文档集隐含主题数相一致的特点,提出了一种以频繁词集网络的社区划分个数用来指定 LDA 主题模型主题输入个数的方法。该方法对文档构建频繁词对,并以此为基础构建词共现网络,然后采用无监督社区划分算法对该词共现网络进行社区划分,并以划分的社区个数作为 LDA 主题模型的主题个数。实验结果表明,该方法可以自动化指定主题个数 K,显著提升主题查准率和查全率,主题独立性更强。

关键词:隐含狄利克雷分布;主题模型;频繁词网络;聚类;社区划分
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2018)08-0001-05
doi:10.3969/j.issn.1673-629X.2018.08.001

Selection Method of LDA Optimal Topic Number Based on Frequent Word Network

LI Fei-fei, WANG Yi-zhi
(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: LDA topic model is widely used in large-scale document processing and usually used for topic extraction, emotional analysis and text reduction. These models use the similar expectation maximum algorithm to extract the low-dimensional semantic distribution from the document collection, and effectively combine each dimension distribution to form the topic. In the model building process, the initial topic number K is very important for the iterative process and result. In order to solve this problem, according to the characteristics that the number of frequent words implied in the network community is consistent with the implied topics of document sets, we propose a method to specify the number of inputs for LDA topic model based on the number of community partition in the frequent word set network. This method builds frequent word pairs of documents, based on which the word co-occurrence network is constructed. And then, the unsupervised community partition algorithm is used to partition the co-occurrence network, and the number of communities is used as the number of topics in the LDA topic model. The experiment shows that this method can automatically specify the number of topic number K, which significantly improves the precision and recall of topic and makes the independence of topic stronger.

Key words: LDA; topic model; frequent word network; clustering; community partition

0 引 言

随着移动互联网的快速发展,网络信息量特别是文本信息呈指数增长,因此如何精准有效地挖掘、组织和利用海量文本背后的有用信息成为一个热门话题。文本聚类技术作为自然语言处理(natural language processing, NLP)的预处理步骤,对文本进一步分析和处理有着重要影响,比如信息检索、生成文档摘要等,在文本聚类方面,主题聚类[1]方法比传统方法更有效。于是,隐含狄利克雷分布(latent Dirichlet allocation, LDA)在挖掘文档中的隐含主题方面得到了越来越多的应用。

LDA[2]主题模型在 2003 年由普林斯顿大学的 David M. Blei 等提出,该模型是一种文本建模方法,能够将每篇文档的主题以概率分布的形式给出,用来识别大规模文本或者视频中隐含的主题信息,在信息检索、文本分类等领域应用非常广泛。目前普遍认为应

收稿日期:2017-09-10 **修回日期**:2018-01-11 **网络出版时间**:2018-04-28
基金项目:国家自然科学基金(K13A300050)
作者简介:李菲菲(1991-),女,硕士,研究方向为移动互联网与机器学习;王移芝,教授,研究方向为计算机网络与数据库技术。

用基于吉布斯采样[3]的LDA的最大问题是无法确定最优主题数目,然而主题数目的选取直接影响到LDA主题模型的性能,导致文档分布表示存在精度误差。

在大规模语料库中,文档集的聚类簇数通常与句子集中隐藏的主题数目一致,前者作为后者的先验知识。根据频繁词网络[4]隐含社区个数与文档集隐含主题数相一致的特点,文中提出了一种以频繁词集网络的社区划分个数用来指定LDA主题模型主题输入个数的方法,使得指定主题数更加接近文档分布表征的隐含主题数目。

1 相关工作

1.1 研究现状

大量研究表明,LDA主题抽取效果[5]与潜在主题的数量直接相关,主题抽取的结果对主题数目非常敏感。在此基础上,国内外大量学者展开了相关研究,通过若干方法确定最优主题数目,常用的有以下几种方法:

(1)Griffiths等提出应用贝叶斯模型[6]确定最优主题结构的方法[7]。该方法依赖于吉布斯采样的过程,但是计算非常耗时,在实际应用中效果不佳。

(2)层次狄利克雷过程(hierarchical Dirichlet processes,HDP)[8]将主题数目进行非参数化处理,它与LDA主题模型的差别在于:该模型可以通过训练得到最优主题个数,并不需要其他参数。HDP虽然通过狄利克雷过程的非参数特征解决了LDA中主题数目的选取问题[9],但是该方法要求为同一个集合同时建立两个模型:HDP模型和LDA模型;并且算法时间复杂性较高,在具体应用中效率不高。

