基于种子词汇的话题标签抽取研究

寇宛秋，李 芳

（上海交通大学 计算机科学与工程系,上海闵行东川路 800 号 200240）

摘要 传统话题模型用词项概率分布表示话题，在可解释性上存在很大的不足。本文在 Latent Dirichlet Allocation（LDA）的结果上提出了一种基于种子词汇的话题标签抽取方法。首先根据提出的权重计算公式抽取每个话题的种子词，然后，采用 bootstrapping 思想，迭代产生包含种子词汇的关键短语集合，最后根据短语的完整性和泛化度选择话题标签。本文对两会报告话题和新闻事件话题进行实验，通过结果展示和人工评测，该方法抽取的话题标签能够较准确地表达话题的语义信息。

关键词 话题标签;种子词抽取;bootstrapping 算法

Topic Label Extraction based on Seed Words

Abstract Traditional topic models use word probability distribution to represent topics. These words are difficult to be understandable and express a consistent meaning. This paper proposed a topic label extraction method based on seed words. The method first extracts topic seed words according to weight formulas, then uses bootstrapping algorithm to generate a key phrase set that contains seed words. Finally, the method selects topic label from the key phrase set according to the integrity and generalization of a phrase. The experiments were made on two corpora. One is topic oriented reports, the other is event based news reports. According to the experimental results, the method work well in extracting a meaningful phrase to represent a topic.

Keywords topic labelling; seed words extraction; bootstrapping method;

1. 引言

当今社会已经进入信息大爆炸的时代，信息量以几何级别速度不断增加。据调查显示，《纽约时报》一周的信息量相当于 17 世纪学者毕生所能接触到的信息量的总和。伴随着信息爆炸的是信息匮乏，海量的信息鱼龙混杂，收集信息所花费的成本已经超过了信息本身的价值。如何获取有价值的信息，已经成为信息爆炸时代极为重要的议题。

话题模型被普遍用来解决这个问题。话题通常被表示成词项的概率分布，话题模型通过对文档集进行降维，将词项空间中的文档变换到话题空间，模拟文档的生成过程。在话题模型中，一个话题用一组关键词来表示，有些话题有一个明确的语义信息，例如，“房价”、“住房”、“土地”、“市场、“上涨”，有些话题没有，例如“网友”、“创意”、“得意”、“广电总局”、“影像”。因此，在实际应用中，需要一个标签来表示话题的语义信息。相对于单个词项，短语能够表示较完整的语义信息，如何从话题模型中得到更具解释性的短语描述，作为话题的标签是本文研究的目的。

本文的组织结构如下：第一部分主要介绍相关工作，第二部分是话题标签抽取方法的描述，第三部分是实验结果和分析，第四部分是结论和展望。

2. 相关工作

话题模型应用最广的是 LDA 模型， 是 DAVID BLEI 在 2003 年提出的[1]。之后很多研究者基于文档特点对 LDA 做了很多拓展,例如 Blei 在 2004 年提出的 Hierarchical LDA [2]， 将话题间的结构描述为树;Hidden Topic Markov Model(HTMM)[2]用句子的分布来表示话题； Author Topic Model(ATM) [3]在话题模型中引入作者信息，用以处理科技文献。

话题标签抽取研究可以分为四种方法，第一种是调整话题模型结果的权重，例如 Weighted latent dirichlet allocation（WLDA）模型[4],在 LDA 模型中，每个单词都被等同看待，而 WLDA 为每个单词赋予一个不同的权重。很多特征权重被用在该模型中，例如 Pointwise Mutual Information (PMI)，CHI 测试，信息增益等。本文方法采用了 WLDA 的思路，利用权重公式对 LDA 建模结果进行权重调整处理。

基金项目：国家自然科学基金项目（60873134）资助

第二种方法是采用短语为单元描述话题，传统话题模型采用单个词语作为话题关键词，而一些研究者用短语取代单个词语。例如 Multiword-Enhanced Author Topic Model[5],该模型根据词性标注信息抽取符合特定短语模式的短语，然后基于这些短语和单词构建话题模型。本文方法采用这一思想，用短语取代单词表示话题。

第三种方法是在话题结果中引入语义信息,例如 POSLDA 模型[6],该模型是 LDA 模型和 HMMLDA 模型[7]的扩展，该模型将文档中的词项分为三个类别，形容词、动词和名词，可以表示话题涉及的事物、动作和描述信息。

