软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(3):845−865 [doi: 10.13328/j.cnki.jos.005898]

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

基于 PSP_HDP 主题模型的非结构化经济指标挖掘*

张奕韬[1,2,3], 万常选[1,3], 刘喜平[1,3], 江腾蛟[1,3], 刘德喜[1,3], 廖国琼[1,3]

[1](江西财经大学 信息管理学院,江西 南昌 330013)

[2](华东交通大学 软件学院,江西 南昌 330013)

[3](数据与知识工程江西省高校重点实验室(江西财经大学),江西 南昌 330013)

通讯作者: 万常选, E-mail: wanchangxuan@263.net

摘 要: 随着经济活动数据的不断丰富,互联网平台上产生了大量的财经文本,其中蕴含了经济领域发展状况的影响因素.如何从这些财经文本中有效地挖掘与经济有关的经济要素,是实现非结构化数据在经济研究中应用的关键.根据人工构建非结构化经济指标的局限性,以及主题模型在非结构化经济指标挖掘中存在的问题,结合已有经济领域分类标准、词语之间的语义关系和词语对主题的代表性,定义了文档的领域隶属度、词语与主题的语义相关度和词语对主题的贡献度,用于分别描述 CRF(Chinese restaurant franchise)中餐厅的菜肴风格、顾客之间对菜肴要求的一致程度和顾客对菜肴的专一程度;结合文档领域属性、词语语义和词语在主题中的出现情况,提出了 PSP_HDP (combining documents' domain properties, word semantics and words' presences in topics with HDP)主题模型.由于 PSP_HDP 主题模型改进了文档-主题与主题-词语的分配过程,从而提高了经济主题的区分度和辨识度,可以更有效地挖掘与经济有关的经济主题和经济要素词.实验结果表明:提出的 PSP_HDP 主题模型不仅在主题多样性、内容困惑度和模型复杂度等评价指标方面的整体性能优于 HDP 主题模型,而且在非结构化经济指标挖掘和经济要素词抽取方面能够得到区分度更好、辨识度更高的结果.

关键词: HDP 主题模型;经济领域分类标准;语义关系;非结构化经济指标;经济要素词

中图法分类号: TP18

中文引用格式: 张奕韬,万常选,刘喜平,江腾蛟,刘德喜,廖国琼.基于 PSP_HDP 主题模型的非结构化经济指标挖掘.软件学报, 2020,31(3):845−865. http://www.jos.org.cn/1000-9825/5898.htm

英文引用格式: Zhang YT, Wan CX, Liu XP, Jiang TJ, Liu DX, Liao GQ. Mining unstructured economic indicators based on PSP_HDP topic model. Ruan Jian Xue Bao/Journal of Software, 2020,31(3):845−865 (in Chinese). http://www.jos.org.cn/1000-9825/5898.htm

Mining Unstructured Economic Indicators Based on PSP_HDP Topic Model

ZHANG Yi-Tao[1,2,3], WAN Chang-Xuan[1,3], LIU Xi-Ping[1,3], JIANG Teng-Jiao[1,3], LIU De-Xi[1,3], LIAO Guo-Qiong[1,3]

[1](School of Information Management, Jiangxi University of Finance and Economics, Nanchang 330013, China)

[2](School of Software, East China Jiaotong University, Nanchang 330013, China)

[3](Jiangxi Key Laboratory of Data and Knowledge Engineering (Jiangxi University of Finance and Economics), Nanchang 330013, China)

* 基金项目: 国家自然科学基金(61972184, 61562032, 61662027, 61762042); 江西省自然科学基金(20152ACB20003)

Foundation item: National Natural Science Foundation of China (61972184, 61562032, 61662027, 61762042); Natural Science Foundation of Jiangxi Province of China (20152ACB20003)

本文由人工智能赋能的数据管理、分析与系统专刊特约编辑李战怀教授、于戈教授和杨晓春教授推荐.

收稿时间: 2019-07-05; 修改时间: 2019-09-10; 采用时间: 2019-11-25; jos 在线出版时间: 2020-01-10

CNKI 网络优先出版: 2020-01-10 13:34:24, http://kns.cnki.net/kcms/detail/11.2560.TP.20200110.1333.002.html

Abstract: With the increasing enrichment of economic activity data, a large number of financial texts have emerged on Internet, which contains the influence factors of the economic development. How to mine these economic factors from these texts is the key to conduct economic analysis based on unstructured data. Due to the limitation of manual selection of economic indicators, and the inaccuracy of modelling economic indicators in unstructured texts, the CRF (Chinese restaurant franchise) allocation processes in HDP topic model are extended to a more efficient pattern. In order to describe the dish style in a restaurant, the existing economic taxonomies are used to determine the domain membership of a document. The semantic similarity between words is exploited to define the semantic relevance between words and topics, which reflect the similarity of customers' requirements for dishes. For each word, its representativeness of each topic is employed to evaluate its contribution to the topic, which explains the loyalty of a customer to each dish. By combining documents' domain properties, word semantics and words' presence in topics with HDP topic model, a novel model, PSP_HDP topic model, is proposed. As the PSP_HDP topic model improves documents-topics and topics-words allocation processes, it increases the accuracy of identifying economic topics and distinctiveness of the topics, which leads to a more effective mining of economic topics and economic factors. Experimental results show that the proposed model not only achieves a better performance in terms of topic diversity, topic perplexity and topic complexity, but also is effective in finding more cohesive unstructured economic indicators and economic factors.

Key words: HDP topic model; economic taxonomy; semantic relevance; unstructured economic indicator; economic factor

随着经济活动数据在数量、质量和表现形式上的不断丰富,以及自然语言处理、数据挖掘和机器学习等技术的迅速发展,经济研究不仅仅局限于结构化数据,也认识到非结构化数据的重要作用.文本数据是非结构化数据的主要形式之一,互联网平台中存在大量与财经领域相关的文本数据(财经文本),其中蕴含了个人或媒体对经济运行和发展中所遇到关键问题的观点和态度,间接反映了经济在投资、消费、进出口、政府财政和人口就业等方面的状况;同时,这类信息的产生和传播速度快,可实时反映经济的发展现状.因此,财经文本在获取经济信息、分析经济实时状况、辅助经济预测等方面有着独特的优势.在这种应用需求背景下,基于非结构化数据的经济研究面临了前所未有的机遇和挑战.

研究者证明:通过文本挖掘可发现与宏观经济有关的潜在影响要素(经济要素或经济要素词),利用搜索指数或情感指数量化这些经济要素,帮助预测经济发展趋势,弥补传统统计指标的时滞性和数据有意造假等问题带来的影响[1].已有研究主要通过人工筛选或结合 LDA 主题模型结果、手工选择这些经济要素词[2,3],再基于领域类别或主题划分,构建经济指标与经济要素词之间的概念隶属关系,称为非结构化经济指标(体系),最后,通过经济要素词的搜索指数量化非结构化经济指标,用于经济分析和预测.针对以上分析,实现非结构化数据在经济研究中的应用需要经过 3 个步骤:(1) 经济要素词的抽取和非结构化经济指标体系的构建;(2) 非结构化经济指标体系的量化;(3) 非结构化经济指标量化值在经济预测模型中的应用.其中,经济要素词的抽取和非结构化经济指标体系的构建,是研究的基础和关键步骤,也是本文主要关注点.

在经济要素词抽取和非结构化经济指标体系构建方面,已有研究主要通过手工或半自动的方式实现,存在抽取效率低、工作量大、人工成本高、数据覆盖率低以及领域的可移植性弱等缺陷.基于主题模型的主题-词语分析可对应于非结构化经济指标体系构建和经济要素词抽取,常用的主题模型包括 LDA(latent dirichlet allocation)和 HDP(hierarchical dirichlet process)模型.由于 LDA 主题模型对主题数目有很大依赖性[4],本文将采用 HDP 主题模型解决这个问题,实现非结构化经济指标体系构建和经济要素词抽取过程的全自动化.

然而,在经济领域中直接使用 HDP 主题模型生成经济主题、抽取经济要素词,存在如下主要问题:(1) 主题模型生成的主题无法体现经济主题的领域性;(2) 经济要素词无法准确对应到经济主题中,经济主题涵义不明确;(3) 很多经济要素词是中低频词,无法被有效地抽取出来.导致这些问题的根本原因是:在主题模型中,文档主题分布是由词语的共现情况决定,主题词是通过统计词语在文档中出现的频繁程度确定.所以,本文预期目标是结合财经领域的的分类信息提高文档主题与经济主题的匹配程度,利用词语之间的语义相似性改善词语在相同领域的共现频率,并基于词语的逆主题频率改进经济要素词在经济主题中的出现频率.

为了提高模型的领域适用性,本文将结合财经领域背景知识和词语之间的语义关系,改进 HDP 主题模型的 CRF(Chinese restaurant franchise)构造过程,实现财经文本中非结构化经济指标体系的自动构建和经济要素词

的自动抽取.基本思路是:根据已有经济领域分类标准定义文档的领域隶属度,改进 HDP 主题模型的文档-主题分配过程,明确财经文本的领域属性;为了提高经济要素词抽取的准确性,根据词语对经济主题的描述能力,基于词向量定义词语与主题的语义相关度,改进 HDP 主题模型的主题-词语分配过程,以便将语义相近的词语尽量分配到相同或相近的经济主题中,即:明确词语的经济主题属性,提高经济主题的区分度;为了进一步凸显经济要素词对经济主题的解释力,根据词语对经济主题的代表性定义词语对主题的贡献度,进一步改进 HDP 主题模型的主题-词语分配过程,以便能够抽取到有效的、中低频的领域主题专用词语,提高经济主题的辨识度.

本文的主要贡献是根据经济领域分类标准、词语对经济主题的描述能力(即词语之间的语义关系)以及词语对经济主题的代表性,定义文档的领域隶属度、词语与主题的语义相关度和词语对主题的贡献度,并分别映射到餐厅的菜肴风格、顾客之间对菜肴要求的一致程度和顾客对菜肴的专一程度.通过餐厅的菜肴风格、顾客之间对菜肴要求的一致程度以及顾客对菜肴的专一程度,改进CRF构造过程,对应于改进文档-主题分配过程和主题-词语分配过程,提出了 PSP_HDP 主题模型,并设计了相应的采样方法实现模型参数推导.在财经文本中进行实验,验证模型在构建非结构化经济指标体系和抽取经济要素词方面的有效性.

本文第 1 节将介绍财经领域中经济要素词抽取和非结构化经济指标体系构建的研究现状,以及主题模型的相关研究进展,并分析使用主题模型及其改进模型在非结构化经济指标体系构建和经济要素词抽取时存在的问题.第 2 节分析 HDP 主题模型的理论基础,并结合第 1 节提出的问题,定义文档的领域隶属度、词语与主题的语义相关度和词语对主题的贡献度等概念,通过概念映射改进CRF构造过程,目的是改进主题模型中的文档-主题和主题-词语分配过程,最终提出 PSP_HDP 主题模型.第 3 节为参数设置说明和实验结果分析.最后总结全文,并对未来值得关注的研究方向进行初步探讨.

1 相关研究

在抽取经济要素词和构建非结构化经济指标体系方面,刘涛雄等人[2]根据经济领域类别人工筛选经济要素词,构建非结构化经济指标;通过百度搜索指数对其量化,综合考虑非结构化经济指标和传统统计经济指标,提出基于“两步法”的经济指标预测模型,实现国内生产总值(gross domestic product,简称 GDP)预测.在金融风险和股票预测方面,类似的方法也用于实现金融指标的量化[3,4].

除了人工筛选经济要素词,Yakovleva[5]利用互联网新闻报道,基于 LDA 模型生成主题-词语分布,从中选取与采购经理人指数(purchase management index,简称 PMI)相关的主题词,结合 SVM(support vector machine)生成主题情感极性,结合主题情感时间序列模拟和预测 PMI,分析经济发展动态.Siegel[6]利用公司管理报告,基于本体的方法和 GRI(global reporting initiative)标准,提取关于社会、经济和生态方面的词语,实现句子分类.

