·CTCIS 2016 推荐论文· DOI:10.15961/j.jsuese.201601032

基于 ICE-LDA 模型的中英文跨语言话题发现研究

陈兴蜀[1,2],罗 梁[2],王海舟[1,2],王文贤[1,2*],高 悦[2]
(1.四川大学 网络空间安全研究院,四川 成都 610065;2.四川大学 计算机学院,四川 成都 610065)

摘 要:近年来互联网在全球化的大背景下飞速发展,针对跨语言的网络数据挖掘成为国内外舆情分析的热点问题,有效实时地检测中英文网络环境下的热点话题对舆情的掌握和舆情的发展有着至关重要的作用。网络新闻作为网络信息舆情中的重要组成部分,由于互联网的大规模普及而成为人们方便快捷获知信息的重要来源。首先,本文选择中文与英文的网络新闻作为数据源进行采集,提出了在 LDA 模型上改进的 ICE-LDA 模型进行跨英汉语言网络环境下的共现话题发现。采用话题向量化的方式,对建模产生的话题进行 JS 距离检测和话题文本分布相似度度量。其次,本文分别对爬虫采集到的中英混合新闻数据分别构建可对比平行语料集和非可对比语料集进行话题建模,在建模过程中利用 TF-IDF 算法对文档提取特征词去噪,提高话题特征表示去除无意义噪音词。最后,分别采用两种不同的话题向量化方式进行跨语言的共现话题发现建模。实验结果表明,在本文设计的爬虫采集构建的真实数据集上,改进后的话题模型不仅能够在不需要先验话题对的情况下对可对比语料集进行跨语言共现话题进行发现,而且能够对语料不平衡的情况进行共现话题发现。

关键词:话题发现;跨英汉文本;ICE-LDA 模型;TF-IDF 特征提取;共现话题
中图分类号:TP391 **文献标志码**:A **文章编号**:2096-3246(2017)02-0100-07

Analysis and Research on Cross Language Topic Discovery in Chinese and English

CHEN Xingshu[1,2], *LUO Liang*[2], *WANG Haizhou*[1,2], *WANG Wenxian*[1,2*], *GAO Yue*[2]
(1. Cybersecurity Research Inst., Sichuan Univ., Chengdu 610065, China; 2. College of Computer Sci., Sichuan Univ., Chengdu 610065, China)

Abstract: With the rapid development of the Internet under the background of globalization, mining network data for cross-language texts has become one of the most popular research fields in public opinion analysis. Detecting hot topics effectively and timely for texts both in Chinese and English plays a crucial role in grasping the development of public opinion. Internet news, as an important part of the Internet public opinion, has become a significant source of information acquisition for netizens. Firstly, Internet news in Chinese and English network were collected. Secondly, the ICE-LDA model based on LDA model was proposed to detect co-occurrence topics of the mixed dataset. Then, the JS distance and cosine similarity of the topic-text distribution were used to calculate the distance between two topics in ICE-LDA model. Thirdly, a contrastive parallel corpus and a non-colligative corpus were constructed respectively for Chinese and English mixed news data. During model building, the TF-IDF algorithm was used to remove noise words of the text. Finally, two kinds of topic vectors were used to detect the co-occurrence topics. The experimental results showed that the improved topic model proposed by us can not only detect topics in the comparison corpus dataset but also in the non-comparison corpus dataset.

Key words: topic model; cross language; ICE-LDA model; TF-IDF feature word extraction; co-occurrence topic

随着近年来互联网的快速普及和蓬勃发展,截至2015年12月,中国网民规模达6.88亿,互联网普及率达到50.3%,网络新闻的用户规模达到5.64亿,网民使用率为82%[1],由此可见通过新闻门户

收稿日期:2016-09-18
基金项目:国家科技支撑计划资助项目(2012BAH18B05);国家自然科学基金资助项目(61272447);四川大学青年教师启动基金(2015SCU11079)
作者简介:陈兴蜀(1968—),女,教授,博士生导师. 研究方向:信息安全;云计算安全等. E-mail:chenxsh@scu.edu.cn
*通信联系人 E-mail:catean@scu.edu.cn

网站发布的网络新闻已经成为人们日常生活不可或缺的信息传播和获取的途径。

网络舆情,指公众在互联网上公开表达的对某种社会现象或社会问题具有一定影响力和倾向性的共同意见[2]。当今世界已经进入到全球化的时代,公众对于新闻热点的兴趣与关注范围往往是世界性的,如“MH370 航班的失联”“香港‘占中’”等事件在国内外均得到广泛持续关注,产生了深远影响。所以,针对跨语言新闻的共现话题进行及时挖掘发现,对于舆情动态的及早监控显得尤为重要。