通过分析以上方法,笔者认为针对LDA最优主题数目的确定,主要存在时间复杂性较高、效率低等问题。基于此,提出一种以频繁词网络的社区划分结果作为文档聚类簇数,并以该簇数确定LDA主题个数K的方法。

1.2 LDA主题模型

LDA主题模型是一种基于EM算法的经典语义分析模型,能够对高维文档降维并提取隐含语义信息。它以词袋(BOW)模型为基础,这种模型将文档转化为一种词向量,有效地将复杂的语义知识转变为易于理解的数字信息。并且该模型的词袋方法并未考虑到词语之间的次序关系,这就简化了模型,使得问题变得简单,也为后人改进LDA模型提供了思路。LDA主题模型在自然语言处理领域中是一种能够进行文本语义挖掘的统计模型[10]。它可以用来发现文档隐藏的主题,将词项空间表达的文档约简为主题空间的低维表示,并实现信息检索、文本分类等。该模型是一种可以随机生成可观测数据的文本建模方法,首先拆解文档为词的集合,通过计算得到文档在主题上的低维向量分布,并根据主题分布对文档分类或计算主题相关性等。关于语料库中的每一篇文章,LDA定义了以下生成过程:

(1)对语料库中的每篇文档,从主题分布中抽取一个主题;

(2)从步骤1中被抽到的主题所对应的单词分布中提取一个单词;

(3)重复步骤1和2,直到覆盖所有单词。

LDA代表的概率模型为:

$$p(\theta,Z,W\mid\alpha,\beta)=p(\theta\mid\alpha)\prod_{n=1}^{N}P(Z_n\mid\theta)P(W_n\mid Z_n,\beta) \tag{1}$$

通过对LDA主题模型的深入讨论可知,LDA模型主要是针对特定语料库,并人工指定语料级别参数α和β,通过一种半监督学习来确定模型,并根据模型生成文档。其中α和β所表示的详细信息如下:

α:分布$p(\theta)$需要一个向量参数,即狄利克雷分布的参数,用于生成一个主题向量;

β:每个主题对应的单词概率分布矩阵表示为$p(w\mid z)$。

1.3 确定最优主题数

由于主题个数能够影响到LDA主题模型的主题抽取查准率、查全率以及主题独立性,如果直接指定主题个数,则可能使得LDA不能得到较好的聚类效果,使其在某些领域的应用效果往往差强人意。所以,需要首先确定最优主题个数K。

文中方法以频繁词网络[11]为基础,其基本思想为:在大规模文档集中,如果两个词(例如苹果与乔布斯、)频繁出现在同一篇文档或同一个句子中,则认为这两个词是语义相关的。共同出现的一组词,通常会被用来表述同一个主题。构建频繁词网络可分为三个步骤:数据预处理[12]、频繁词集挖掘、构建频繁词网络。数据预处理部分主要对文章进行分词[13],并过滤掉一些停用词和常见的噪音,利用频繁词集挖掘算法挖掘文本中的频繁词集,然后利用频繁词集中词的共现关系构建词共现网络[14]。最后采用社区发现算法对词共现网络进行社区划分[15],划分出的每一个社区就是一个主题。具体过程如下文所述。

1.4 FWN频繁词网络算法

1.4.1 数据预处理

文档数据的预处理包括对文本分词及过滤停用词、噪声等[16]。

(1)分词。

中文语义理解中,词是最小的语义表达单位。但

是中文句子与英语或其他语言有所不同,单词与单词之间并没有分隔符,所以中文分词相对比较困难,因此,分词是文本处理中的关键一步。分词的方法有很多种,中文分词工具也有很多种,比如中科院的 NLPIR/ICTCLAS 汉语分词系统、Java 开源工具 Jcseg 等等。

文中使用的是经典汉语分词工具(开源)IKAnalyzer。IKAnalyzer 结合词典分词和文法分析算法的中文词组组件,对中文分词支持较好。它具有以下特性:分词速度较快,对大规模文本分词具有重要意义;可以对多种格式的数据进行处理(中文、英文、数字、日期等);支持用户词典扩展定义。

(2)过滤停用词、噪声。

停用词是指在自然语言处理中具有一定作用,但是又没有什么实际意义的词语。这些词通常以较高的频率出现,从而影响文本的处理。此外,文本中经常会出现一些高频词,如“了”、“的”、“是”、“也”等,这些高频词会对文本造成强烈的干扰,因此也应该和停用词一块儿过滤掉。文中采用搜狗发布的停用词库作为基本停用词库。