综上可知,结合机器学习和自然语言处理技术抽取经济要素词和构建非结构化经济指标体系的研究还处于初级阶段,其中,主题模型提供了关键技术支持.本节接下来先综述主题模型的发展动态,再分析主题模型在构建非结构化经济指标体系和抽取经济要素词时存在的问题.

主题模型是基于概率的生成式模型,一篇文档的每个词语是通过“以一定概率选择了某个主题,并从这个主题中以一定概率选择某个词语”的过程得到.如果已知文档的词语分布,则可以通过概率推导得到文档的主题分布和主题的词语分布.主题模型是挖掘文档主题的重要工具,主要包括参数贝叶斯模型和非参数贝叶斯模型.

1.1 参数贝叶斯模型

Blei 等人[7]在文档-主题和主题-词语的先验分布中引入 Dirichlet 分布、设置相应超参数,构建 3 层贝叶斯模型,即 LDA 模型;利用主题分布差异和文档中词语共现信息,采用贝叶斯推断方法(Gibbs 采样或变分推断)计算文档-主题和主题-词语的后验概率分布,生成文档-主题分布和主题-词语分布.关于 LDA 主题模型的研究主要有以下几类.

(1) 基于领域知识的 LDA 模型

为了抽取文档中与领域相关的影响因素,即领域特征或领域词语,基于领域知识的主题模型 MDK-LDA (LDA with multi-domain knowledge)[8]和 AMC(automatically generated must-links and cannot-links)[9]通过频繁项

集、同义词等方法获取领域知识,目的是把语义相近的词语放在 must-link 集合中,不共现的词语放在 cannot-link 集合中,用于约束主题-词语的分配过程.AKL(automated knowledge LDA)模型[10]首先从不同领域的商品评论中自动获取先验知识,然后在主题模型中加入先验知识,指导商品特征的识别,改善商品特征的提取.

(2) 基于词向量的 LDA 模型

随着神经网络方法的广泛使用,词向量对词语的表达变得更加丰富,使得词语之间的语义关系可以度量得更加准确.基于词向量的主题模型可以增强语义相似的词语在同一主题上的分配概率,提高主题模型的性能.人们在分析文档中某个词语的涵义时,不仅考虑与其共现的词语,还考虑词语之间的语义关系.主题模型主要根据词语的共现信息计算主题-词语的概率分布,缺少词语之间的语义信息.GPU-DMM(generalized polya urn dirichlet multinomial mixture)模型[11]针对主题模型中存在的上述问题,从大规模语料中学习词语的语义关系,在主题模型的采样过程中,通过 Bernoulli 分布决定是否利用词向量抽取文档在某个主题下的相关词语,目的是将语义相近的词语尽量分配在相同的主题中,同时剔除无关词语,提高主题词对主题的解释效果,改善主题模型的效果.

由于 LDA 主题模型中文档-词语的生成概率是由文档-主题和主题-词语的概率分布共同决定,导致一些在主题-词语概率分布中主题隶属概率较高的词语,即高频词,在主题中占有明显优势.为了捕获与目标主题相关的中低频词语,TWE(topic word embeddings)模型[12]基于 LDA 主题模型建立主题-词语关系,并且在构建词向量的同时引入主题向量,使得这两种向量处于同一向量空间;通过比较词向量和主题向量之间的相似性,选取与文档主题最相关的词语.

(3) 基于 LDA 模型的微博主题挖掘

在微博主题挖掘中,Das 等人[13]通过挖掘主题短语或特征词发现当前关注度较高的热点话题,用于趋势分析和观点挖掘.张晨逸等人[14]利用微博中的转发和回复信息与被转发和被回复信息之间存在相似主题的特点,综合考虑微博的联系人关系,改进主题分配概率,提出了 MB-LDA(micro blog LDA)模型.庞雄文等人[15]利用微博用户之间的转发、对话、点赞和评论关系计算微博之间的相关性,基于同一用户、相邻时间片段间微博主题具有强相关性的假设,改进文档主题概率分布,提出了 MRT-LDA 模型,改善微博主题聚类效果.

然而,这类 LDA 主题模型都需要预先设定主题数目,在不具备任何先验知识的情况下,很难准确给定固定的主题数目;随着时间推移,文档主题会有新旧主题的更替,增加了主题数目的不确定性;通过不断测试来确定主题数目的方法,将耗费大量的时间和精力,限制了模型的应用推广.

1.2 非参数贝叶斯模型

HDP 主题模型[16]是 LDA 主题模型的非参数形式推广,可实现主题共享,用于构建无穷多个主题的混合模型;给定文档集,通过后验概率推导可以自动确定主题数目,增强了模型的适用性.HDP 主题模型假设数据是可交换的:一是“对内”,假设文档内部词语次序的交换并不影响主题的概率分布;二是“对外”,假设文档之间的次序与主题分布无关.然而,可交换性假设破坏了文档之间的依赖性和时序性.关于 HDP 主题模型的研究主要有以下几类.

(1) 考虑依赖关系的 HDP 模型

Kim 等人[17]考虑到数据“外部信息”之间的依赖关系,提出了 ddCRP(distance dependent Chinese restaurant process)模型,对于不可交换的、序列数据的主题聚类,在“餐桌-菜肴”分配的先验概率中考虑了相邻文档之间的距离关系,通过衰减函数调整距离对主题划分的影响,改善文档-主题概率分布,降低了模型的内容复杂度.Li 等人[18]在文档内容的基础上添加关键词、引用信息、链接信息和合作者信息,计算文档间距离,改进 CRP 结构,提出了 SID-CRP(side information dependent CRP)模型,用于提高文档集主题聚类效果.与此同时,Blei 等人[19]从数据“内部关系”的角度,根据数据的时间和空间距离关系决定“顾客-餐桌”分配过程,基于数据本身的依赖关系改善数据聚合机制.

Ahmed 等人[20]考虑了文本之间的时间依赖关系,解决 HDP 主题模型中数据的可交换问题,提出了 iDTM (infinite dynamic topic models)模型.该模型通过在 CRF 中加入固定的时间段设置,以便发现主题-词语的时序分

布、动态的主题数和主题的流行度.Ma 等人[21]利用 HDP 模型分析固定间隔的时间片段间主题和主题词的变化规律,获取主题进化模式.Zhang 等人[22]提出了 Evo-HDP(evolutionary hierarchical dirichlet process)模型,考虑到文本之间的时间依赖性,为每个固定时间段的多个语料库同时建立 HDP 模型,基于马尔可夫链假设,相邻时间段的聚类模式是强相关的,用 HDP 实现语料库内与不同语料库之间在不同时间段的主题共享和主题演化模式.为了挖掘和建模主题演化的分支结构,Wang 等人[23]提出了 EDP(evolving dirichlet processes)和 EHDP(evolving hierarchical dirichlet processes)模型,用于构建时序文本数据集的非线性主题进化轨迹.

(2) 基于领域知识的 HDP 模型

在领域主题挖掘方面,刘少鹏等人[24]提出了 MB-HDP(micro blog HDP)模型,在微博主题挖掘中考虑了微博发布时间、用户信息和话题标签,用 ddCRP 改进 HDP 的顶层结构,把发布时间、用户信息和话题标签相同的文档聚在相同主题下,改进文档-主题的分配过程,解决短文档主题聚类问题.Qian 等人[25]提出了 sHDP(social HDP)模型,结合文档内容和社会网络结构信息,利用不同的社会群体结构设计文档-主题参数的混合权重,改善短文档主题聚类效果.Yang 等人[26]提出了 HDPauthor 模型,结合作者列表信息,在 HDP 模型中增加一层用于表示作者之间共享的主题分组,并将混合的文档-作者-主题分布代替 HDP 模型中的文档-主题分布,构建基于作者列表信息的生成模型,用于挖掘相关文档中作者的主题兴趣.

在财经领域应用方面,有监督的主题模型 HDP-IR[27]可用于分析产品质量和预测产品销售;向量自回归无限隐马尔可夫模型[28]利用自动学习的状态数捕获股票市场收益对未来经济增长率的预测能力.

在主题结构研究方面,由于 LDA 主题模型假设主题是孤立的,导致模型无法研究主题之间的关联关系.Blei 等人[29]基于 nCRP(nested CRP)的原理提出了 HLDA(hierarchical LDA)模型,解决 LDA 模型中主题孤立问题,结合改进的采样算法提取文档集的主题层次结构.为了改进 nCRP 的推导效率,Chen 等人[30]在主题共享的基础上分析主题之间的关联关系,结合 LDA 主题模型实现主题层次结构的提取,提出可扩展的推导过程(scalable inference),实现大规模文本语料库主题结构和数目的自动推导.

目前,HDP 主题模型除了为文本主题挖掘提供有效的分析工具,还在视频监控分析、图像理解与标注、认知研究等方面得到了广泛应用[31].

上述 HDP 改进模型主要利用文本的时间信息、用户信息等附加信息改善主题聚类效果,并实现主题共享和主题数目自动生成.但是这类模型仍然存在一些不足:一是在确定文档-主题分布时没有考虑文档的领域特性,导致主题的领域类别不明显;二是在生成主题-词语的分布时没有考虑词语对主题的描述能力和代表性,导致无法提取到财经文本中蕴含的大量经济要素词.具体表现如下.

- 词语存在领域归属问题,例如“外卖”“快递”“手机”“奶粉”“酒”“书”等能够反映经济指标在消费领域的表现,“贷款”“融资”“利率”“基金”“房价”等能够反映经济指标在投资领域的表现,“石油”“关税”“芯片”“期货”“奶粉”“大豆”等能够反映经济指标在进出口领域的表现,“收入”“税收”“增值税”“教师”“税率”“农民工”等能够反映经济指标在政府财政领域的表现,“兼职”“费率”“社保”“个税”“工伤”等能够反映经济指标在人口就业领域的表现;
- 有些词语并不能描述经济主题.例如,某财经文档的分词结果为“甲醛 房租 蛋壳 租金 居室 房子 房租 房屋 时间 拆除 身体 红包 机构 房间 甲醛 人员 理论 对方 时间 App 甲醛 房租 蛋壳 租金”,该文档反映了经济指标在消费领域的主题信息,其中,“甲醛”“房租”“房子”等词语与该文档主题存在较明确的语义相关性,但“时间”“红包”“机构”“人员”“理论”“对方”等词语与该文档主题的语义相关性并不大;
- 领域主题专用词语在主题-词语概率分布中分配了较低的概率.例如在人口就业领域:一方面,“企业”“人员”“个人”“制度”“社会”“公司”等高频词语与该领域有关联,但是这些词语无法凸显该领域的主题内容;另一方面,“税法”“待遇”“个税起征点”“医保”“补贴”等中低频的领域主题专用词语,能够较好地代表该领域的主题信息.

基于上述问题,一方面,本文将利用已有的经济领域代表性词语集指导经济主题划分过程,建立主题和经济

领域之间的相关关系,明确主题在经济领域的涵义;另一方面,结合词语对主题的描述能力(即词语之间的语义关系)以及词语对主题的代表性,辅助经济要素词的筛选,不仅提高经济主题之间的区分度,也提高经济主题的辨识度.