早期的文本聚类主要是在单语言环境下进行,多语言文本聚类技术依托于传统的聚类技术,同时适应了多语言的信息环境,能够较好地满足人们跨语言环境的信息需求[3]。对于多语言文本,先转换再聚类的方法是目前研究的主流[4],分为基于翻译的方法与基于语义分析的方法。而对文本特征提取之后再翻译更加合理,因为翻译工具在词语方面的翻译性能和效果显然优于全文翻译[5]。Dumais 等使用潜在语义标引对英法双语,利用了英法平行语料作为训练集产生语义空间[6]。Wim 等提出用 LDA 话题模型的方法构建英语和荷兰语这两种语言的中间语 LDA 模型,通过语料训练获取中间话题模型,通过选取话题分布和命名实体分布中最大值构成相异度函数对跨语言文本进行聚类[7]。Ni 等针对 Wiki 百科英汉双语语料库提出 ML-LDA 模型[8],通过训练集得到广义空间上的话题单元,但是没有针对实际新闻语料集进行话题发现。陆前等也在中英文可比语料库的基础上提出了 CLU-LDA[9]模型,将两种语言分开建模采样求得跨语言的联合话题分布,在可比语料库上取得了不错的效果。但 CLU-LDA 模型,发现共现话题需要依靠先验的话题对才能发现跨英汉语言空间的共现话题,并且该模型只针对可对比语料集做话题发现,而针对无先验话题对和可对比语料集的开放型英汉混合数据集实验则难以发现跨语言语料集中的共现话题。高盛祥等基于全局局部话题对在时间序列上的贡献,对汉越双语新闻事件进行检索[10]。

本文提出了一种面向中英文混合文本热点话题发现的改进模型——ICE-LDA(improved Chinese-English LDA)。该模型采用话题向量化的方式,针对实际采集的中英文语料集进行话题发现的同时,能够不依赖先验话题对进行跨语言共现话题发现,同时对实际情况中更多的非平行语料集也能够通过计算话题向量相似度进行话题对齐。

1　话题模型介绍

1.1　LDA 话题模型简介

话题模型是一种无监督学习模型,能够利用大量现有的互联网数据,产生的话题易于人类理解,能够发现文档集中隐含的语义特点[11]。

概率话题模型起源于潜在语义索引 LSI[12] 以及随后出现的概率潜在语义索引 pLSI[13]。Blei 等在 2003 年首次提出了 LDA(latent Dirichlet allocation)[14]模型。

LDA 属于文档生成模型。Griffiths 等对话题 - 单词分布施加了 Dirichlet 先验分布的假设[15],从而使 LDA 成为一个完全的概率生成模型如图 1 所示。

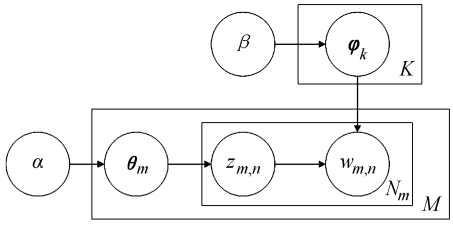

图 1　LDA 模型

Fig. 1　LDA model

其中,α 表示文档 - 话题分布的先验参数,β 表示话题 - 单词分布的先验参数。在 LDA 话题模型中,超参数 α 和 β 都是固定值,由用户事先指定,文档 - 话题分布 θ_m 和话题 - 单词分布 $\boldsymbol{\varphi}_k$ 是 LDA 话题模型中需要求解的参数,k 表示话题总数,M 表示文档总数。

1.2　ICE-LDA 模型

由于 LDA 模型是对文本的潜在语义关系进行建模,对于不同语言空间处理乏力,因此本文提出了一种的面向中英文混合文本热点话题发现的改进模型 ICE-LDA。该模型将话题分布做了向量化处理,能够在不利用先验知识的情况下发现中英文文本中存在的共现话题。