第四种方法是对 LDA 生成的话题结果进行组合处理，例如 Turbo Topic[8],该方法基于 LDA 的结果抽取可能的短语。算法步骤如下：

（1） 对文档进行 LDA 建模，得到文档-话题分布，词项-话题分布和每个单词所属话题的词对：

$$(w_1, z_2), (w_2, z_2) \ldots$$

（2） 对每一个单词，判断该单词周围的单词是否和该单词属于同样的话题，如果属于，则这两个单词可能组成一个短语，再根据似然估计，判断它们是否可以组成短语，如果可以，则加入到短语集合中

（3） 重复步骤（b），直到找不出新的短语。

本文综合了以上几种方法，引入了特征权重、词性分析，短语表示等因素，产生话题的标签，有效提高了话题模型结果的可解释性。

3. 方法介绍

话题标签信息是话题内容的概括与总结，能够综合地反映话题内容，增强话题的可解释性。表 1 展示了 LDA 建模生成的话题信息和采用本文方法抽取的话题标签信息。

表 1 话题信息与对应的话题标签

话题	话题中概率最大的 10 个词语	话题标签
话题 1	减排 指标 节能 完成 排放 环境 约束性 考核 总量 目标	完成节能减排指标
话题 2	安全 食品 管理 监管 国家 药品 监督 组建 职责 生产	食品安全监管

下表为本文使用到的主要符号和定义

表 2 话题标签抽取研究涉及的符号

符号	符号描述
z	话题
w	词项
d	文档
θ_{ij}	文档d_j中话题z_i的权重
ϕ_{ij}	话题z_j中词项w_i的权重
α	LDA 先验参数
β	LDA 先验参数
P	短语集合
p	表示短语或单词
K	话题个数
D	文档个数

话题标签抽取方法主要包括四个步骤：话题建模，种子词抽取，关键短语抽取和话题标签选择。话题建模是利用 LDA 模型对输入的文本集合进行建模，种子词抽取是对 LDA 话题结果进行重排序，选择权重最大的前三个词作为种子词，关键短语生成是根据种子词和其它词汇出现次数等信息生成短语，话题标签选

择是从这些短语中选择最终话题标签。

3.1 种子词抽取

根据文献[9]提出的 LDA 结果重排序方法，根据下面公式对 LDA 结果，调整话题词项的权重，进行重排序。

(a) Term frequency - Inverse document frequency (TF-IDF)
TF-IDF 被广泛用于评估词项在文档中的重要性。词项在文档中出现的次数越多，包含该词项的文档数目越少，就越重要。w_i在话题z_j中的重要性权重计算如下：

$$\mathrm{TF-IDF}_{i,j} = \phi_{i,j} * (\phi_{i,j} / \sum_{k=1}^{K} \phi_{i,k}) \quad (1)$$

(b) 话题覆盖度
用于计算一个话题在文档集合上的覆盖程度，覆盖度高的话题中词项的权重更大。话题覆盖度用一个话题在所有文档中的概率之和除以总文档数来表示：

$$\mathrm{coverage(j)} = \frac{\sum_{d=1}^{D} \theta_{j,d}}{D} \quad (2)$$

(c) PMI
PMI 统计概率分布中两个变量的相关性，公式如下

$$\mathrm{pmi}(w_i, w_j) = \log p(w_i, w_j) / p(w_i) p(w_j) \quad (3)$$

词汇w_i与同一话题（top-10）中其它 9 个词汇越相关，则该词汇的权重越高，某一词汇的关联度计算用 PMI 的平均值。

因此，结合 TF-IDF，覆盖度以及和 PMI，权重计算公式如下：

$$\mathrm{weight}_{i,j} = \phi_{i,j} * \left(\frac{\phi_{i,j} * \mathrm{coverage(j)}}{\sum_{k=1}^{K} \phi_{i,k} * \mathrm{coverage(k)}} \right) * \left(\frac{1}{9} * \sum_{j \neq i} \mathrm{pmi}(w_i, w_j) \right) \quad (4)$$