2 HDP 主题模型建模

2.1 HDP主题模型

HDP 主题模型包括两层 DP(Dirichlet process,狄利克雷过程):第 1 层 DP 以基分布 H 和超参数γ为参数,抽样生成全局随机概率测度,记为 G_0;第 2 层 DP 以全局随机概率测度 G_0 和超参数α为参数,为第 j 篇文档抽样生成一个概率测度,记为 G_j.如公式(1)所示:

$$\begin{aligned} &G_0 \mid \gamma, H \sim DP(\gamma, H) \\ &G_j \mid \alpha, G_0 \sim DP(\alpha, G_0) \end{aligned} \tag{1}$$

由公式(1)可知,第 j 篇文档的概率测度 G_j 均来源于参数为 G_0 的 DP 过程,保证了文档之间的主题共享.设 $\{\theta_{j1},\theta_{j2},\ldots,\theta_{ji}\}$是服从 G_j 的独立同分布的随机变量序列,该序列的先验分布来源于基分布 H,其中,θ_{ji} 对应词语 x_{ji} 的主题分布参数,$F(\theta_{ji})$表示在给定参数 θ_{ji} 下词语 x_{ji} 的主题分布,如公式(2)所示:

$$\begin{aligned} &\theta_{ji} \mid G_j \\ &x_{ji} \mid \theta_{ji} \sim F(\theta_{ji}) \end{aligned} \tag{2}$$

HDP 主题模型通过 CRF(Chinese restaurant franchise)构造方法实现采样过程,进而推断模型参数的后验概率分布.假设所有餐厅共用一份菜单,每张餐桌只供应一道菜肴,顾客-餐桌-菜肴分配过程如下:首先给新顾客分配餐桌,当餐桌中分配了第 1 位顾客后,需要为该餐桌分配菜肴;后续顾客可选择已有餐桌,享用已有菜肴,也可选择新的餐桌,并为该餐桌选择新的菜肴或已有菜肴.整个分配过程允许多个餐厅中的多个餐桌分配相同菜肴.

为了叙述方便,先基于 CRF 构造方法定义如下符号.

- J 表示所有餐厅的集合,j 表示某个餐厅;
- X_j 表示第 j 个餐厅中的顾客集合,x_{ji} 表示 j 餐厅第 i 个顾客,则 $X=\{x_{ji}|1\leqslant j\leqslant|J|,1\leqslant i\leqslant|X_j|\}$表示所有顾客集合,$|\cdot|$表示集合中的元素个数;
- T 表示所有餐桌的集合,T_j 表示第 j 个餐厅中已分配顾客的餐桌集合;
- t_{ji} 表示第 j 个餐厅第 i 个顾客所坐的餐桌, $X_j^t=\{x_{ji}\mid x_{ji}\in X, t\in T, t_{ji}=t\}$ 表示第 j 个餐厅中就座餐桌 t 的顾客集合;
- θ_{ji} 表示分配第 j 个餐厅中第 i 个顾客就座餐桌的分布参数,顾客选择一张新的餐桌就座,则需要分配菜肴,记分配菜肴的分布参数为 $\varphi_j^{t_{ji}}$,因此结合餐桌和菜肴的对应关系可知,θ_{ji} 与 $\varphi_j^{t_{ji}}$ 一一对应;
- Φ表示菜单上菜肴的集合,ϕ^k 表示菜肴 k 的分布参数,T^k 表示供应菜肴 k 的餐桌集合;
- $K=\{k_j^t\mid k_j^t\in\Phi, 1\leqslant j\leqslant|J|, t\in T\}$ 表示所有餐桌已供应的菜肴集合,由于每个餐桌只供应一道菜肴,因此 φ_j^t 与 $\phi^{k_j^t}$ 一一对应;
- $\delta_{\varphi_j^t}$ 表示顾客在第 j 个餐厅中就座餐桌 t 的概率分布参数;
- δ_{ϕ^k} 表示分配菜肴 k 的概率分布参数.

在 CRF 中,为新顾客分配餐桌时,新顾客被分配到已有餐桌的概率与该餐桌所坐顾客的数量成正比,被分配到新餐桌的概率与超参数α成正比,生成第 j 个餐厅中第 i 个顾客 x_{ji} 就座餐桌的分布参数 θ_{ji},如公式(3)所示;为新餐桌分配菜肴时,分配已有菜肴的概率与供应该菜肴的餐桌数成正比,被分配到新菜肴的概率与超参数γ成正比,生成第 j 个餐厅中餐桌 t 供应菜肴的分布参数 φ_j^t,如公式(4)所示:

$$\theta_{ji}\mid\theta_{j1},\theta_{j2},...,\theta_{j(i-1)},\alpha,G_0\sim\sum_{t\in T_j}\frac{\mid X_j^t\mid}{i-1+\alpha}\delta_{\varphi_j^t}+\frac{\alpha}{i-1+\alpha}G_0 \tag{3}$$

$$\varphi_j^t\mid\varphi_1^1,...,\varphi_1^t,\varphi_2^1,...,\varphi_2^t,...,\varphi_j^1,...,\varphi_j^{t-1},\gamma,H\sim\sum_{k\in K}\frac{\mid T^k\mid}{\mid T^k\mid+\gamma}\delta_{\phi^k}+\frac{\gamma}{\mid T^k\mid+\gamma}H \tag{4}$$

整个 CRF 分配过程与文档-主题分析中的主题-词语分配、文档-主题分配过程是对应的,通过采样和参数后验概率推导构造 HDP 主题模型,生成文档-主题概率分布和主题-词语概率分布.

2.2 PSP_HDP主题模型

在财经文本挖掘中,假设一篇文档只关注一个经济领域,这种假设对于大部分财经文本主题分析是合理的.结合财经文本领域属性、词语语义和词语在主题中的出现情况改造 HDP 主题模型,构建 PSP_HDP(combining documents' domain properties, word semantics and words' presences in topics with HDP)主题模型.假设餐厅对应文档、顾客对应词语、菜肴对应经济主题,餐厅的菜肴风格对应财经文本的领域属性.根据 HDP 主题模型在财经文本主题分析中存在的问题,结合 HDP 主题模型的 CRF 构造过程,定义以下概念.

(1) 文档的领域隶属度

在CRF构造餐桌-菜肴的分配过程中,新餐桌菜肴的分配仅取决于已有菜肴被分配的餐桌数,没有考虑餐桌所在餐厅的菜肴风格.给某个餐厅的餐桌分配菜肴之前,应分析该餐桌所在餐厅的菜肴风格,目的是对同一餐厅的餐桌以及具有相同菜肴风格的其他餐厅的餐桌分配相同风格的菜肴.相当于在文档-主题分配时,应明确文档所属经济领域类别.

参照已有研究[2]对经济领域类别的划分,记经济领域类别集合为Ω,给定$\rho\in\Omega$,d_ρ表示经济领域类别ρ的代表性词语集合;利用文档与经济领域类别词语集的相似度,划分文档的经济领域类别.设 d_j 和 $d_{j'}$分别表示文档 j 和 j'的词语集合,I 表示指示函数,sim 表示词语集之间的相似度计算函数;根据与文档 j 的经济领域类别相同的文档数,定义文档 j 隶属于该经济领域类别的程度,即文档 j 的领域隶属度,记为 A_j,如公式(5)所示:

$$A_j=\sum_{j'\neq j}I(\arg\max_{\rho\in\Omega}(sim(d_{j'},d_\rho))=\arg\max_{\rho\in\Omega}(sim(d_j,d_\rho))) \tag{5}$$

财经文档的领域隶属度描述了文档的领域属性,明确了财经文档的领域类别归属,建立了财经文档主题与经济领域间的相关性.但是文档词语对主题的解释能力依然不明确,因此需要进一步改进顾客-餐桌的分配过程,改善经济主题的区分度和辨识度.

(2) 词语与主题的语义相关度

一个经济领域类别下会有多个经济领域主题,如消费领域类别下可能会有家电消费、餐饮消费等消费领域主题,相当于一个菜肴风格下可能会有多个菜肴类别.在 CRF 构造顾客-餐桌的分配过程中,新顾客的餐桌分配取决于已有餐桌的顾客数,忽略了顾客之间对菜肴类别要求的差异性.新顾客分配餐桌时,应根据新顾客与已分配餐桌的顾客之间对菜肴类别要求的一致程度,决定顾客所坐的餐桌,即对菜肴类别要求一致或相近的顾客应该分配在相同的餐桌.在财经文本中,词语之间的语义相似性可反映词语在经济领域主题空间上的距离信息,即语义相似性较高的词语在潜在经济领域主题空间中应该距离更近,在主题分配时更应该被分配在相同的经济领域主题中.缺少词语语义分析,将导致一些与经济领域主题语义相关性不高的词语,也分配到该主题中,降低了经济领域主题的区分度.本文把词语之间的语义相似性对应成顾客之间对菜肴类别要求的一致性,语义相似性高的词语表示对菜肴类别要求一致或相似的顾客.

新顾客选择餐桌时,需要度量新顾客与已分配餐桌的顾客之间对菜肴类别要求的一致程度,选择与自己菜肴类别要求相同或相近的餐桌就座.因此,本文需要在当前已分配餐桌的顾客中,计算每张餐桌中与新顾客对菜肴类别要求一致的顾客数.

为了考察词语对经济主题的描述能力,本文通过词向量计算词语之间的语义相似性,统计待分配主题的词语 w_i 在主题 k 的词语集 W_k 中语义相似的词语个数,称为词语 w_i 与主题 k 的语义相关度,记为 $A(w_i,W_k)$,如公式(6)所示.其中,sr 表示词语之间的语义相似性,ξ表示语义相似性的阈值:

$$A(w_i,W_k)=|\{w_j|sr(w_i,w_j)\geqslant\xi,w_j\in W_k,w_j\neq w_i\}| \tag{6}$$

在主题-词语的分配过程中,将待分配主题的词语 w_i 分配给语义相关度最大的主题.词语与主题的语义相关度描述了词语的主题属性.在财经文本中,通过计算词语与主题的语义相关度,可以将描述相同经济领域主题的词语分配在同一主题中,提高主题模型中词语对主题的描述能力,提升经济主题的区分度.

(3) 词语对主题的贡献度

顾客-餐桌和餐桌-菜肴分配完成后,主题模型按照顾客在每个菜肴中的出现频次计算顾客享用每个菜肴的概率,用于区分不同菜肴的顾客群体.这种计算方式使得一些出现频率高、菜肴喜好特点不明显的顾客在群体划分时将占很大优势.然而,不同顾客对菜肴的喜好存在不同的专一程度,即:有些顾客会频繁出现在不同餐桌中品尝各种菜肴,这类顾客属于普通顾客;也有些顾客只会在特定的餐桌中出现,品尝特定菜肴,这类顾客属于专一顾客.对应于主题-词语分配,有些词语可用在文档的各种主题中,虽然出现频次较高,但对辨识领域主题的贡献不大,属于通用词语;有些词语只在文档的特定领域主题中出现,虽然总体出现频次不高,但能明确凸显领域主题的内容,反映了词语对领域主题的代表性,属于领域主题专用词语.所以,在明确文档的领域属性和词语的主题属性后,通过分析词语在领域主题中的代表性,改进主题-词语分配概率计算方法,提高中低频的领域主题专用词语分配到相应经济领域主题中的概率,凸显领域主题专用词语对领域主题的贡献,进一步提高经济领域主题的辨识度.

设主题-词语分配的后验概率分布为 phi,$phi(w_i,k)$表示词语 w_i 在主题 k 中出现的概率,$N(phi)$表示后验概率推导中得到的主题数目,$N_{w_i}(phi)$ 表示在分布 phi 生成的所有主题中出现词语 w_i 的主题数目;$\log(N(phi)/N_{w_i}(phi))$ 表示词语的逆主题频率,用于描述词语对主题的代表性.用词语 w_i 在主题 k 中的出现概率 $phi(w_i,k)$乘以词语 w_i 的逆主题频率 $\log(N(phi)/N_{w_i}(phi))$ (即词语 w_i 在主题 k 中的权重概率值),表示词语 w_i 对主题 k 的贡献度,记为 $C(w_i,k)$,如公式(7)所示:

$$C(w_i,k)=\begin{cases}phi(w_i,k)*\log\dfrac{N(phi)}{N_{w_i}(phi)}, & \text{if } \log(N(phi)/N_{w_i}(phi))\geqslant\eta\\ \sigma, & \text{else}\end{cases} \tag{7}$$

其中,η表示词语对主题的代表性阈值,σ表示主题中词语的概率均值.当词语 w_i 对主题的代表性不强,即 $\log(N(phi)/N_{w_i}(phi))<\eta$,则用$\sigma$表示词语 w_i 对主题 k 的贡献度.