图 2 为本文提出的 ICE-LDA 模型架构图。

其中,α、β、γ 表示先验分布的超参数,$\boldsymbol{\varphi}_{k,\mathrm{C}}$ 和 $\boldsymbol{\varphi}_{k,\mathrm{E}}$ 分别代表两种语言集上的单词 - 话题分布,下标 C 表示中文语料集,E 表示英文语料集,K 代表话题个数,$\boldsymbol{\theta}$ 代表经过话题距离计算后的共现话题对分布,$\boldsymbol{\varphi}$ 代表经过距离计算后的共现话题对,Z 与 W 分别代表话题编号与单词。

模型的基本思想是对已有的数据集进行文档语言空间进行标注后进行建模求解,即在已知文档属于英文或中文的情况下,进行求解,得到联合文档 -

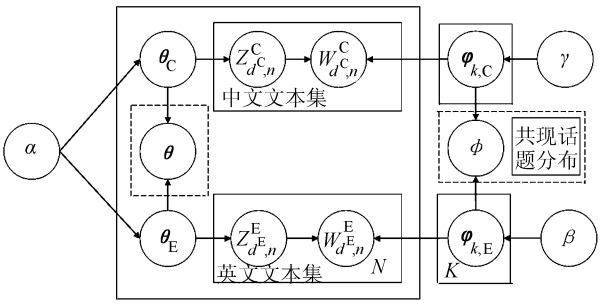

图 2 ICE-LDA 模型

Fig. 2 ICE-LDA model

话题分布 $\{\boldsymbol{\theta}_C, \boldsymbol{\theta}_E\}$，以及各自语言空间上的单词 - 话题分布 $\{\boldsymbol{\varphi}_C, \boldsymbol{\varphi}_E\}$，然后通过将话题向量化的方式计算话题间的距离及相似度来得到共现话题。本文使用上述模型，针对混合中英文话题集，采用两种方法进行话题发现：

第 1 种方法是首先调用翻译工具，将单一语言空间的文档经过翻译过后映射在另一语言空间上，使每一篇文档能够在语义上在双语空间上都有映射，构建可对比的平行语料集；然后利用 ICE-LDA 模型进行建模，得到联合话题分布后通过计算话题 - 文档分布的列向量的 JS 距离设定阈值，进行共同话题发现。百度在线翻译 API 支持较大规模的文本翻译，单次请求字符长度达 100 万，长文本不必再做字符截断[16]，因此本文采用百度在线翻译作为翻译的工具。

第 2 种方法是首先利用 ICE-LDA 模型分别在双语空间上进行聚类，然后在双语空间上两两取话题。例如在中文上取某话题，在英文上取话题，取这两个话题下分布排名前 30 的单词，将英文话题下的单词通过在线翻译词典映射到同一语言空间上，之后分别对这两个话题按单词进行向量化处理，计算向量间的余弦值来衡量话题相似度。这样就可以不需要有先验话题对的情况来发现中英文混合语料集中的共现话题。

由于精确推断 ICE-LDA 模型中的参数十分复杂，本文采用蒙特卡罗方法（Markov chain monte carlo，MCMC）[17] 的一种特例——Gibbs 采样算法[18] 推断本文提出的高维数据模型。Gibbs 采样方法具有采样效率高、适合高维空间采样的特点。Gibbs 算法在固定其它维的情况下，依次在各个维度上采样。Gibbs 采样算法依次在各个维度上进行采样，将文本集中第 i 个词所属的话题编号记为 z_i，其中，$i=(m,n)$ 表示文档集中第 m 篇文档中的第 n 个词。其中，$\neg i$ 表示除去下标为 i 的词。那么按照 Gibbs 采样算法的要求，下标为 i 的词对应的条件分布为 $p(z_i = k \mid z_{\neg i}, w))$，最终得到的采样公式如下：

$$p(z_i = k \mid z_{\neg i}, w) \propto \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_{m,\neg i}^{(t)} + \alpha_k)} \cdot \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V}(n_{k,\neg i}^{(t)} + \beta_t)} \quad (1)$$

一篇新闻本身可能会存在大量的无关紧要的词语，话题建模所需要的是能代表这篇文档的特征词。TF-IDF 是一种特征权重方法[19]。为了简化单词话题距阵 $\boldsymbol{\varphi}_{k,C}$、$\boldsymbol{\varphi}_{k,E}$ 的维度，去除噪声词，对文本的预处理中先使用 NLPIR 分词器，保留名词、动词与实体标注词，计算单词 TF-IDF 值，按权重保留前 75% 的单词实现对文本的特征提取并实现降维效果。文档中词的 TF-IDF 权值计算公式如下所示：

$$w_{i,j} = \frac{f_{i,j}}{\max\{f_{1,j}, f_{2,j}, \cdots, f_{|V|,j}\}} \cdot \lg\frac{N}{df_i + 1} \quad (2)$$