1.4.2　挖掘频繁词集

频繁项集是文本挖掘领域内一项重要的处理手段。该模型假设,在大规模的文档集合中,如果某些词频繁出现在同一篇文档或句子中,则可以近似认为这些词是语义相关的,并且它们出现的频次越高,相关程度越高。共同出现的一组词,通常会用来表述同一个主题。例如,在同一时间段的文本中,“数学”、“政治”、“英语”、“专业课”这组词以较高频率共同出现在同一篇文章中,包含这四个词的文本很可能和考研这一主题相关。

文中使用 FP-GROWTH 算法[17]挖掘 k 频繁词集。在进行频繁词集的挖掘时,根据输入教育类数据的文本数量来调整 SUP 支持度的大小。k 为 1 的频繁词集没有体现出特征词的共现关系,因此将其忽略,k 为 2 的频繁词集包含了大量噪声,故对教育文本进行观察时发现,绝大部分文本至少有三个特征词,因此将 k 为 2 的频繁项集忽略,只保留 k 大于或等于 3 的频繁词集。

1.4.3　构建频繁词共现网络

利用 $k(k \geqslant 3)$ 频繁词集中的共现关系构建网络,得到基于频繁词集的词共现网络。社区(community)是复杂网络中一种常见的现象,社区内节点联系紧密,社区间节点联系稀疏。在社交网络中,社区表现为一系列具有相似兴趣的用户群体;在科学引文网络中,社区表现为描述相似内容的引文集合;在生物医学网络中,社区表现为描述同一病症的 DNA 或染色体集合。而在以频繁词集为基础构建的词共现网络中,同一社区内的词,通常描述同一主题,即主题以社区的结构出现。

文中对文档集中的 k ($k \geqslant 3$)频繁词集构建频繁词共现网络,即当两个词出现在同一频繁词对时,将这两个词量化为图中的两个点,并认为这两个点之间有一条边。这样,根据频繁词对构建了文档集对应的频繁词共现网络。

1.5　基于模块度的社区划分

文中使用基于模块度[18]的社区划分算法(BGLL)[19]对词共现网络进行社区划分。作为经典的社区划分算法,BGLL 有如下优点:

(1)BGLL 是一种无监督社区划分算法,不需要指定社区数目;

(2)相对于 Fast-Newman[20]等算法,BGLL 算法可以处理百万级节点。文中以 1.4.3 生成的频繁词共现网络为基础,采用 BGLL 算法对该词共现网络进行社区划分。

2　实验与分析

2.1　实验数据

实验数据来源于实验组网络抓取,采用网络爬虫技术在 20 余个教育网站上抓取各类教育信息(文章),共计 79 463 条数据。按内容完整性、主题性强弱、主题分布是否均匀等原则,筛选出 3 318 条数据作为实验数据,其中 10% 的数据集用于测试集评估模型,剩余数据用于训练模型。

通过对文档集标题、简介、关键词等数据的分析,课题组对 3 318 条数据进行主题打标签,专家鉴定,共获得主题 22 个。

2.2　实验过程与实验结果

由于课程简介最能体现出课程的特点,因此以课程简介作为语料库进行实验。首先进行频繁词集的挖掘,选取最小支持度 MINSUP 为 0.01,频繁词集挖掘部分结果见表 1。

表 1　频繁词集部分结果

编号	频繁词集	SUP
1	重点 基础 讲解	24
2	课程 提高 掌握	23
3	知识 大纲 考点	23
4	课程 部分 视频	24
5	教材 老师 大纲	25
6	强化 基础 司法考试	24
7	学习 强化 考研	27
8	设计 开发 技术	25

续表 1

编号	频繁词集	SUP
9	学习 方法 获得	23
10	课程 老师 知识点	29
11	技巧 老师 主讲	27
12	课程 学习 语法 知识	24

然后对频繁词集构建词共现网络,使用 BGLL 对词共现网络进行社区划分。取模块度为 0.48 时(模块度最大)的社区划分方式作为最后的实验结果,此次实验共发现 18 个社区,故以 18 作为 LDA 主题个数的指定。其中部分主题结果如表 2 所示。

表 2　部分主题结果

主题	主题词
考研	英语 培训 数学 肖秀荣 重点 理解
求职应聘	开发 工程师 财务 运维 产品
英语	托福 口语 听力 雅思 留学 四级
计算机	操作 比尔盖茨 系统 云计算 数据库
成人自考	学历 证书 专业 毕业 报名 成人

除了使用 Fast-Newman 算法对词共现网络图做社区划分,文中则采用 BGLL 对词共现网络进行社区划分,其中 Fast-Newman 算法得出的社区个数为 16,BGLL 算法得出的社区个数为 18 个。