根据公式 5，对每个话题前十个单词进行权重重排序，选出前三个单词作为关键短语抽取的种子词。

3.2 关键短语集合生成

初始关键短语集合等于种子词集合，运用 bootstrapping 算法迭代生成短语，当短语的权重大于阈值，则加入到关键短语集合中。用W_{seed}表示种子词集合，用 P 表示关键短语集合（初始阶段等于W_{seed}），用W_{LDA}表示 LDA 话题前十个词。短语（p_1, p_2）同时满足下述条件，则为关键短语：

（1） p_1, p_2 是属于$P \cup W_{LDA}$ 中的任意短语或单词
（2） p_1, p_2中至少有一个属于 P
（3） (p_1, p_2)的权重大于阈值

算法 1 描述了关键短语生成的过程：

算法 1 话题关键短语生成算法

Topic keyphrase generation algorithm:
Input : P - 关键词集合，初始为种子词汇,词汇权重为用公式 4 重排序后的权重
W_{seed} - 种子词汇
W_{LDA} - LDA top-10 words
Output : P - 话题的关键短语集合
Set flag as true
Set P = W_{seed}
计算 P 中各短语的权重 phraseweight(p_i)
While flag
For p_i in P
For p_j in W_{LDA}

生成短语(p_i, p_j)，并统计其共现次数 $n(p_i, p_j)$

计算短语权重

$$\text{phraseweight}(p_i, p_j) = \text{weight}(p_i) * \frac{n(p_i, p_j)}{n(p_i)} + \text{weight}(p_j) * \frac{n(p_i, p_j)}{n(p_j)}$$

End for

End for

选择权重最高的短语(p_i, p_j)

If $\text{phraseweight}(p_i, p_j) \geq$ 阈值

将短语(p_i, p_j)作为$p_{|P|+1}$放入关键短语集合 P 中,

$$\text{weight}(p_{|P|+1}) = \text{phraseweight}(p_i, p_j)$$

Else

把 flag 设置为 false

End if

End while

3.3 话题标签选择

在抽取出关键短语后，需要从关键短语集合中最终选出解释性强的短语作为话题标签。本文提出两种标准选择话题标签：短语的完整性和泛化度。

3.3.1 短语完整性标准

根据实验结果，有些权重最高的关键短语缺乏关键信息，例如，关键短语“卡恩涉嫌”、“同比增长”、“中方支持”。这些短语在语义上并不完整，“卡恩涉嫌”缺少宾语，“同比增长”缺少主语，“中方支持”缺少宾语。大部分不完整的短语均是动词性短语。因此，短语完整性规则如下：如果关键短语集合中权重最高的短语是动词词组，而且缺少主语或宾语，则按照完整性规则，在关键短语集合中重新选择。

判断以及选择方法如下：

假设关键短语集合 P 中权重最高的短语为p_{max}，

（1） 如果该短语第一个词为动词，或者第一个动词前没有名词， 则判定短语p_{max}缺乏主语。

（2） 如果该短语最后一个词为动词或者最后一个动词后面没有名词，则判定p_{max}缺乏宾语。

对于判定缺乏主语或宾语的短语p_{max}，在关键短语集合 P 中，按权重从高到低的顺序搜索满足如下条件的短语 p，作为最后的标签:

（1） p 包含短语p_{max}

（2） p 中含有主语（动词前的名词）或宾语（动词后的名词）

3.3.2 概念泛化标准

实验发现了另一种现象，即权重最高的关键短语只是描述话题特定的方面，例如“治理北京大气污染”，而其它的关键短语为“大气污染” “大气污染防治”，更好的描述短语是“大气污染”。这类短语一般是名词性短语，为了解决这种问题，本文引入概念泛化规则：关键短语集合中权重最高的短语，如果是名词短语，则根据该集合中其它词汇进行泛化，选择关键短语最大的公共子串作为该话题的标签。

具体步骤如下：

计算关键短语集合 P 中短语 p 的泛化度：

a) 对于同时满足条件 i 和条件 ii 的短语 p，按照公式 6 计算泛化度

i. 短语 p 属于 P 中权重最高的三个短语或者权重前三的短语包含 p。

ii. P 中至少存在两个包含 p 的短语

$$\text{phrasecoverage（p）} = \text{phraseweight(p)} + \sum \text{phraseweight（}p_{contain}\text{）} \quad (5)$$