其中，$f_{i,j}$ 表示文档 j 中词 i 的频率，df_i 表示词 i 在多少文档中出现过，N 表示文档数目。实际上，IDF 也有不足之处，如果一个实体特征词条在一个类的文档中频繁出现，则说明该词条能够很好代表该类的文本的特征，应该给它赋予较高的权重，并选作该类文本的特征词以区别与其他类文档。所以根据此不足，本文在进行 TF-IDF 权重值计算时对实体名词与标题中出现的词进行了 TF-IDF 值加倍的处理，尽量保证这些词语不被过滤掉。

英文文本与中文文本混合导致的主要问题是在计算 IDF 值时，中文单词或英文单词会因为语言空间的不同，出现偏离算法本意的结果，比如，“计算机”这个单词进行 IDF 计算，由于只会在中文新闻中出现，而在英文新闻中一般不会出现汉语，导致 IDF 值分母词频统计偏离本身所期望的语义关联而出现偏差导致得出错误的计算值。

针对以上不足，本文使用的 ICE-LDA 模型，针对中英文文本先进行各自语言空间上的处理和聚类，避免上述情况出现。

2 中英双语文本采集系统

本文在采集文本数据时开发了一款基于 Nutch 框架的爬虫系统，整个流程如图 3 所示。Nutch 是基于 Java 实现的爬虫框架，是 Apache 软件基金会下的开源项目[20]。底层基于 Hadoop 平台，采用 MapReduce 计算框架实现，可部署于大规模集群。具有较高的可扩展性，拥有高度模块化的架构设计和易扩充的插件机制，便于用户基于需求定制特定的功能模块，方便地进行二次开发。

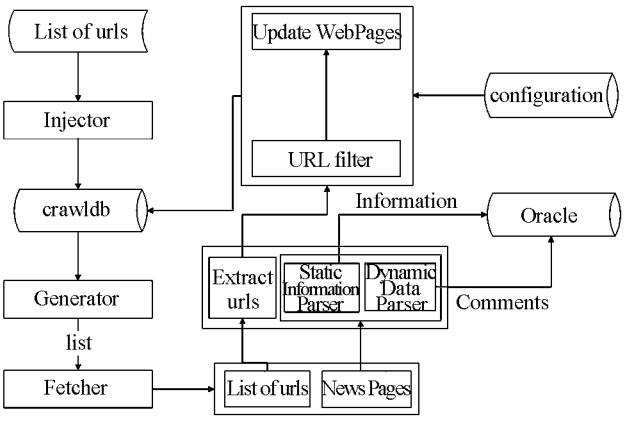

图 3　爬虫流程

Fig. 3　Crawler process

爬虫运行简要流程为:在 Nutch 爬虫配置文件 url. txt 中配置所需爬行的种子链接和一轮采集的更新周期,在 Regex-url. txt 文件中按域名填写所需抓取的 URL 的正则表达式;在爬虫运行的过程中程序首先将种子 URL 加入到抓取状态数据库 crawldb 中,再从 crawldb 中选择出 *topN* 个 URL 生成抓取列表,交由 Fetcher 模块来施行网页抓取,之后对爬取到的网页进行解析,在这个阶段采用了开源的 NewsParserFactory 类,根据不同 URL 的域名来调用相应网站的解析器,利用解析器抽取实验所需正文、作者、标题等信息,同时在解析过程中完成对该 URL 页面下的外链提取等操作,更新 crawldb 数据库中已抓取的 URL 状态,并将新产生的 URL 插入到 crawldb 中,如此重复抓取过程直至 crawldb 中所有的 URL 均被采集。

本文主要针对“腾讯新闻网”“新浪新闻网”“华盛顿邮报”“Chinadaily 日报”的英文版面进行定点采集;并针对特定的主题关键词事件通过调用百度以及谷歌搜索引擎的方式进行新闻数据的采集。总共采集新闻 86 108 篇,其中,英文 9 888 篇,中文 76 220 篇。针对 2014 ~ 2015 年特定事件的采集筛选以及随机抽取新闻构建语料集。最终用于跨语言话题检索的可对比语料文档集 1 205 对,共 2 410 篇语料,构建方式为使用在线百度翻译 API,将原随机抽取的 1 205 篇中英文语料按英汉互译的形式进行翻译,得到可对比语料集;随机抽取非可对比混合语料集新闻 1 949 篇,其中,英文 927 篇,中文 1 022 篇。