对照文档、主题、词语与餐厅、菜肴、顾客之间的映射关系,将文档的领域隶属度对应成餐厅的菜肴风格,用于在餐桌分配菜肴时考虑其所在餐厅的菜肴风格;将词语与主题的语义相关度对应成顾客之间对菜肴类别要求的一致程度,目的是将当前待分配顾客分配在符合其菜肴类别要求的餐桌中.根据餐厅的菜肴风格和顾客之间对菜肴类别要求的一致程度,改进 CRF 分配过程,如图 1 所示.

对于图 1,在顾客-餐桌层 CRP(Chinese restaurant process)中,根据顾客之间对菜肴类别要求的一致程度决定顾客的餐桌选择,其中,长方形表示餐厅,大圆表示餐桌,小圆、小平行四边形、小正方形表示顾客的分布参数,不同形状代表顾客对菜肴风格要求的不同;在餐桌-菜肴层 CRP 中,结合餐厅的菜肴风格决定餐桌的菜肴选择,其中,大圆表示菜肴,小圆表示餐桌,不同背景表示不同菜肴风格的餐桌;通过以上两层 CRP 结构,构建 PSP_HDP 主题模型的 CRF 过程.

根据改进的 CRF 构造过程,得到 PSP_HDP 主题模型中对应参数的概率分布.首先,当新顾客分配餐桌时,新顾客被分配到已有餐桌的概率与该餐桌中已有顾客和新顾客之间对菜肴类别要求的一致程度成正比,被分配到新餐桌的概率与超参数α成正比,生成第 j 个餐厅中第 i 个顾客 x_{ji} 就座于第 j 个餐厅餐桌 t 的分布参数 θ_{ji}.根据公式(6)的定义,将第 j 个餐厅 t 餐桌中已有顾客和新顾客之间对菜肴类别要求的一致程度记为 $A(x_{ji},X_j^t)$,X_j^t 表示第 j 个餐厅中就座餐桌 t 的顾客集合,顾客餐桌分配过程的参数如公式(8)所示:

$$\theta_{ji}\,|\,\theta_{j1},\theta_{j2},...,\theta_{j(i-1)},\alpha,G_0\sim\sum_{t\in T_j}\frac{A(x_{ji},X_j^t)}{\sum_{t\in T_j}A(x_{ji},X_j^t)+\alpha}\delta_{\varphi_j^t}+\frac{\alpha}{\sum_{t\in T_j}A(x_{ji},X_j^t)+\alpha}G_0 \tag{8}$$

因此,新顾客可以选择与自己的菜肴类别要求一致程度较高的餐桌就座,即θ_{ji}对应φ_j^t,且餐桌编号为 t;也可选择新餐桌就座,并由 G_0 采样生成新餐桌.

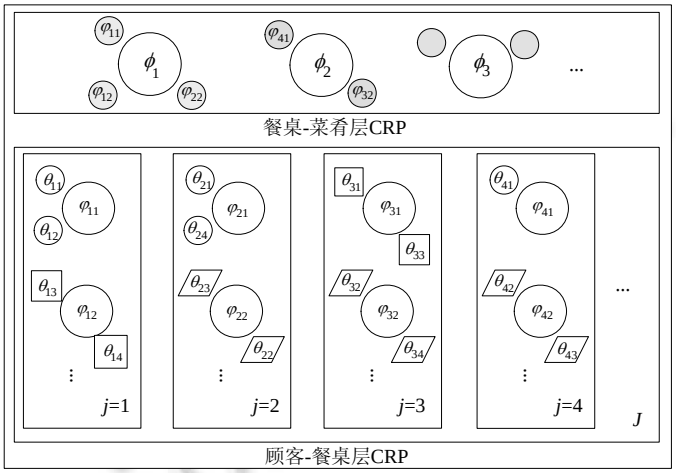


Fig.1 Depiction of CRF for PSP_HDP topic model

图 1 PSP_HDP 主题模型的 CRF 结构

如果新顾客选择新餐桌就座,则需要为新餐桌分配菜肴.按照餐桌-菜肴的分配原则,新餐桌分配已有菜肴的概率与相同菜肴风格的餐厅中已有菜肴的供应数成正比,分配到新菜肴的概率与超参数γ成正比.结合公式(5)的文档领域隶属度定义和餐桌-菜肴分配原则,计算与餐厅 j 的菜肴风格相同的所有餐厅中供应菜肴 k 的餐桌数 A_j^k,如公式(9)所示,其中,k_{jt} 为餐厅 j 中餐桌 t 供应的菜肴,$k_{j't'}$为餐厅 j'中餐桌 t'供应的菜肴:

$$A_j^k=\sum_{\substack{j'\neq j,t\in T_j,t'\in T_{j'}\\ \text{s.t. } k_{jt}=k\wedge k_{j't'}=k}} I\left(\arg\max_{\rho\in\Omega}(sim(d_{j'},d_\rho))=\arg\max_{\rho\in\Omega}(sim(d_j,d_\rho))\right) \tag{9}$$

结合 A_j^k 的定义改进餐桌-菜肴分配过程,生成第 j 个餐厅中餐桌 t 供应菜肴的分布参数 φ_j^t.因此,新餐桌可分配与其所在餐厅菜肴风格一致的已有菜肴;也可分配由 H 采样生成的新菜肴,如公式(10)所示:

$$\varphi_j^t\mid\varphi_1^1,...,\varphi_1^t,\varphi_2^1,...,\varphi_2^t,...,\varphi_j^1,...,\varphi_j^{t-1},\gamma,H\sim\sum_{k\in K}\frac{A_j^k}{\sum_{k\in K}A_j^k+\gamma}\delta_{\phi^k}+\frac{\gamma}{\sum_{k\in K}A_j^k+\gamma}H \tag{10}$$

根据顾客-餐桌分配、餐桌-菜肴分配的结果,生成菜肴-顾客的概率分布.因此,结合菜肴-顾客的概率分布和顾客在不同菜肴中的出现情况,计算顾客对菜肴的专一程度,形成 PSP_HDP 主题模型.在财经文本主题分析中,顾客-餐桌分配过程对应词语的参数概率分布,餐桌-菜肴分配过程对应文档-主题概率分布,用公式(8)和公式(10)改进经济要素词的抽取和经济主题的生成.顾客对菜肴的专一程度对应词语对主题的贡献度,用公式(7)更新词语在主题中的概率取值,以便抽取能较好地代表经济主题的、中低频的领域主题专用词语.

2.3 模型采样

已知文档词语和主题的先验分布,基于改进的 CRF 分配过程,完善模型参数 Gibbs 采样过程,生成模型参数的后验概率分布.参照 Whye 等人[16]的采样策略,不直接对θ_{ji}和φ_j^t采样,通过采样对应的索引变量 t_{ji}和k_j^t,并利用这些索引变量和ϕ^k重构生成θ_{ji}和φ_j^t.

(1) 计算变量 x_{ji}和 x_{jt}的条件概率

在采样索引变量 t_{ji}和k_j^t时,需要分别以单个顾客 x_{ji}和单张餐桌顾客 X_j^t为采样对象,因此首先需要计算变量 x_{ji}和 X_j^t的条件概率,其中,$x_{ji}\in X$.在文档-主题-词语中,已知 F 和 H 的密度函数分别为 f 和 h,在采样过程中,给

定主题 k 以及除词语 x_{ji} 以外的其他所有词语,定义 x_{ji} 的条件概率,如公式(11)所示:

$$f_k^{-x_{ji}}(x_{ji})=\frac{\int f(x_{ji}\mid\phi^k)\prod_{j'i'\neq ji,z_{j'i'}=k}f(x_{j'i'}\mid\phi^k)h(\phi^k)\mathrm{d}\phi^k}{\int\prod_{j'i'\neq ji,z_{j'i'}=k}f(x_{j'i'}\mid\phi^k)h(\phi^k)\mathrm{d}\phi^k} \tag{11}$$

这里,将除某个变量以外的剩余变量部分通过上标中的减号(–)表示.

根据基分布函数 H 和主题-词语分布函数 F 的共轭特性,化简公式(11)得公式(12):

$$f_k^{-x_{ji}}(x_{ji})=\begin{cases}\dfrac{n\_kv[v]}{|X^k|}, & k\in K\\ \dfrac{1}{|X|}, & k=k_{new}\end{cases} \tag{12}$$

其中,X 表示所有词语集合,X^k 表示主题 k 的词语集合,v 是采样过程中词语 x_{ji} 对应的索引变量,$n_kv[v]$表示主题 k 中索引值为 v 的词语数.

类似地,给定主题 k 以及词语所在分组以外的其他所有词语,定义第 j 个文档中 t 组词语集合 X_j^t 的条件概率,其中一组词语对应于 CRF 分配过程中一张餐桌的顾客,如公式(13)所示:

$$f_k^{-X_j^t}(X_j^t)=\begin{cases}\dfrac{\Gamma(|X^k|)}{\Gamma(|X^k|+|X_j^t|)}\dfrac{\prod_v\Gamma(n\_kv[v]+n\_jtv[v])}{\prod_v\Gamma(n\_kv[v])}, & k\in K\\ \dfrac{\Gamma(|X|\beta)}{\Gamma(|X|\beta+|X_j^t|)}\dfrac{\prod_v\Gamma(\beta+n\_jtv[v])}{\prod_v\Gamma(\beta)}, & k=k_{new}\end{cases} \tag{13}$$

其中,X^k 表示主题 k 的词语集合,$n\_jtv[v]$,$n_kv[v]$分别表示词语集合 X_j^t,X^k 中索引值为 v 的词语数,β表示主题分布参数.

(2) 采样 t

索引变量 t_{ji} 对应参数θ_{ji},t_{ji} 的后验概率由 t_{ji} 的先验概率乘以 x_{ji} 的条件概率得到.当顾客选择已有餐桌时,t_{ji} 的先验概率由 $A(x_{ji},X_j^t)$ 决定,x_{ji} 的条件概率为 $f_{k_j^t}^{-x_{ji}}(x_{ji})$,由公式(12)计算;当顾客选择新餐桌时,$t_{ji}$ 的先验概率由 α决定,x_{ji} 的条件概率可结合公式(10)中新餐桌菜肴分配过程计算,记为 $p(x_{ji}|t_{ji}=t_{new},t^{-ji},k)$,如公式(14)所示:

$$p(x_{ji}\mid t_{ji}=t_{new},t^{-ji},k)=\sum_{k\in K}\frac{A_j^k}{\sum_{k\in K}A_j^k+\gamma}f_{k_j^t}^{-x_{ji}}(x_{ji})+\frac{\gamma}{\sum_{k\in K}A_j^k+\gamma}f_{k_{new}}^{-x_{ji}}(x_{ji}) \tag{14}$$

其中, $f_{k_{new}}^{-x_{ji}}(x_{ji})=\int f(x_{ji}\mid\phi_{k_{new}})h(\phi_{k_{new}})\mathrm{d}\phi_{k_{new}}$.

综上,t_{ji} 的后验概率表示如公式(15)所示:

$$p(t_{ji}=t\mid t^{-ji},k,x)\propto\begin{cases}A(x_{ji},X_j^t)f_{k_j^t}^{-x_{ji}}(x_{ji}), & t_{ji}\in T_j\\ \alpha p(x_{ji}\mid t_{ji}=t_{new},t^{-ji},k), & t_{ji}=t_{new}\end{cases} \tag{15}$$

如果顾客选择新的餐桌就坐,需要继续为新的餐桌分配菜肴 $k_j^{t_{new}}$.结合餐桌-菜肴分配机制,新餐桌分配已有菜肴的概率与相同菜肴风格的餐厅中已有菜肴的供应数成正比,分配到新菜肴的概率与超参数γ成正比,如公式(16)所示:

$$p(k_j^{t_{new}}=k\mid t,k_j^{-t_{new}})\propto\begin{cases}A_j^k f_k^{-x_{ji}}(x_{ji}), & k\in K\\ \gamma f_{k_{new}}^{-x_{ji}}(x_{ji}), & k=k_{new}\end{cases} \tag{16}$$

因为在顾客分配餐桌及餐桌分配菜肴的采样过程中,t_{ji} 的更新将导致一些餐桌的顾客数变为 0,导致一些餐桌变成空桌,因此采样过程中需要更新所有已分配顾客的餐桌信息.类似地,t_{ji} 的更新也会导致一些菜肴没有被

分配,导致一些菜肴需要下架,因此采样过程中需要更新所有已分配餐桌的菜肴信息.