其中$p_{contain}$是在 P 中包含 p 的所有短语。

b) 对于不满足 a)中条件的短语 p，按照公式 7 计算泛化度。

$$\text{phrasecoverage（p）} = \text{phraseweight(p)} \quad (6)$$

话题标签根据如下规则得出：

a) 当条件 i 或条件 ii 满足时，选择P 中泛化度最高的词汇p_{genmax}为话题标签

i. p_{genmax}的泛化度大于关键短语集合 P 中最大权重短语p_{max}的泛化度。

ii. p_{genmax}的泛化度小于关键短语集合 P 中最大权重短语p_{max}的泛化度且p_{max}包含p_{genmax}，同时

$$\text{phrasecoverage}(p_{max}) < 60\% * phrasecoverage(p_{genmax}) \quad (7)$$

b) 选择P 中权重最高的词汇p_{max}为话题标签，当且仅当 a)中条件不能满足

4. 实验结果分析

4.1 实验语料

实验共选取了两个语料集进行测试：2013 年两会新闻数据集和在 09 至 13 年发生的新闻事件集合。先预处理，分词并抽取名词动词形容词，去掉单个字以及高频低频词，然后用 LDA 对其进行建模。实验设置参数$\alpha = \frac{50}{K}$ $\beta = 0.01$, 其中 K 为话题数目。我们采取了一套自适应的话题数目计算方法，根据新闻文本数目以及信息量随时间的变化趋势确定话题数目[9]。表 3 是不同事件对应的新闻数目、词汇数目、话题数目等信息，按照话题数目从小到大的顺序展示。

表 3 实验语料说明

事件名称	新闻数目	词汇数目	话题数目	事件名称	新闻报道数目	词汇数目	话题数目
巴勒斯坦申请加入联合国	150	6070	3	神州 9 号	213	14728	3
多省区暴雨洪水	193	7164	3	2013 年春运	102	9337	3
李双江之子无照驾车	67	5170	3	台湾领导人选举	186	12093	4
胡锦涛出席 APEC	88	3567	3	朝鲜发射卫星	168	9905	4
京沪高铁	149	10909	3	美国丹佛枪击案	144	8101	5
2010 年美国中期选举	195	9861	3	2009 日本众议院	373	12273	6
菲律宾台风	23	1561	3	2011 春节	400	20874	6
里约联合国可持续发展论坛	151	9708	3	台湾塑化剂事件	240	9559	6
金正日逝世	249	13900	3	朝鲜半岛局势恶化	497	9578	8
雅安地震	126	11112	3	2013 年两会	3541	18287	60
马云卸任 CEO	50	7288	3				

4.2 实验结果展示

话题标签抽取方法对新闻语料进行处理，主要包括三个步骤：种子词抽取；关键短语集合生成；话题标签选择。表 4 和表 5 分别展示了事件语料话题标签抽取实验和两会语料话题标签抽取实验各步骤的结果。

表 4 事件语料话题标签抽取结果

事件	话题编号	LDA 前十个单词	种子词	关键短语集合	话题标签
台湾领导人选举	话题 0	台湾 宋楚瑜 马英九 地区 日本 民众 领导人 投票 松涛 主持人	主持人 地区 吴敦义	台湾地区领导人选举 台湾地区领导人 台湾地区	台湾地区领导人选举
	话题 1	总统 政党 国民党 选举 台湾	政党 区域 获得	区域当选席次	区域当选席次

		委员会 主席 代表 选举 选票		区域席次 区域选票	
	话题 2	台湾 两岸 马英九 关系 经济 大陆 九二共识 政策 英文 和平	两岸 关系 台湾	两岸和平 两岸关系 台湾选举	两岸关系
	话题 3	民进 英文 国民党 表示 立委 主席 报道 选举 当选 候选人	民进 英文 竞选	竞选团队 英文竞选 面对民进	竞选团队
2011 春节	话题 0	中国 群众 胡锦涛 生活 人民 工作 社会 村民 总理 发展	胡锦涛 群众 中国	中国人民 胡锦涛来到 胡锦涛来到村民	中国人民
	话题 1	庙会 游客 文化 摊位 中国 记者 活动 传统 地坛 北京	庙会 地坛 龙潭	地坛公园 龙潭公园 地坛庙会	地坛公园
	话题 2	烟花 爆竹 燃放 安全 北京 火灾 发生 部门 记者 空气	烟花 燃放 爆竹	燃放烟花 烟花燃放 安全燃放管理工作	燃放烟花爆竹
	话题 3	过年 记者 拜年 回家 网友 老人 工作 父母 孩子 没有	拜年 回家 过年	回家过年 拜年方式 过年回家	回家过年
	话题 4	旅游 游客 同比 增长 接待 景区 地区 黄金 假日 公园	旅游 同比 景区	同比增长 接待游客同比增长 接待游客同比增长	接待游客同比增长
	话题 5	记者 价格 市场 企业 食品 市民 增加 商品 餐饮 销售	企业 餐饮 消费	餐饮企业 消费市场 企业销售	餐饮企业