3　共现话题定义与发现

共现话题是在不同语料空间上共同出现的话题。文中,共现话题定义为在中英文混合语料集中,既出现在中文语料集这,同时也出现在英文语料集上的话题。这符合平时舆情爆发时的实际情况,即有影响力的国际事件必然会既出现在中文的报道中,也会大量出现在国外的媒体传播网站上。

在 CLU-LDA 模型[9]中,发现共现话题需要依靠先验的话题对才能发现跨英汉语言空间的共现话题,并且使用该模型针对可对比语料集做的实验才能够发现共现话题,而针对无先验话题对和可对比语料集的开放型英汉混合数据集实验则难以发现跨语言语料集中的共现话题。

本文采用提出的 ICE-LDA 模型,针对英汉混合语料集进行实验,同时构建了两种混合语料集,一种为可对比平行语料集,即英汉文本在双语空间内一一对应;另一种为非可对比英汉混合语料集,即中英文文本不是一一对应的。对于两种不同的语料集,本文使用 ICE-LDA 进行建模后,得到中英文各自语言空间上的单词 - 话题分布和话题 - 文本方法分布。

针对可对比英汉混合语料集的建模结果,本文采用文本 - 话题分布矩阵进行计算,将每一个话题按照文本在该话题的分布进行向量化处理,得到每一个不同语言空间上编号为 i 的话题向量表示如式(3):

$$\boldsymbol{K}_i^L = (\theta_{i,1},\theta_{i,2},\theta_{i,3},\cdots,\theta_{i,m}) \tag{3}$$

其中,i 表示话题编号,L 表示语言空间,$\theta_{i,m}$ 表示在该语言空间上的第 i 个话题的第 m 篇文本的分布。

话题相似性计算一般采用 KL 距离(Kullback-Leibler divergence)[2],它度量了两个话题在词集上分布的差异性。对于编号为 M 和 N 两个话题分布和,KL 距离定义如式(4):

$$KL(\varphi_M,\varphi_N) = \sum_{i=1}^{V} \varphi_{M,i} \lg \frac{\varphi_{M,i}}{\varphi_{N,i}} \tag{4}$$

因为 KL 距离是不对称的,但两个话题的距离应该是对称的。本文采用 Jensen-Shannon 距离度量两个话题分布的差异性。JS 距离由 KL 距离定义,对于话题分布和,JS 距离如式(5):

$$JS(\varphi_M,\varphi_N) = \frac{1}{2}\left[KL\left(\varphi_M,\frac{\varphi_N+\varphi_M}{2}\right)+KL\left(\varphi_N,\frac{\varphi_N+\varphi_M}{2}\right)\right] \tag{5}$$

因为在可对比语料集中中英文的文本是一一对应的,那么可以将每一篇文本对于双语空间共现话题的贡献在各自语言空间上看成是相近的。这样就可以将不同语言空间的所有话题向量化以后,通过

计算话题向量间的距离,设定阈值来计算两个话题是否是共现话题。

针对非可对比的英汉混合语料集的聚类结果,本文采用单词 - 话题分布矩阵进行计算,将每一个话题按照在其上分布排名前 30 的单词进行向量化处理如下:

$$\boldsymbol{K}_i^L = (w_1, w_2, w_3, \cdots, w_{30}) \tag{6}$$

其中,i 表示话题编号,L 表示语言空间,w 表示单词。

由于中文与英文单词分属两种不同的语义空间,需要对其进行映射。有道翻译 API 对字符串长度限制为 200,但没有次数频率上的限制且翻译准确度高,满足本文对翻译词语的要求。所以,本文通过调用有道翻译 API[22] 在线翻译词典对英文单词进行翻译,从而将中文与英文单词就映射在了同一语义空间,话题向量之间就能够通过计算向量相似度来判断是否为共现话题。

对于非平行语料集上的话题,本文采用向量空间模型[23](vector space model, VSM)表示,并采用向量夹角余弦相似度作为话题相似度的计算方法。对于两个话题向量夹角余弦相似度计算公式如下:

$$\cos(\boldsymbol{d}_m, \boldsymbol{d}_n) = \frac{\sum_{i=1}^{|V|} d_{mi} \cdot d_{ni}}{\sqrt{\sum_{i=1}^{|V|} d_{mi}^2} \cdot \sqrt{\sum_{i=1}^{|V|} d_{ni}^2}} \tag{7}$$