(3) 采样 k

由于 t_{ji} 的更新会影响菜肴的分配信息,采样 k 时,结合单张餐桌顾客 X_j^t 的条件概率,计算菜肴分配参数 k_j^t 的后验概率.索引变量 k_j^t 对应参数 φ_j^t,k_j^t 的后验概率由 k_j^t 的先验概率乘以 X_j^t 的条件概率得到.当餐桌分配已有菜肴时,k_j^t 的先验概率由 A_j^k 决定,X_j^t 的条件概率为 $f_k^{-X_j^t}(X_j^t)$;当餐桌分配新菜肴时,k_j^t 的先验概率由γ决定,X_j^t 的条件概率可结合公式(13)计算得到.因此,k_j^t 的后验概率表示如公式(17)所示:

$$p(k_j^t=k\mid t,k_j^{-t})\propto\begin{cases}A_j^k f_k^{-X_j^t}(X_j^t), & k\in K\\ \gamma f_{k_{new}}^{-X_j^t}(X_j^t), & k=k_{new}\end{cases} \tag{17}$$

3 实 验

3.1 数据集和参数设置

由于媒体微博的发布方具有一定的权威性和代表性,相较于个人微博更加正式、可信度更高,更利于验证模型效果,因此本文选择媒体发布的财经类微博文本为实验数据.数据集包括 2012 年 9 月~2018 年 8 月财经网的微博文本,共计 92 747 篇文档.在利用主题模型提取经济要素词之前,需要进行数据预处理.

- 首先采用百度词法分析(http://ai.baidu.com/tech/nlp_basic/lexical)对微博文档实现分词以及词性标注,并将文档中出现次数低于 10 或高于 7 000 的词语删除;然后,由于本文是挖掘经济要素词,因此仅保留了词性为名词(含普通名词、动名词和专有名词)的词语;最后,经预处理后微博文本数据集的详细特征描述见表 1.其中,不同的词语总数量为 267 722 个,不同的名词总数量为 95 213 个(含不同的普通名词 47 185 个、不同的动名词 12 574 个和不同的专有名词 35 454 个);
- 根据刘涛雄等人[2]给出了 85 个代表性词语,我们通过词语间点互信息(point mutual information,简称 PMI)扩充该代表性词语集,构成经济领域类别的代表性词语集合,扩充后的代表性词语集共有 1 447 个词语.

Table 1 Feature description of dataset

表 1 数据集特征描述

微博数量	平均每篇微博字数	平均每天微博篇数	词语总数量	普通名词数量	动名词数量	专有名词数量
92 747	88	42.66	8 167 165	1 159 093	352 961	173 635

本文将 HDP 模型和改进的 HDP 模型中的超参数γ,α和β分别设置为 0.01,1.50 和 0.50.利用词语之间的语义相似性可计算待分配词语与当前主题中词语的语义相似性,可用于分析待分配词语与当前主题的语义相关度,通过词语与主题的语义相关度改进主题-词语的分配过程.过高的语义相似性阈值将导致主题中的词语大部分是语义相近词语,使得反映主题的词语不够丰富;过低的语义相似性阈值将导致主题中的词语太宽泛,使得经济领域的主题内涵不明确.经过调试,本文将词语之间的语义相似性阈值ξ设置为 0.3.

词语对主题的贡献度反映了词语对主题的代表性,需要考虑词语在所有主题中的出现情况,本文用参数η表示词语对主题的代表性阈值;如果词语在大部分主题中均出现,说明该词语对主题的代表性不强,即词语的逆主题频率越低,词语的主题代表性越弱;结合主题-词语后验概率分布 *phi*,把在 20%以上的主题中出现的词语定义为通用词语,设置词语在主题中的出现频率阈值为 0.2,结合主题数目计算该类词语的逆主题频率值约为 0.6,将其设置为参数η的取值;因此,本文将词语的逆主题频率低于 0.6 的词语定义为通用词语.参数σ表示通用词语对主题的贡献度,一般通用词语比领域主题专用词语在主题中的概率值高,为了凸显领域主题专用词语在主题中的概率值,本文将σ值设置为当前主题中词语的概率均值.

PSP_HDP 主题模型通过文档的领域隶属度、词语与主题的语义相关度和词语对主题的贡献度,改进文档-主题分配过程和主题-词语分配过程.为了对比不同因素对抽取经济要素词和构建非结构化指标体系的影响,本

文将考虑了文档领域隶属度的HDP模型记为P_HDP,将在P_HDP模型基础上再考虑词语与主题语义相关度的HDP模型记为PS_HDP,将在PS_HDP模型基础上再考虑词语对主题贡献度的HDP模型记为PSP_HDP.

3.2 实验结果

本文将从两个方面比较模型的主题挖掘效果:首先,按照主题模型常用的评价标准,包括主题多样性(KL距离)、内容困惑度和模型复杂度3个评价指标,比较不同主题模型的效果;然后,从模型抽取经济要素词和构建非结构化经济指标体系的效果评价主题模型.

3.2.1 主题模型的评价

对于主题中未出现的词语,标准的KL距离和内容困惑度均采用给定的默认值参与计算.然而,在HDP主题模型中考虑了文档的领域隶属度、词语与主题的语义相关度和词语对主题的贡献度之后,主题间的重复词语会变得越来越少,因此采用系统设定的默认值计算主题之间的KL距离和主题内容困惑度时,词语在主题中的概率取值大部分来源于默认值,导致两个主题的概率分布具有较大相似性,进而使得主题之间的KL距离越来越小、主题内容困惑度越来越大,也就是说,KL距离和内容困惑度已经无法真实地评估模型的优劣了;换句话说,对于主题中未出现的词语,系统分配的默认值无法真实地描述词语在主题中的概率分布,导致与经济领域主题无关的词语被分配较高的概率,混淆了经济要素词对经济领域主题的辨识度和区分度,无法真实地反映经济领域主题之间的差异性.因此,在计算以上2个评价指标时,本文仅考虑在主题中出现的词语.

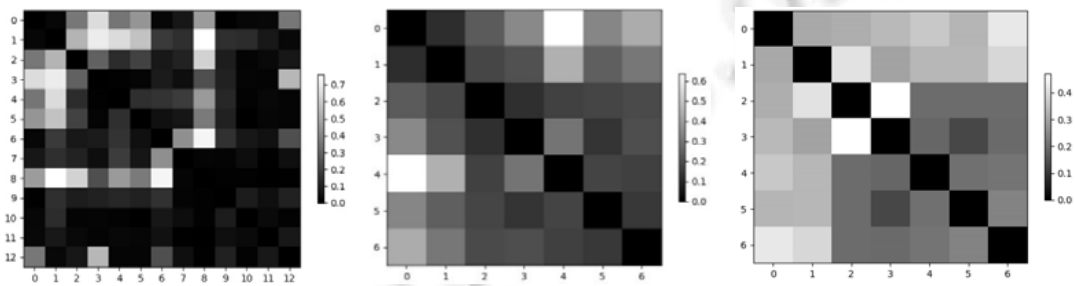
(1) 主题多样性

通过主题之间的KL距离(kullback-leibler divergence)评价主题模型的主题多样性,当KL距离为0时,表示两个主题是相同的;当KL距离为1时,表示两个主题是完全不相同的.由于实验中仅考虑主题中出现的词语,原始的KL距离计算公式可改写成公式(18):

$$KL(W_{k_1},W_{k_2})=\sum_{x_{ji}\in W_{k_1}\cap W_{k_2}} phi(x_{ji}\,|\,W_{k_1})\log\frac{phi(x_{ji}\,|\,W_{k_1})}{phi(x_{ji}\,|\,W_{k_2})} \tag{18}$$

其中,W_{k_1}和W_{k_2}分别表示主题k_1和k_2的词语集合,phi表示主题词语概率分布.

由于PSP_HDP模型生成的主题中已经剔除了大部分的重复词语,使得任意两个主题中词语集合的交集均接近空集,继而导致任意两个主题之间的KL距离都接近0,此时的KL距离已无法真实反映主题之间的区别.因此,只针对HDP,P_HDP和PS_HDP模型对比主题之间的KL距离,结果如图2所示.



(a) HDP 主题模型 (b) P_HDP 主题模型 (c) PS_HDP 主题模型

Fig.2 KL divergence of topic models

图2 主题模型的KL距离对比图

图2的横、纵坐标均为模型生成的主题数(主题编号),主题之间的KL距离用灰度的深浅表示,灰度越深,表示KL距离越近、主题之间的差异越小.对比发现:在生成的主题数方面,P_HDP和PS_HDP模型相较于HDP模型更符合已有领域分类标准要求,说明考虑了文档的领域隶属度对文档集的主题生成有指导作用;在主题的差异性方面,PS_HDP模型的KL距离分布比前两者更均匀和明确,说明考虑了文档的领域隶属度和词语与主题的语义相关度,可以提高主题词语的差别性以及主题之间的差异性.

(2) 内容困惑度

内容困惑度可以度量主题模型的效果,内容困惑度越低,表示主题模型的效果越好.计算词语 x_{ji} 在文档中的概率需要结合文档-主题概率分布和主题-词语概率分布,由于实验中只考虑主题中出现的词语,所以修改内容困惑度计算公式,如公式(19)所示:

$$perplexity(X)=\exp\left(-\frac{1}{|J|}\sum_{j\in J}\sum_{x_{ji}\in X_j,x_{ji}\in W_k}phi(x_{ji},W_k)theta(W_k,j)\right) \tag{19}$$

其中,J 表示所有餐厅的集合,X_j 表示第 j 个餐厅中的顾客集合,W_k 表示主题 k 的词语集合,*phi* 表示主题-词语概率分布,*theta* 表示文档-主题概率分布.

HDP,P_HDP,PS_HDP 和 PSP_HDP 主题模型在迭代过程中的内容困惑度变化情况如图 3 所示.

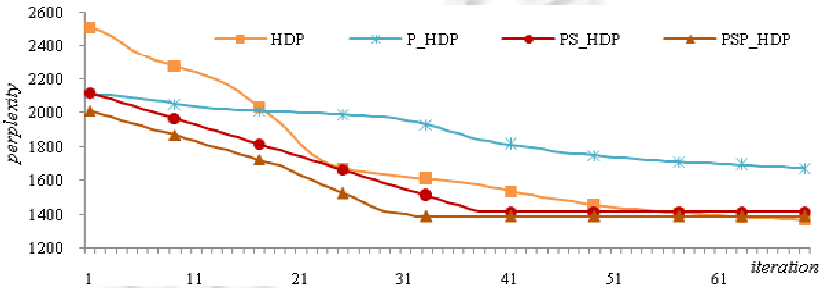


Fig.3 Perplexity of topic models

图 3 主题模型的内容困惑度对比图

从图 3 可看出,P_HDP 模型的内容困惑度收敛值最大,HDP 模型的内容困惑度在迭代 60 次附近时达到了收敛值.虽然 PS_HDP,PSP_HDP 和 HDP 模型的内容困惑度收敛值相差不大,但是 PS_HDP 模型在迭代 38 次时可达到与 HDP 内容困惑度相近的收敛值,而 PSP_HDP 模型则在迭代 30 次时就达到了相似的效果,不仅提高了主题模型的分析效率,而且主题词和主题的提取效果比 HDP 更好,这点在分析非结构化经济指标体系构建和经济要素词抽取的效果时再具体阐述.

从图 3 还可看出,仅考虑文档的领域隶属度,并不能有效地改善模型的内容困惑度,反而减缓了模型内容困惑度的下降速度.分析原因,主要是由于不同领域的文档中存在部分共有词语,这些共有词语的语义信息不明确,降低了主题之间的辨识度,提高了模型的内容困惑度.另外,增加文档的领域隶属度和词语与主题的语义相关度,不仅明确了文档的领域信息,而且还明确了词语的语义信息,对提高主题之间的区分度有促进作用,起到了降低模型内容困惑度和加快迭代收敛速度的效果;继续考虑词语对主题的贡献度,则可进一步识别不同主题之间词语的差别性,提高主题辨识度、降低模型内容困惑度,并进一步加快迭代收敛速度.