表 5 两会语料话题标签抽取结果

话题编号	LDA 前十个单词	种子词	关键短语集合	话题标签
话题 5	处理 只能 火化 存在 垃圾 无害化 新闻 掩埋 互动 及时	只能 火化 处理	垃圾无害化处理 无害化处理掩埋 无害化处理	垃圾无害化处理
话题 7	医疗 卫生 医生 基层 服务 医院 主任 医药 基本 看病	卫生 医疗 医药	医药卫生体制 基层医疗卫生服务 医疗卫生	医疗卫生
话题 12	建设 推进 加强 完善 机制 提高 加快 促进 社会 体系	完善 加强 加快	完善社会保障制度 服务体系建设加快 完善社会保障	完善社会保障制度
话题 18	生态 环境 保护 建设 文明 美丽 绿色 国家 全国 工程	生态 文明 美丽	建设全国生态文明 生态文明建设 生态文明	生态文明
话题 24	城市 城镇化 人口 城乡 城镇 建设 推进 农村 农业 实现	城市 城镇化 城镇	推进城镇化 农村城镇化 城镇建设	推进城镇化

话题30	经济 产业 转型 加快 结构 投资 促进 优势 创新 环境	产业 转型 经济	经济转型 经济转型创新 加快转型升级	经济转型
话题33	收入 保障 分配 居民 社会 制度 提高 公平 基本 民生	收入 分配 居民	收入分配 收入分配制度 提高居民收入	收入分配制度
话题34	安全 食品 监管 管理 药品 部门 生产 标准 奶粉 责任	安全 监管 奶粉	食品安全 食品药品安全 质量安全监督管理	食品安全
话题35	污染 环境 治理 大气 空气 质量 北京 城市 排放 监测	污染 大气 空气	大气污染 大气污染防治 治理北京大气污染	我国大气污染
话题58	文化 传统 保护 历史 民族 艺术 故宫 活动 交流 博物馆	文化 观众 遗产	传统文化 文化遗产 民族文化遗产	传统文化

实验结果显示，种子词抽取方法能够有效去除话题背景词，抽取相关的重要词汇。例如台湾领导人选举话题2，话题关键词中有很多背景词，如“台湾”“马英九”等，根据 3.1 提出的权重公式计算后，降低了背景词的权重，提高了“两岸关系”等词汇的权重，更能反映话题的语义信息。

关键短语生成步骤可以产生有效的话题关键短语，例如台湾领导人选举事件中能够生成和事件有关的“台湾领导人选举”、“台湾领导人”等短语；2011 年春节事件中能够生成“回家过年”、“燃放烟花”等短语；两会事件话题 58 中能够生成和文化领域相关的“传统文化”、“文化遗产”等。

根据完整性和泛化规则选择的标签可以给出话题特定的语言信息，例如 2011 春节话题 4，“接待游客同比增长”而不是缺乏主语的“同比增长”。另一方面，台湾领导人话题 2“两岸关系”作为标签，“两岸关系”的泛化程度比“两岸和平”高；例如两会话题 30，话题标签“经济转型”更能概括话题关键短语的信息。

4.3 话题标签实验评测

4.3.1 精度评测

人工评测话题的标签是否符合话题的语义。评测需要的数据是话题标签以及该话题所占权重最大的文档标题。评测者根据新闻题目人工总结出关键短语，并和自动抽取的话题标签进行比较，语义相关的判定话题标签正确，评分为1，部分相关的评分0.5，不相关的为0。例如人工总结的短语是“两岸和平”，计算机抽取的是“两岸关系”，则该标签的精度为0.5；例如人工总结的短语是“救援情况”，计算机抽取的标签是“登陆美国”，则该标签的精度是0。