余弦值约接近 1 说明向量距离越近,说明话题相似度越高。

4 实验结果及分析

4.1 模型性能度量

困惑度是自然语言处理中常用的一种指标,广泛用于话题模型的性能度量[14]。困惑度衡量了话题模型对于未观测数据的预测能力,困惑度越小,模型预测能力越强,模型的推广性越高。困惑度计算如下:

$$perplexity(D_{\text{test}}) = \exp\{-\frac{\sum_{d=1}^{M} \lg p(w_d)}{\sum_{d=1}^{M} N_d}\} \tag{8}$$

式中,D_{test} 代表测试集,N_d 表示文档 d 中的可观测单词序列,w_d 表示文档 d 的单词数目。实验从数据集中随机选择 10% 的文档作为测试集,其余文档作为训练集进行 LDA 模型推断后计算模型困惑度。

本文采用 TF-IDF 特征词提取的方式进行降维去噪,随机从采集到的新闻数据集中选取了500、1 000、1 500、2 000、2 500 篇数量新闻构成 5 个实验数据集,设定聚类话题 K 数目为 100 个,先验参数 α 设置为 0.5,先验参数 β 设置 0.01,采样迭代次数设定为 1 000,计算去噪先后的困惑度如图 4 所示。

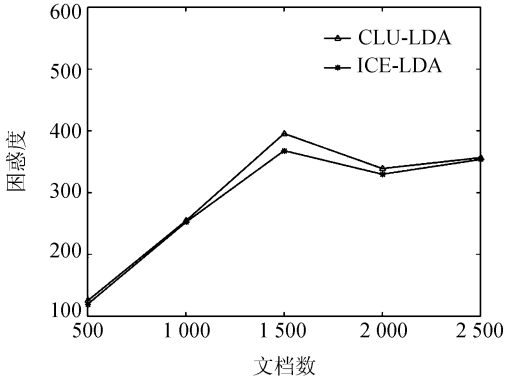

图 4 实验文本集困惑度

Fig. 4 Perplexity of datasets

并使用内存 6 G,Core(TM) i3 CPU M390 主频 2.67 GHz 配置电脑测试降维后耗时,平均消耗时间如图 5 所示。

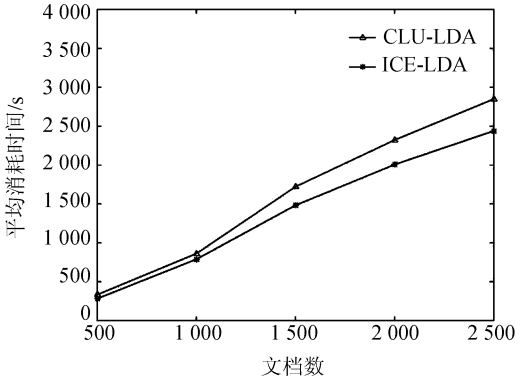

图 5 平均时间消耗

Fig. 5 Average time-consuming

可以看出经过降维去噪后的模型困惑度有一定的优化,并且建模时间缩短,文本集数量越多,节省时间越多。

4.2 中英文平行语料集话题相似度度量与共现话题对发现

对于中英文混合平行语料集,采用本文第 3 节介绍的方法,将 ICE-LDA 模型建模得到的中英文文档 - 主题分布进行向量化处理并两两计算 JS 距离,距离值越小说明话题越相近。得到每一个话题之间的最近距离分布,取距离阈值为 1 以内的话题为共现话题,发现话题对之间距离如表 1 所示,共现话题对如表 2 所示。

表 1　话题间距离

Tab. 1　Distance of topics

话题编号	话题编号	JS 距离
17	85	0.522 746 547 199 532 1
92	40	0.776 378 815 634 420 9
70	99	0.753 317 802 450 852 9
98	31	0.978 166 875 283 895 5