(3) 模型复杂度

模型复杂度的计算方法参考 Kim[17]和 Ahmed[20]等人提出的方法,定义为模型主题个数与所有文档中不同主题个数之和,计算公式如公式(20)所示:

$$complexity=|K|+\sum_{j\in J}\sum_{k\in K}I\left(\left(\sum_{t\in T_j}I(k_j^t=k)\right)>0\right) \tag{20}$$

其中,J 表示所有餐厅的集合;K 表示所有餐桌已供应的菜肴集合,即模型主题集合;T_j 表示第 j 个餐厅中已分配顾客的餐桌集合;k_j^t 表示第 j 个餐厅 t 餐桌分配菜肴的索引值.

在评价主题模型效果时,当主题模型的内容困惑度差别不大时,如果模型复杂度越低,则主题模型就越有效.4 个主题模型的模型复杂度如图 4 所示.

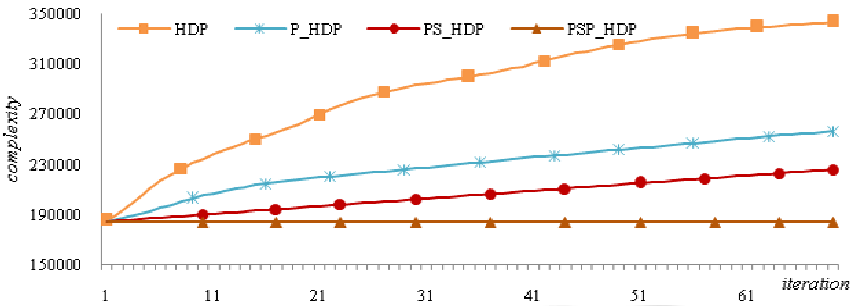


Fig.4 Complexity of topic models

图 4 主题模型的模型复杂度对比图

从图 4 可以看出,随着迭代次数的增加,4 个模型的模型复杂度都在变大,并且 HDP 模型的模型复杂度的增大速度明显快于 P_HDP,PS_HDP 和 PSP_HDP 模型.PSP_HDP 模型的模型复杂度最低,且变化范围不大,基本趋于平稳状态.在分析模型的内容困惑度时,PS_HDP,PSP_HDP 和 HDP 模型的内容困惑度差别不大;结合模型复杂度发现,P_HDP,PS_HDP 和 PSP_HDP 模型的模型复杂度明显低于 HDP 模型,说明考虑文档的领域隶属度、词语与主题的语义相关度和词语对主题的贡献度,可以改善主题模型的整体性能.

3.2.2 抽取经济要素词和构建非结构化经济指标体系的效果评价

(1) 定性评价

衡量主题模型效果的另一种方法是通过对比主题模型生成的主题和主题词,考察财经文本中非结构化经济指标体系构建和经济要素词抽取的效果.刘涛雄等人[2]以 2006 年~2014 年的百度指数网站的百度搜索词语为参照,根据经济领域类别人工筛选了 85 个代表性词语,结果见表 2.其中,共有 12 个词语是无法通过主题模型提取到的,包括带实下划线的 8 个没有出现在微博文本数据集中的词语、带虚下划线的 3 个在微博文本数据集中出现少于 10 次的词语(将会在数据预处理时被删除)、带波浪下划线的 1 个没有被正确分词的词语.主要原因是,经济运行发展中反映经济指标的词语发生了一些变化.本文将人工划分的经济领域类别及其筛选的代表性词语[2],以及主题模型生成的主题和主题词,分别对应于非结构化经济指标和经济要素词.

Table 2 Manual partition of the economic sector categories and representative words

表 2 人工划分的经济领域类别及筛选的代表性词语

经济领域类别	代表性词语
投资	华夏基金;证券;国债;创业;利率;邮票;融资;投资;期权;房地产;收藏品;保险;贵金属;贷款;外汇交易;银行;股市;基金;炒股;理财产品
进出口	运费;进口葡萄酒;出口;出口信贷;进口;中国进出口贸易网;进出口;进出口代理;出口许可证;海关编码查询;成品油;来料加工;汽车零部件;PMI;出口退税;海运费;铁矿石;关税查询;石油;原油;外汇;进出口银行;期货;钢材;进口化妆品;集装箱;大豆
消费	户外;音响;首饰;餐巾纸;消费;享受;装饰;时装;电器;演唱会;促销;维修;书;健身房;茶;玩具;酒店;团购;自助;减肥;出国;美容
政府财政	政府采购;工程招标;差旅费;工资;税收;招投标;公务员工资;增值税;基础设施建设
人口就业	兼职;劳动合同法;招聘;小时工;智联招聘网;失业保险;个人所得税

采用 HDP 主题模型共生成了 13 个主题,主题之间的领域信息不明确,并且主题之间存在大量的重复高频词语,具体结果见附录 1.P_HDP 主题模型生成的主题和主题词在投资、进出口和政府财政这 3 个方面具有一定的代表意义,其他领域的主题词语较为杂乱,语义信息不明确.PS_HDP 主题模型生成包括投资、进出口、消费、政府财政和人口就业这 5 个方面的主题,保留每个主题的前 30 个词语,结果见表 3.其中,带下划线的词语表示人工判断明显不属于该非结构化经济指标下的经济要素词,共 76 个,在投资、进出口、消费、政府财政和人口就业这 5 个方面的主题中,分别有 10,17,15,17 和 17 个.本文共选用具备较强财经知识的 2 位博士研究生和 1 位教师作为判断者;当出现不一致的判断结果时,则由 3 人讨论确认判断结果.

Table 3 Topics and topic words that represent economic indicators based on PS_HDP

表 3 PS_HDP 主题模型生成的非结构化经济指标和经济要素词

非结构化经济指标	经济要素词
投资	银行;投资;房地产;市场;经济;贷款;公司;企业;政策;股市;金融;政府;资金;融资;利率;风险;住房;项目;房价;钱;人民币;业务;基金;城市;个人;沪指;交易;价格;社会;行业
进出口	企业;价格;进口;市场;经济;人员;公司;文件;汽油;账户;产品;政府;油价;情况;数据;记者;信息;社会;事件;人民币;出口;政策;制造业;时间;措施;消息;手机;警方;关税;钢铁
消费	消费;男子;酒店;警方;人员;价格;社会;手机;公司;孩子;学生;游客;记者;市场;女子;产品;企业;事件;政经;时间;书;微博;服务;钱;消费者;政府;情况;信息;媒体;经济
政府财政	工资;企业;收入;税收;政府;公司;员工;人员;公务员;钱;增值税;疫苗;项目;经济;社会;个人;改革;土地;价格;地方;单位;情况;政策;市场;财政;职工;住房;采购;官员;标准
人口就业	企业;人员;养老保险;个人;改革;收入;养老金;招聘;制度;社会;社保;缴费;路线;公司;政策;违约;单位;员工;政府;职工;个税;兼职;市场;专家;工资;经济;劳动者;养老;情况;退休

对比表 3 和附录 1 的结果,PS_HDP 主题模型生成的经济要素词对主题的解释能力更强、非结构化经济指标的内涵相对更清晰.对比表 3 和表 2,在非结构化经济指标的构建方面,PS_HDP 主题模型生成的非结构化经济指标和人工划分的结果比较一致.在经济要素词抽取方面,随着时间变化,非结构化经济指标对应的经济要素词会有一定的变化,表 3 中的词语包括表 2 中的部分词语;但是,表 3 对应的经济要素词中存在部分高频的、与经济指标不太相关的词语.

PSP_HDP 模型是在 PS_HDP 模型的基础上,增加了对主题具有较高代表性的中低频词语在主题中的影响作用,抽取的经济要素词以及对应的非结构化经济指标见表 4.表中只取对应指标的前 30 个主题词,其中带下划线的词语表示人工判断明显不属于该非结构化经济指标下的经济要素词,共 27 个,在投资、进出口、消费、政府财政和人口就业等 5 个方面的主题中分别有 3,3,3,11 和 7 个.

对比表 4 和表 2,在非结构化经济指标的构建方面,PSP_HDP 主题模型生成的非结构化经济指标和人工划分的结果比较一致.对比表 4 和表 3 中的经济要素词,表 4 中的经济要素词更加具有代表性;对比表 4 和表 2 中的经济要素词,表 4 不仅包含了人工筛选的代表性词语,还包括一些随着经济和社会发展而新产生的经济要素词.如在投资方面,目前主要侧重于国债、ICO、MLF、艺术品、M2、快捷支付、国债期货、购汇、创业板等词语;在进出口方面,目前更具体地体现在汽油、成品油、芯片、水稻、高粱、棉花、钢企、矿机等;在消费方面,更实时地反映了当前人民群众的关注领域,包括餐馆、大众点评、相亲、舞、酒席、面膜、奥数等;在政府财政方面,出现了一些与政府政策实时相关的词语,包括消费税、营改增、竞标、营业税、招待费、投标、税种等;在人口就业方面,出现了一些与目前就业政策相关的、且更具体的词语,包括养老保险、违约、五险一金、工伤、养老院、参保、年金等.

Table 4 Topics and topic words that represent economic indicators based on PSP_HDP

表 4 PSP_HDP 主题模型生成的非结构化经济指标和经济要素词

非结构化经济指标	经济要素词
投资	融资;股市;房价;基金;利率;存款利率;收益率;外汇;跨境;国债;ICO;贷款基准利率;基点;刚需;MLF;军工;艺术品;知乎;M2;同业存单;净资产;快捷支付;存单;房贷利率;降息;古董;国债期货;钱庄;购汇;创业板
进出口	汽油;成品油;PMI;出口;原油;大豆;柴油;转基因;芯片;汽柴油;进出口;危险品;期货;油箱;转基因作物;调价;水稻;反制;钢;高粱;棉花;运费;钢企;菜农;燃煤;口岸;用电;矿机;集装箱;玉米
消费	酒店;游客;餐馆;大众点评;相亲;格力;舞;减肥;玩具;模特;时尚;演唱会;酒席;手镯;神医;果汁;音响;面膜;设计师;奥数;茶;青蛙;整形;健身房;团购;茶叶;首饰;KTV;整容;健身
政府财政	工资;税收;增值税;采购;消费税;效价;营改增;政府采购;招投标;竞标;营业税;经济舱;药厂;中标;招待费;使用率;投标;头等舱;小金库;差旅费;创收;住宿费;税种;医疗器械;副职;大老虎;确权;局级;干部;媒体
人口就业	养老保险;违约;招聘;劳动者;缴费;岗;个税起征点;五险一金;工伤;税法;费率;养老院;招聘会;参保;BOSS 直聘;应聘者;园林;求职者;年金;居家;教职;义工;启事;运管;基本养老保险;失业保险;医疗保险;退休;服务;项目

(2) 定量评价

为了定量、客观地分析主题模型抽取经济要素词和构建非结构化经济指标体系的效果,将从表 2 人工筛选

的 85 个代表性词语集中去除 12 个未在微博文本数据集中出现的词语后的代表性词语集记为 N,即$|N|$=73;将主题模型生成的经济要素词集记为 M,将从 M 中人工剔除明显不属于非结构化经济指标的主题词之后的经济要素词集记为 $\bar{M}$;将 M 中包含 N 中词语的数量与 N 中词语数量的占比定义为提取代表性词语的召回率 R,如公式(21)所示,定量评估模型提取经济领域类别中代表性词语的效果.

$$R=\frac{|M\cap N|}{|N|} \tag{21}$$

将 $\bar{M}$ 中的词语数量与 M 中词语数量的占比定义为抽取经济要素词的准确率 P,如公式(22)所示,定量评估模型抽取经济指标中经济要素词的效果.

$$P=\frac{|\bar{M}|}{|M|} \tag{22}$$

如果主题模型抽取 top_n 个主题词作为经济要素词,则 HDP,P_HDP,PS_HDP 和 PSP_HDP 主题模型提取代表性词语的召回率 R、抽取经济要素词的准确率 P 的结果分别如图 5 和图 6 所示.