本文实验中有两位评测者对全部语料进行评测。计算出的精度如表6所示。结果显示，话题标签抽取方法在两会语料的精度可以达到39.5%，在事件语料上的精度可以达到27.9%。

表6 实验评测结果

语料	标签抽取精度/评测者1	标签抽取精度/评测者2	平均精度
事件语料	28.8%	26.9%	27.9%
两会语料	37.5%	41.6%	39.5%

根据实验评测结果，可以得到如下结论：

（a） 话题标签抽取方法能较好的总结话题内容，所抽取的标签短语由话题关键词组成，能够表示特定的语义信息。

（b） 两会语料的精度要高于事件语料，主要因为两会语料讨论的是话题，有一些固定的主题，例如“国防军事”“教育”“住房问题”等等，两会语料中抽取的话题标签往往由名词性短语组成。而事件的话题信息比较特定，包括与事件有关的信息，事件语料中抽取的话题标签有很多包含动词短语，反

映事件特定的信息。

线索标签抽取方法存在不足，最主要是精度较低，这是因为本文提出的关键短语作为话题标签，短语更能反映话题的语义信息，但人工评测时，短语比词汇更容易错误。另一方面，不同人对同一类文档总结的标签也不相同，很难得出一个正确的答案。表 7 展示了部分错误的话题标签。

表 7 错误结果分析

事件	话题	本文方法话题标签	人工话题标签（评测者 1）	人工话题标签（评测者 2）
胡锦涛出席 APEC	话题 1	经合组织领导人增长	胡锦涛 APEC 演讲	胡锦涛 APEC 发表演讲
2010 年美国中期选举	话题 1	美国经济	中美关系	美国大选影响
金正日逝世	话题 1	中国新闻	金正日关注农民	金正恩评价
马云卸任 CEO	话题 1	阿里巴巴集团	阿里巴巴高德上市	阿里巴巴投资
台湾塑化剂事件	话题 0	暂停进口台湾问题	台湾塑化剂暂停进出口	暂停进口
2013 年两会	话题 13	国家最高权力机构	经济增长	经济发展
2013 年两会	话题 20	新华网记者	海洋旅游	海南旅游
2013 年两会	话题 29	山西生产	安全生产	西部发展
2013 年两会	话题 40	产能过剩	清洁能源	能源问题
2013 年两会	话题 50	交通运输	铁道部火车票改革	铁道部改革

从错误结果可以看出，错误原因包括以下几个方面：

（1） 部分 LDA 话题结果语义不明确，例如两会话题 20，话题关键词为“旅游”、“新华网”、“全国”、“江苏”、“建设”、“市长”、“人大代表”、“老百姓”、“记者”、“游客”，并不具有明显的语义信息。生成的关键短语只有“新华网记者”。

（2） 对动词词组的处理不完善，例如事件“2011 年春节”话题 3，关键短语为“回家过年”、“拜年方式”、“过年回家”，方法判定“回家过年”缺乏宾语，判定错误。方法在判断包含动词的短语和动词性短语的关系上有所欠缺。

（3） 部分短语泛化性偏高或偏低。例如事件“马云卸任 CEO”，抽取的标签为“阿里巴巴集团”，过于概括，不能表示具体的话题信息。例如两会事件话题 29 抽取的标签为“山西生产”，过于具体。方法在选择适中的泛化度上有待提升。

4.3.2 对比实验

本文方法同[9]中提出的方法进行了比较，均根据 LDA 话题结果生成话题标签短语。

表 8 对比实验结果

事件	话题	[9]话题标签	本文方法话题标签
巴勒斯坦申请加入联合国	话题 2	中方支持	中方支持巴勒斯坦
2010 年美国中期选举	话题 0	共和党众议员领袖	共和党赢得众议员控制权
台湾领导人选举	话题 2	两岸和平	两岸关系
台湾领导人选举	话题 0	台湾地区领导人选举	台湾地区
2011 春节	话题 4	同比增长	接待游客同比增长
2013 年两会	话题 5	无害化处理	垃圾无害化处理
2013 年两会	话题 12	完善	完善社会保障制度
2013 年两会	话题 14	老人	社区养老服务