以查准率和查全率来评价主题抽取效果,查准率用于评价 LDA 主题抽取正确主题数占所有抽取主题数的比例,查全率用于评价 LDA 主题抽取正确主题数占专家预先标注的主题个数。

公式如下:

$$P = \frac{T_{\text{correct}}}{T_{\text{extract}}} \tag{2}$$

$$R = \frac{T_{\text{correct}}}{T_{\text{standard}}} \tag{3}$$

其中,T_{extract} 表示 LDA 抽取的有效主题数;T_{correct} 表示 LDA 抽取的正确主题数;T_{standard} 表示专家标注的主题个数。

不同主题个数下对应的查准率与查全率如表 3 所示。

表 3　不同主题个数下的查准率和查全率

主题数	T_{extract}	T_{correct}	T_{standard}	查准率/%	查全率/%
10	9	6	22	66.7	27.3
16	14	10	22	71.4	45.4
18	16	14	22	87.5	63.6
20	17	14	22	82.4	63.6
30	27	16	22	59.3	72.7
40	17	11	22	50.0	77.3

由表 3 可知,通过 BGLL 算法得出的主题个数 k(18)略好于 Fast-Newman 算法得出的主题个数(16),具有最高的查准率和较高的查全率,效果相对较好。

在概率语言模型中,困惑度[21]是一种用来评估语言模型优劣的指标,其基本思想是给测试集赋予较高概率值的语言模型,且较小的困惑度意味着模型对新文本具有较好的预测作用,并且在一定程度上,困惑度会随着潜在主题数量的增加呈现递减的规律。文中以困惑度评价 LDA 主题模型的优劣。在 LDA 主题模型中,困惑度的计算公式如下所示:

$$\text{Perplexity}(D) = e^{\{-\sum_{d=1}^{M}\log^{p(w_d)}/\sum_{d=1}^{M}N_d\}} \tag{4}$$

其中,D 表示待训练文档数据集,一共有 M 篇文档;N_d 表示每篇文档 d 中的单词数量;w_d 表示文档 d 中的词;$p(w_d)$ 表示词 w_d 在文档 w 中的频率。

实验取主题个数范围为[10,200],步长为 10,分别在训练数据上训练模型,最终的困惑度-主题个数曲线如图 1 所示。

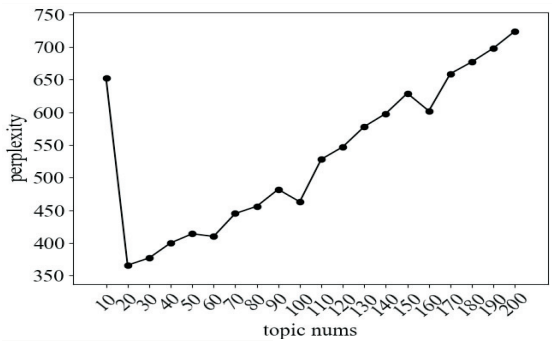

图 1　不同 k 值下的 LDA 模型困惑度

从图 1 可以看出,当主题数目 k 为 19 时,LDA 的困惑度指标达到最小值,这与文中实验的结果 18 个主题较为接近,从而证明了文中方法的有效性。

2.3　对照实验

按照文献[5]科技情报分析中 LDA 主题模型最优主题个数确定方法的研究,参数设置和算法过程如文中所述。不同 k 值下的 LDA 模型方差值如图 2 所示,不同 k 值下的 LDA 模型 Perplexity-Var 值如图 3 所示。