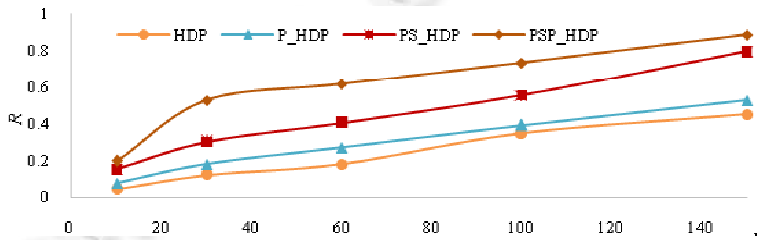


Fig.5 R values of representative words based on top_n's topic words in topic models

图 5 主题模型抽取的 top_n 个主题词中代表性词语的召回率

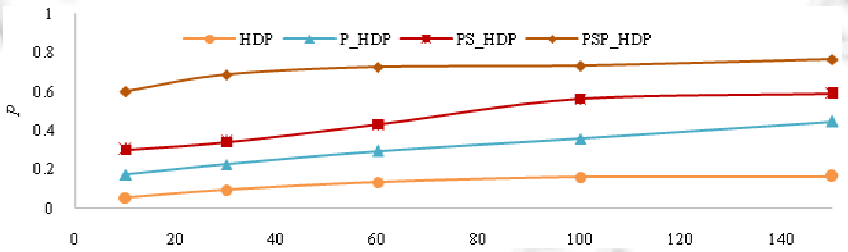


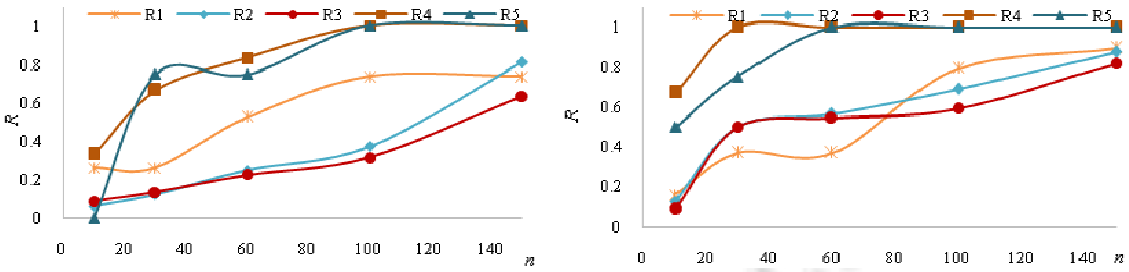
Fig.6 P values of economic factors based on top_n's topic words in topic models

图 6 主题模型抽取的 top_n 个主题词中经济要素词的准确率

从图 5 和图 6 可知,P_HDP,PS_HDP 和 PSP_HDP 主题模型的 R 和 P 值均高于 HDP 主题模型,说明改造后的 HDP 主题模型不仅能够较好地提取到经济领域类别中的代表性词语,而且也能抽取到非结构化经济指标中更多的经济要素词.

如果将主题模型中非结构化经济指标对应到人工设定的经济领域类别,分析每个非结构化经济指标的经济要素词中代表性词语的召回情况,结果如图 7 所示.由于 4 种模型中只有 PS_HDP 和 PSP_HDP 主题模型生成的非结构化经济指标与人工设定的经济领域类别较为一致,因此图 7 只展示这 2 种模型的代表性词语在不同经济领域类别中的召回情况,图 7(a)、图 7(b)分别对应 PS_HDP 和 PSP_HDP 主题模型的结果.其中,R1~R5 分别对应投资、进出口、消费、政府财政和人口就业领域的召回率情况.

从图 7 可知,总体上,政府财政和人口就业领域的召回率明显高于其他 3 个领域,主要原因是政府财政和人口就业领域中人工筛选的代表性词语较少,导致公式(21)中的分母较小,R 值整体偏大;进出口和消费领域中代表性词语召回率整体偏低,主要原因是经济运行发展中进出口和消费领域的关注点变化较快;PSP_HDP 主题模型的代表性词语召回率,在各个领域大部分都优于 PS_HDP 主题模型的结果.



(a) PS_HDP 主题模型 (b) PSP_HDP 主题模型

Fig.7 *R* values of representative words in different economic indicators based on PS_HDP and PSP_HDP

图 7 PS_HDP 和 PSP_HDP 主题模型不同领域代表性词语的召回率

通过上面的定性和定量分析可以看出,PSP_HDP 主题模型在考虑了文档的领域隶属度、词语与主题的语义相关度和词语对主题的贡献度之后,能够抽取更加全面、实时和具有代表性的经济要素词,体现了非结构化经济指标的内涵变化,提高了经济主题的区分度和辨识度.

另外,PS_HDP 主题模型和 PSP_HDP 主题模型还生成了一个关于社会事件方面的主题,其对应的前 30 个主题词见表 5.

Table 5 Additional topics and topic words based on PSP_HDP

表 5 PSP_HDP 主题模型生成的附加主题及对应的主题词

主题	主题词
社会事件	医院;地震;爆炸;违法;航班;互联网;犯罪;患者;雾霾;污染;客机;环境;死亡;疫苗;暴力;高校;救援;执法;犯罪嫌疑人;网瘾;硫氰酸钠;遇害案;中华鲟;马航;病毒;汛期;脑瘫;冻肉;侦察机;可疑物

分析表 5 发现,除了生成对应于人工划分的投资、进出口、消费、政府财政、人口就业这 5 个经济领域类别的主题之外,PS_HDP 和 PSP_HDP 主题模型还生成了一个“社会事件”主题,该主题与经济存在着密切关联.因为一些与人民生命财产相关的事件,例如自然灾害、事故灾害、公共卫生事件和社会安全事件,在给社会造成严重危害的同时,也会直接或间接地影响经济发展.相关研究也证明了这类社会事件与经济之间的相互影响关系[32].

实验结果说明:通过 PS_HDP 和 PSP_HDP 主题模型不仅能挖掘与经济指标相关的确定性因素,还能挖掘与经济相关的不确定性因素.

4 总结与展望

不仅财经文本具有领域类别属性,而且与经济指标相关的词语(即经济要素词)也具有领域主题差异性,因此,通过原始的主题模型挖掘出来的主题和主题词无法准确地反映这些领域或领域主题特性.人工构建的非结构化经济指标体系中蕴含了经济指标的领域类别划分标准和对应的代表性词语,且同一经济领域的词语具有丰富的语义关联关系,利用这些信息可指导 HDP 主题模型挖掘经济主题及其对应的经济要素词.

本文提出的 PSP_HDP 主题模型考虑了经济领域划分标准和对应的代表性词语,通过文档的相似性,计算文档的领域隶属度,改进 CRF 的餐桌-菜肴分配过程,指导文档-主题分配;考察了词语对经济主题的描述能力,利用词向量对词语语义信息的描述,计算词语与主题的语义相关度,改进 CRF 的顾客-餐桌分配过程,以便将语义相近的词语尽量分配到相同或相近的主题,提高经济主题的区分度;考察了词语对经济主题的代表性,根据词语对主题的贡献度,进一步改进顾客在不同菜肴风格群体中的代表性,以便能够抽取到有效的、中低频的领域主题专用词语,提高经济主题的辨识度.通过改进采样方法,实现非结构化经济指标体系的构建和经济要素词的抽取.

从实验结果中发现,综合主题多样性、内容困惑度和模型复杂度,PSP_HDP 主题模型整体性能明显优于

HDP 主题模型.在非结构化经济指标体系构建和经济要素词抽取方面,PSP_HDP 主题模型不仅可以自动地挖掘出人工抽取的经济指标和经济要素词,而且还能挖掘出随着经济社会发展而产生的新颖的经济要素词,同时还能挖掘出潜在的、与经济指标相关的其他主题和主题词,验证了模型的有效性.

下一步工作将继续分析财经文本中经济要素词的分布特点,改进主题模型的挖掘效果,提取更丰富的、具有代表性的、实时的经济要素词.对于新的经济主题,需要进一步细化经济要素词的分类.本文研究主要侧重于构建经济主题的两层结构,后续研究可构建经济主题的多层结构.

References:

[1] Einav L, Levin J. Economics in the age of big data. Science, 2014,346(6210):715−719.

[2] Liu TX, Xu XF. Can Internet search behavior help to forecast the macro economy? Economic Research, 2015,12:68−83 (in Chinese with English abstract).

[3] Moat HS, Curme C, Stanley HE, Preis T. Anticipating stock market movement with Google and Wikipedia. In: Proc. of the Int'l Conf. on NATO Science for Peace and Security Series C: Environmental Security. Springer-Verlag, 2013. 47−59.

[4] Luo P, Chen YG, Xu CH. Baidu search, risk perception and risk prediction—A perspective of behavioral finance. Finance Forum, 2018,1:39−51 (in Chinese with English abstract).

[5] Yakovleva K. Text mining-based economic activity estimation. Russian Journal of Money and Finance, 2018,77(4):26−41.

[6] Siegel M. Text mining in economics. Semantic Applications, 2018, 63−73.

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

[8] Chen Z, Mukherjee A, Liu B, Hsu MC, Castellanos M, Ghosh R. Leveraging multi-domain prior knowledge in topic models. In: Proc. of the 23rd Int'l Joint Conf. on Artificial Intelligence. 2013. 2071−2077.

[9] Chen Z, Liu B. Mining topics in documents: Standing on the shoulders of big data. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2014. 1116−1125.

[10] Chen ZY, Mukherjee A, Liu B. Aspect extraction with automated prior knowledge learning. In: Proc. of the Association for Computational Linguistics. Association for Computational Linguistics, 2014. 347−358.

[11] Li C, Wang H, Zhang Z, Sun AX, Ma ZY. Topic modeling for short texts with auxiliary word embeddings. In: Proc. of Int'l ACM SIGIR Conf. on Research & Development in Information Retrieval. ACM, 2016. 165−174.

[12] Liu Y, Liu Z, Chua TS, Sun M. Topical word embeddings. In: Proc. of the 29th AAAI Conf. on Artificial Intelligence. AAAI, 2015. 2418−2424.

[13] Das A, Kannan A. Discovering topical aspects in Microblogs. In: Proc. of the 25th Conf. on Computational Linguistics. Association for Computational Linguistics, 2014. 860−871.

[14] Zhang CY, Shun JL, Ding YQ. Topic mining for microblog based on MB_LDA model. Journal of Computer Research and Development, 2011,48(10):1795−1802 (in Chinese with English abstract).

[15] Pang XW, Wan BS, Wang P. Micro-Blog's text classification based on MRT_LDA. Computer Science, 2017,44(8):236−241 (in Chinese with English abstract).

[16] Whye TY, Jordan MI, Beal MJ, Blei DM. Hierarchical dirichlet processes. Journal of the American Statistical Association, 2006, 101:1566−1581.

[17] Kim D, Oh A. Accounting for data dependencies within a hierarchical dirichlet process mixture model. In: Proc. of the ACM Int'l Conf. on Information and Knowledge Management. ACM, 2011. 873−878.

[18] Li C, Rana S, Phung D, Phung D, Venkatesh S. Data clustering using side information dependent Chinese restaurant processes. Knowledge and Information Systems, 2016,47(2):463−488.

[19] Blei DM, Frazier PI. Distance dependent Chinese restaurant processes. Journal of Machine Learning Research, 2012,12(1): 2461−2488.

[20] Ahmed A, Xing EP. Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. In: Proc. of the 26th Conf. on Uncertainty in Artificial Intelligence. Association for Uncertainty in Artificial Intelligence, 2010. 20−29.

[21] Ma T, Qu D, Ma R, Feng W, Li K. Online topic evolution modeling based on hierarchical dirichlet process. In: Proc. of the IEEE 1st Int'l Conf. on Data Science in Cyberspace. IEEE, 2016. 400−405.