表 2　共现话题对与关键词

Tab. 2　Pairs of topic and keywords

话题编号	关键词(top 15)
17	飞机 复兴 台湾 台北 图片 救援 照片 法新社 人员 起飞 死亡 坠毁 基隆 ATR 乘客 航班
85	TransAsia river Taiwan Taipeiplane people Photo crash crashed clip ATR bridge YEH Getty capital
92	警察 抗议 芝加哥 黑人 拍摄 视频 曼纽尔 威胁 城市 暴力 麦当劳 警官 弗格森 密苏里 局长
40	police Chicago city shooting shot rights protests Emanuel protesters protest year footage Dyke McDonald station
70	巴黎 袭击 法国 攻击 死亡警察 自杀 分子 炸弹 爆炸 警方 比利时 餐厅 人质 音乐厅
99	Paris attacks France attack people killed terror Bataclan restaurant hall victims stadium Stade concert
98	项目 管道 土库曼斯坦 建设 铁路 天然气 阿塞拜疆 工程 电力 运行 出口 阶段 恢复 资源 提高
31	project gas Turkmenistan pipeline construction land transportation train network cost water export projects railway resources

从表 1 与 2 中可以看出,对于中英文混合平行语料集,采用话题按照文档 - 话题分布进行向量化处理后能够通过计算话题间的距离来得到双语空间上的共现话题。

4.3　中英文非平行语料集话题相似度度量与共现话题对发现

对于非平行语料集采用 ICE-LDA 模型进行建模,得到中英文单词 - 话题分布。采用第 3 节提到的方法,将英文话题分布值较高的前 30 个单词通过在线翻译词典进行翻译,并使用余弦相似度进行中英文语料集之间的话题相似度计算,得到话题间的相似度如表 3 所示,相似度越大说明向量间越相似。经过实验得到符合条件话题对 9 对,得到部分话题对如表 4 所示。

表 3　话题间相似度

Tab. 3　Similarity of topics

话题编号	话题编号	相似度
61	23	0.355 600 355 600 533
35	44	0.254 000 254 000 381
98	30	0.228 600 228 600 342
58	39	0.228 600 228 600 342

表 4　共现话题对应关键词

Tab. 4　Pairs of topic and keywords

话题编号	话题关键词(top 15)
61	埃博拉 病毒 疫苗 感染 爆发 疫情 治疗 试验 死亡 几内亚 传播 西非 感染者 患者 利比里亚
23	Ebola virusAfrica outbreak health cases disease Guinea spread Leone patients infected Health deaths Organization
35	MH370 搜寻 飞机 马来西亚 客机 澳大利亚 航班 区域 搜索 马航 残骸 乘客 发现 亚航失事
44	plane search aircraft flight Malaysia crash system pilot air Flight March Malaysian Boeing aviation MH370
98	两岸 会面 领导人 关系 台湾 大陆 共识 新加坡 交流 和平 历史 马英九 同胞 张志军 基础
30	Taiwan meeting China Xi Ma cross Strait Ying Jinping meet relations Singapore jeou War sides
58	巴黎 袭击 法国 发生 分子 爆炸 死亡 恐怖主义 伊斯兰 剧院 体育场 组织 奥朗德 枪击 遇难
39	Paris France people Bataclan killed concert suicide stadium attack de hall police Stade night terror

从以上表 3 和 4 中可以看出,对于中英文混合非平行语料集,采用在线翻译词典将话题的主题词映射到中文语义空间上,进行向量化处理后能够通过计算话题间的相似度来得到双语空间上的共现话题。

5　结　论

本文主要的贡献在于,针对中英文混合平行语料集与中英文混合非平行语料集,提出 ICE-LDA 模型进行建模,并采用特征提取的方式对建模过程中的矩阵进行降维,提高计算效率,采用将话题按不同分布进行向量化的方式,计算话题向量间的相似度,从而发现中英文跨文本集上的共现话题。

本文研究也尚有不足之处,主要是由于英文单词的一词多义,目前的逐词翻译可能会导致该词的词义与在文本语句中的词义有所不同,造成计算文档相似度时产生一定误差的情况,但是一般实体名词的翻译效果由于唯一性比较准确。

参考文献:

[1]CNNIC 中国互联网络发展状况统计报告[EB/OL]. [2016-01-22]. http://www.cnnic.com.cn/hlwfzyj/hlwxzbg/201601/P020160122469130059846.pdf.

[2]Xu Xiaori. Study on the way to solve the paroxysmal public feelings on internet[J]. Journal of North China Electric Power University(Social Sciences),2007(1):89-93. [徐晓. 网络舆情事件的应急处理研究[J]. 华北电力大学学报(社会科学版),2007(1):89-93.]

[3]Wan Jiexi. Research on multilingual text cluster[D]. Nanjing:Nanjing University. 2013. [万接喜. 多语言文本聚类研究[D]. 南京:南京大学,2013.]