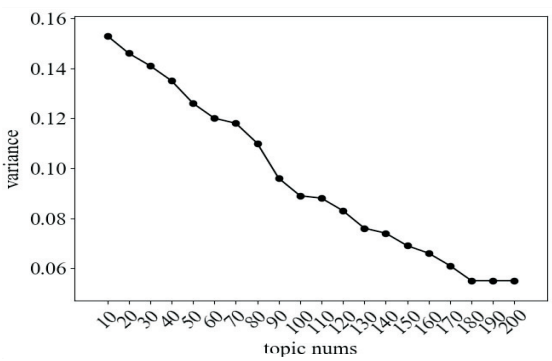

图 2　不同 k 值下的 LDA 模型方差值

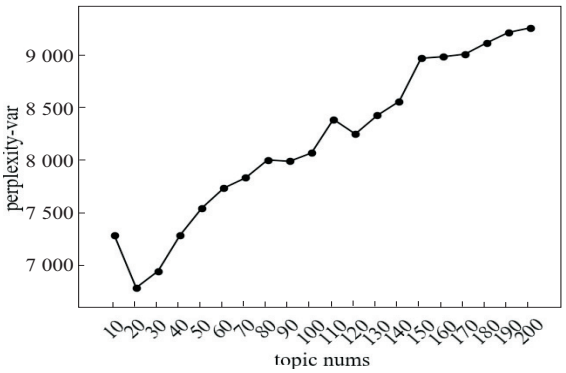

图 3　不同 k 值下的 LDA 模型 Perplexity-Var 值

由图 3 可知,k 取 20 时 LDA 主题效果最好,这与文中方法取得的最优 k 值 18 基本一致,验证了该方法的正确性。

3　结束语

在大数据背景下,为了更准确地分析和处理海量文本数据,根据频繁词网络隐含社区个数与文档集隐含主题数相一致的特点,提出了一种以频繁词集网络的社区划分个数用来指定 LDA 主题模型主题输入个数的方法。实验结果表明,该方法可以以一种无监督聚类方式得到文档聚类簇数并对主题数目做指定,显著提升了主题查准率和查全率,并得到较好的聚类效果,相对智能地得到最优的主题分布。但文中只是针对教育网站上抓取的数据进行了实验分析,并未针对其他类型的数据集进行方法的验证,如微博新闻、短文本文档等。接下来,将着重于在多种数据集上验证该方法的正确性。

参考文献:

[1] LAI J Z C,HUANG T J,LIAW Y C. A fuzzy k-means clustering algorithm using cluster center displacement[J]. Pattern Recognition,2009,42(11):2551-2556.

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

[3] HEINRICH G. Parameter estimation for text analysis[R]. Darmstadt,Germany:Fraunhofer IGD,2004.

[4] 王永恒,贾　焰,杨树强. 基于频繁词集聚类的海量短文分类方法[J]. 计算机工程与设计,2007,28(8):1744-1746.

[5] 关　鹏,王日芬. 科技情报分析中 LDA 主题模型最优主题数确定方法研究[J]. 现代图书情报技术,2016,32(9):42-50.

[6] GRIFFITHS T L,CHATER N,KEMP C,et al. Probabilistic models of cognition:exploring representations and inductive biases[J]. Trends in Cognitive Sciences,2010,14(8):357-364.

[7] NELSON G C,VALIN H,SANDS R D,et al. Proceedings of the national academy of sciences of the united states of America (PNAS)[J]. Powder Metallurgy & Metal Ceramics,2013,16(1):63-65.

[8] 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.

[9] 曹　娟,张勇东,李锦涛,等. 一种基于密度的自适应最优 LDA 模型选择方法[J]. 计算机学报,2008,31(10):1780-1787.

[10] 邹晓辉,孙　静. LDA 主题模型[J]. 智能计算机与应用,2014,4(5):105-106.

[11] CHOW C K,KANEKO T. Automatic boundary detection of the left ventricle from cineangiograms[J]. Computers & Biomedical Research,1972,5(4):388-410.

[12] 陈宝树,党齐民. Web 数据挖掘中的数据预处理[J]. 计算机工程,2002,28(7):125-127.

[13] 龙树全,赵正文,唐　华. 中文分词算法概述[J]. 电脑知识与技术,2009,5(10):2605-2607.

[14] 刘则渊,尹丽春. 国际科学学主题共词网络的可视化研究[J]. 情报学报,2006,25(5):634-640.

[15] 刘志雄,贾彩燕. 面向用户兴趣与社区关系的微博话题检测方法[J]. 智能系统学报,2016,11(3):294-300.

[16] 李　伟. 基于频繁词集词共现网络的短文本聚类方法[D]. 北京:北京交通大学,2016.

[17] HAN Jiawei,PEI Jian,YIN Yiwen,et al. Mining frequent patterns without candidate generation:a frequent-pattern tree approach[J]. Data Mining & Knowledge Discovery,2004,8(1):53-87.

[18] 刘绍海,刘青昆,谢福鼎,等. 复杂网络基于局部模块度的社团划分方法[J]. 计算机工程与设计,2009,30(20):4708-4710.

[19] CHATURVEDI P,DHARA M,ARORA D. Community detection in complex network via BGLL algorithm[J]. International Journal of Computer Applications,2012,48(1):32-42.

[20] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E,2004,69(6):066133.

[21] 贺　亮,李　芳. 基于话题模型的科技文献话题发现和趋势分析[J]. 中文信息学报,2012,26(2):109-115.