2013 年两会	话题 16	我国国防	建设信息化军队
2013 年两会	话题 33	收入	收入分配制度

实验结果可以看出，本文的方法得到的短语能够表示特定的语义信息，例如两会话题 33，[9]标签为“收入”，而本文选择了“社区养老服务”，语义上更为完整；例如台湾领导人选举话题 2，本文标签为“两岸关系”比[9]“两岸和平”更泛化和确切。本文方法部分实验结果不如[9]中方法，例如台湾领导人选举话题 0，本文标签“台湾地区”泛化度偏高，不如[9]“台湾地区领导人选举”。

根据同样的标准答案，表 9 是两种方法精度的对比结果。可以看出本文方法的精度要高于[9]，在两会语料中提高精度 12%，在事件语料上提高精度 4%。说明短语的完整性以及泛化度考虑方法的合理性。

表 9 对比评测结果

语料	对比方案精度	本文方案精度
事件语料	23.5%	27.9%
两会语料	27.9%	39.5%

5 结论和展望

本文提出了一种基于种子词的话题标签抽取方法。方法首先根据提出的权重计算公式抽取每个话题的种子词，然后，采用 bootstrapping 思想，迭代产生包含种子词汇的关键短语集合，最后根据短语的完整性和泛化度选择话题标签。

本文对新闻事件语料和两会报告语料进行了实验，结果表明本文方法能够有效地抽取出话题标签，相对于方法[9]，本文抽取的短语完整性和概括性更高。本文主要的贡献是：将种子词抽取与 bootstrapping 方法引入到话题标签抽取的研究中；利用词性标注与短语结构信息抽取话题标签；根据短语的完整性和泛化原则，抽取表达力更强的标签短语。

本文的方法还存在很多不足之处，后续工作包括以下三个方面：研究题目信息与话题之间的关系；使用更有效的 LDA 结果重排序公式；将话题标签抽取工作融合进话题模型中，以短语为基本词汇单元，同时引入词性标注信息等信息

参考文献

[1] Blei David, Ng Andrew, Jordan Michael. Latent Dirichlet Allocation[J]. The Journal of Machine Learning Research, 2003, 3: 993-1022.

[2] 徐戈，王厚峰。自然语言处理中主题模型的发展，计算机学报[J]，2011-8 , 34(8): 1423-1436.

[3] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, 2004: 487-494.

[4] Ruifeng XU, Lu YE. Reader’s Emotion Prediction Based on Weighted Latent Dirichlet Allocation and Multi-label k-nearest Neighbor Model[J]. Journal of Computational Information System 9:6,2013.

[5] Johri N, Roth D, Tu Y. Experts' retrieval with multiword-enhanced author topic model. Proceedings of the NAACL HLT 2010 workshop on semantic search[C]. Association for Computational Linguistics, 2010: 10-18.

[6] William Darling, Fei Song.Probabilistic Topic and Syntax Modeling with Part-of-Speech LDA[C]. Association for Computational Linguistics. 2005

[7] Griffiths T L, Steyvers M, Blei D M, et al. Integrating topics and syntax[J]. Advances in neural information processing systems, 2005, 17: 537-544.

[8] Allison J.B. Chaney, David M.Blei. Visualizing Topic Models[C]. Association for the

Advancement of Artificial Intelligence. 2012
[9] 闫泽华. 基于 LDA 的新闻线索抽取研究 上海交通大学 2012 年硕士论文
[10] Teh Y W, Jordan M I, Beal M J, et al. Hierarchical dirichlet processes[J]. Journal of the American Statistical Association, 2006, 101(476).
[11] Blei D M, Lafferty J D. Visualizing topics with multi-word expressions[J]. arXiv preprint arXiv:0907.1013, 2009.
[12] Wallach H M. Topic modeling: beyond bag-of-words[C]//Proceedings of the 23rd international conference on Machine learning. ACM, 2006: 977-984.
[13] Wang X, McCallum A, Wei X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval[C]//Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. IEEE, 2007: 697-702.
[14] Nallapati R, Feng A, Peng F, et al. Event threading within news topics[C]//Proceedings of the thirteenth ACM international conference on Information and knowledge management. ACM, 2004: 446-453.
[15] Lau J H, Newman D, Karimi S, et al. Best topic word selection for topic labelling[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010: 605-613.
[16] Carmel D, Roitman H, Zwerdling N. Enhancing cluster labeling using wikipedia[C]//Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2009: 139-146.
[17] Song Y, Pan S, Liu S, et al. Topic and keyword re-ranking for LDA-based topic modeling[C]//Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009: 1757-1760.