[4]Zhang Chenzhi, Wang Linghui. Survey on multilingual documents clustering[J]. New Technology of Library and Information Service,2009,25(6):31-36. [章成志,王惠临. 多语言文本聚类研究综述[J]. 现代图书情报技术,2009,25(6):31-36.]

[5]Leftin L J. Newsblaster russian-english clustering performance analysis [R]//Computer Science Technical Report Series. New York:Columbia University,2003.

[6]Littman M L, Dumais S T, Landauer T K. Automatic cross-language information retrieval using latent semantic indexing [M]//Cross-LanguageInformation Retrieval. Heidelberg: Springer-Velag,1998:51-62.

[7]de Smet W, Moens M F. Cross-language linking of news stories on the web using interlingual topic modelling[C]//Proceedings of the 2nd ACM Workshop on Social Web Search and Mining (SWSM'09). Hong Kong:Association for Computing Machinery,2009:57-64.

[8]Ni X, Sun J T, Hu J, et al. Mining multilingual topics from Wikipedia[C]//Proceedings of the 18th International Conference on World Wide Web(WWW09). Madri:Association for Computing Machinery,2009:1155-1156.

[9]陆前. 英、汉跨语言话题检测与跟踪技术研究[D]. 北京:中央民族大学,2013.

[10]Gao Shengxiang, Yu Zhengtao, Long Wenxu, et al. Chinese-vietnamese blingual news event storyline analysis based on words co-occurrence distribution[J]. Journal of Chinese Information Processing,2015,29(6):90-96. [高盛祥,余正涛,龙文旭,等. 基于全局/局部共现词对分布的汉越双语新闻事件线索分析[J]. 中文信息学报,2015,29(6):90-97.]

[11]Guo Jiacheng. Supervised topic model[D]. Shanghai: Shanghai Jiao Tong University,2010. [郭佳骋. 监督学习的话题模型[D]. 上海:上海交通大学,2010.]

[12]Scott D, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science,1990,41:391-407.

[13]Hofmann T. Probabilistic latent semantic indexing[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Boston, MA: Association for Computing Machinery,2015:56-73.

[14]Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 2(3): 993-1022.

[15]Griffiths T L, Steyvers M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences of the United States of America,2004,101 (Suppl 1):5228-5235.

[16]百度翻译开放平台[DB/OL]. [2016-08-10]. http://api.fanyi.baidu.com/api/trans/product/apidoc.

[17]Liu Leping, Gao Lei, Yang Na. Development of MCMC methods and revial of modern baysian celebrating 250 Years of Bayes's Theorem[J]. Status and Information Forum, 2014,29(2):3-11. [刘乐平,高磊,杨娜. MCMC 方法的发展与现代贝叶斯的复兴——纪念贝叶斯定理发现250周年[J]. 统计与信息论坛,2014,29(2):3-11.]

[18]Witten I H, Paynter G W, Frank E, et al. KEA: Practical automatic key phrase extraction[C]//Proceedings of the 4th ACM Conference on Digital Library. New York: Association for Computing Machinery,1999:254-255.

[19]Voorhees E M. Variations in relevance judgments and the measurement of retrieval effectiveness[J]. Information Processing & Management,2015,36(5):697-716.

[20]Mcgibbney L J. Nutch Wiki nutch tutorial [EB/OL]. (2016-11-21)[2016-08-10] http://wiki.apache.org/nutch/NutchTutorial.

[21]Zhou Gang, Zou Hongcheng, Xiong Xiaobing. MB-Single-Pass: Microblog topic detection based on combined similarity [J]. Computer Science,2012,39(10):198-202.

[22]有道翻译 API[DB/OL]. [2016-08-10]. http://fanyi.youdao.com/openapi.

[23]Salton G, Wong A, Yang C S. A vector space model for automatic indexing[M]//Readings in Information Retrieval. San Francisco: Morgan Kaufmann Publishers Inc,1997:273-280.

(编辑 赵 婧)

引用格式: Chen Xingshu, Luo Liang, Wang Haizhou, et al. Analysis and research on cross language topic discovery in chinese and english[J]. Advanced Engineering Sciences,2017,49(2):100-106. [陈兴蜀,罗梁,王海舟,等. 基于 ICE-LDA 模型的中英文跨语言话题发现研究[J]. 工程科学与技术,2017,49(2):100-106.]