[22] Zhang J, Song Y, Zhang C, Liu SX. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2010. 1079−1088.

[23] Wang P, Zhang P, Zhou C, Li Z, Yang H. Hierarchical evolving dirichlet processes for modeling nonlinear evolutionary traces in temporal data. Data Mining & Knowledge Discovery, 2017,31(1):32−64.

[24] Liu SP, Yin J, Ouyang J, Huang Y, Yang XY. Topic mining from microblogs based on MB_HDP model. Chinese Journal of Computers, 2015,38(7):1408−1419 (in Chinese with English abstract).

[25] Qian J, Gong Y, Zhang Q, Huang XJ. Hierarchical dirichlet processes with social influence. In: Proc. of the National CCF Conf. on Natural Language Processing and Chinese Computing. Springer-Verlag, 2017. 490−502.

[26] Yang M, Hsu WH. HDPauthor: A new hybrid author-topic model using latent dirichlet allocation and hierarchical dirichlet processes. In: Proc. of the 25th Int'l Conf. Companion on World Wide Web. ACM, 2016. 619−624.

[27] Li W, Yin J, Chen HC. Supervised topic modeling using hierarchical dirichlet process-based inverse regression: Experiments on e-commerce applications. IEEE Trans. on Knowledge and Data Engineering, 2018,30(6):1192−1205.

[28] Yang Q. Stock returns and real growth: A Bayesian nonparametric approach. Social Science Electronic Publishing, 2018,3:1−38.

[29] Blei DM, Griffiths TL, Jordan MI, Tenenbaum JB. Hierarchical topic models and the nested Chinese restaurant process. In: Proc. of the Int'l Conf. on Neural Information Processing Systems. MIT, 2003. 17−24.

[30] Chen J, Zhu J, Lu J, Liu SX. Scalable inference for nested Chinese restaurant process topic models. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2017. 1−9.

[31] Zhou JY, Wang FY, Zeng DJ. Hierarchical Dirichlet process and their application: A survey. Acta Automatica Sinica, 2011,37(4): 389−407 (in Chinese with English abstract).

[32] Farhi E, Gabaix X. Editor's choice: Rare disasters and exchange rates. Quarterly Journal of Economics, 2016,131(1):1−52.

附中文参考文献:

[2] 刘涛雄,徐晓飞.互联网搜索行为能帮助我们预测宏观经济吗?经济研究,2015,12:68−83.

[4] 罗鹏,陈义国,许传华.百度搜索、风险感知与金融风险预测——基于行为金融学的视角.金融论坛,2018,1:39−51.

[14] 张晨逸,孙建伶,丁轶群.基于 MB-LDA 模型的微博主题挖掘.计算机研究与发展,2011,48(10):1795−1802.

[15] 庞雄文,万本帅,王盼.基于 MRT-LDA 模型的微博文本分类.计算机科学,2017,44(8):236−241.

[24] 刘少鹏,印鉴,欧阳佳,黄云,杨晓颖.基于 MB-HDP 模型的微博主题挖掘.计算机学报,2015,38(7):1408−1419.

[31] 周建英,王飞跃,曾大军.分层 Dirichlet 过程及其应用综述.自动化学报,2011,37(4):389−407.

附录 1 HDP 主题模型的生成结果

主题	主题词
Topic0	公司;企业;市场;信息;用户;手机;平台;人员;网络;社会;价格;服务;行为;业务;投资;产品;媒体;微博;员工;时间;情况;银行;个人;数据;交易;政府;股东;记者;项目;消息;机构;学生;警方;管理;监管;经济;钱;董事长;公告;人士;新闻;事件;城市;政策;金融;行业;互联网;图片;内容;资金;网站;孩子;广告;官方;发展;方面;股票;微信;方式;收入;风险;政经;人民币;影响;系统;投资者;制度;原因;法律;技术;iPhone;阿里;男子;改革;消费者;上市公司;措施;活动;专家;销售;标准;A 股;结果;违法;报告;金额;IPO;资本;法院;学校;股价;账户;计划;股份;司机;负责人;品牌;股权;案件;地方
Topic1	男子;人员;警方;孩子;学生;社会;事件;企业;政经;公司;记者;信息;情况;行为;女子;微博;网络;政府;市场;媒体;学校;老师;个人;司机;时间;家长;图片;手机;政策;官员;新闻;钱;经济;老人;民警;儿子;官方;消息;教育;服务;女生;原因;乘客;事故;用户;女儿;平台;父母;干部;教师;医院;项目;改革;领导;专家;价格;女孩;女性;法院;关系;收入;案件;内容;活动;管理;产品;网站;措施;违法;结果;单位;母亲;妻子;城管;负责人;机构;法律;家庭;群众;游客;过程;父亲;飞机;方面;方式;局长;生活;数据;组织;犯罪;地方;制度;家属;儿童;职务;电梯;员工;书记;城市;影响
Topic2	城市;企业;经济;政府;市场;收入;政策;房价;价格;改革;土地;公司;住房;社会;投资;人口;房地产;地方;项目;个人;数据;制度;发展;工资;情况;政经;人员;居民;建设;资金;财政;银行;钱;家庭;金融;服务;官员;成本;管理;GDP;人民币;图片;中央;房;消费;住宅;报告;时间;农村;信息;方面;风险;比例;记者;行为;措施;增速;媒体;楼市;产品;农民;影响;关系;人士;专家;方式;利益;学生;中方;政治;员工;贷款;平台;新闻;行业;公务员;交易;商品;规模;标准;微博;就业;机构;产业;环境;消息;生活;涨幅;资产;调控;手机;单位;原因;地区;资本;计划;领域;业务;监管;国企

附录 1 HDP 主题模型的生成结果(续 1)

主题	主题词
Topic3	人员;警方;飞机;男子;事故;公司;事件;社会;企业;信息;乘客;政经;记者;情况;航班;市场;消息;微博;爆炸;客机;时间;媒体;手机;政府;司机;学生;原因;新闻;用户;女子;官员;行为;网络;钱;经济;图片;游客;银行;项目;官方;孩子;员工;民警;个人;家属;人士;平台;车辆;投资;活动;产品;数据;服务;案件;价格;系统;技术;警察;方面;专家;旅客;交警;老人;管理;组织;资金;交易;过程;金融;法院;结果;影响;领导;医院;网站;全部;业务;城市;机构;改革;关系;救援;政策;微信;群众;人民币;地方;收入;发展;失联;方式;机场;机;措施;行业;女儿;报告;电话;儿子;法律
Topic4	住房;政策;经济;企业;市场;政府;贷款;城市;银行;居民;家庭;房地产;房价;个人;金融;改革;价格;公司;投资;人民币;风险;社会;比例;人员;资金;收入;情况;地方;项目;楼市;数据;制度;户籍;监管;管理;措施;公积金;产品;业务;商品;人士;房;土地;首付;发展;调控;影响;图片;机构;交易;钱;股市;购房;时间;财政;资产;政经;信息;媒体;行业;需求;专家;服务;记者;融资;人口;货币;方式;行为;官员;利率;规模;消息;中央;条件;社保;标准;资本;增速;商品房;本市;建设;原因;工资;调整;事件;GDP;方案;债务;方面;警方;微博;新闻;基金;住宅;意见;单位;报告;房产;存款
Topic5	市场;企业;公司;人员;社会;政府;经济;政经;警方;记者;学生;微博;媒体;情况;事件;信息;男子;孩子;改革;手机;行为;时间;新闻;价格;消息;网络;产品;政策;钱;官员;收入;银行;城市;管理;个人;发展;项目;官方;行业;数据;图片;服务;领导;电影;机构;投资;地震;用户;活动;制度;人士;人民币;员工;原因;事故;学校;业务;平台;金融;法律;影响;资金;地方;方式;网站;专家;结果;方面;游客;标准;成本;法院;组织;监管;女子;生活;单位;教育;演员;措施;资本;内容;报告;交易;风险;工资;政治;互联网;医院;关系;负责人;人口;广告;会议;董事长;群众;家长;过程;干部;系统
Topic6	经济;企业;市场;政府;银行;政策;改革;公司;价格;城市;金融;利率;人民币;投资;房地产;风险;收入;地方;社会;数据;项目;房价;增速;GDP;情况;资金;贷款;发展;债务;货币;人员;股市;行业;国企;居民;钱;宏观;政经;影响;个人;措施;财政;产品;百分点;时间;资本;土地;增长;监管;人士;资产;融资;制度;记者;事件;规模;媒体;报告;图片;压力;消息;官员;员工;专家;住房;货币政策;管理;信息;原因;机构;成本;方面;污染;房贷;沪指;金融机构;方式;微博;业务;环境;A 股;目标;工资;行为;流动性;官方;建设;调整;警方;董事长;交易;服务;涨幅;地区;中央;调控;历史;存款;股东;人口
Topic7	警方;男子;医院;人员;医生;孩子;社会;患者;事件;政经;女子;家属;记者;学生;情况;老人;病例;公司;手机;微博;企业;女儿;钱;政府;儿童;民警;女孩;专家;事故;官方;儿子;消息;妻子;母亲;市场;父亲;媒体;原因;时间;行为;司机;村民;结果;新闻;信息;父母;官员;价格;案件;法院;图片;死者;警察;手术;乘客;产品;H7N9 禽流感;死亡;学校;网络;医疗;刀;家长;女生;嫌疑人;群众;家人;丈夫;药品;药;嫌犯;经济;病人;过程;数据;影响;员工;爆炸;生命;雾霾;报告;女性;治疗;婴儿;城市;市民;银行;老师;关系;律师;全部;个人;食品;研究;人士;机构;男童;女童;方式;政策
Topic8	警方;男子;人员;社会;学生;政经;事件;司机;记者;孩子;女子;事故;微博;情况;媒体;公司;民警;政府;企业;时间;手机;钱;消息;游客;汽油;信息;新闻;价格;官方;乘客;原因;行为;交警;网络;老人;市场;官员;医院;车辆;图片;群众;家属;警察;结果;车主;活动;爆炸;儿子;经济;员工;平台;学校;领导;村民;产品;专家;油价;女儿;法院;中方;案件;干部;个人;措施;母亲;银行;汽车;政策;女孩;过程;市民;飞机;酒店;项目;城市;女生;老师;关系;地方;方面;城管;柴油;影响;医生;父母;人士;数据;方式;女性;犯罪;交通;网站;父亲;对方;家长;负责人;枪;水;单位;组织
Topic9	地震;人员;城市;政府;企业;市场;社会;公司;警方;经济;价格;男子;政策;情况;政经;房价;微博;记者;项目;收入;时间;消息;改革;信息;事件;媒体;官员;钱;地方;银行;投资;数据;司机;手机;住房;图片;新闻;行为;房地产;学生;个人;交通;事故;官方;居民;服务;乘客;孩子;管理;原因;游客;影响;资金;人士;制度;专家;行业;网络;措施;领导;产品;土地;交易;女子;车辆;地铁;地区;金融;风险;员工;建设;老人;深度;发展;监管;人口;方式;报告;单位;机构;活动;工资;人民币;标准;系统;雾霾;用户;平台;出租车;震源;案件;生活;业务;楼市;调控;过程;市民;负责人;房;结果
Topic10	学生;警方;男子;人员;社会;政经;孩子;企业;政府;事件;记者;公司;情况;微博;学校;媒体;市场;钱;女子;时间;经济;收入;官员;老人;标准;医院;新闻;教育;专家;手机;官方;原因;领导;奶粉;信息;女生;行为;消息;价格;患者;民警;医生;网络;个人;图片;司机;游客;项目;事故;儿子;城市;家长;食品;产品;政策;结果;村民;地方;改革;老师;家属;干部;活动;生活;父母;网站;女儿;影响;女性;数据;报告;员工;教师;女孩;银行;工资;发展;管理;关系;市民;案件;妻子;群众;组织;母亲;居民;过程;历史;机构;服务;父亲;方式;法律;单位;乘客;高校;投资;制度;爆炸;人士
Topic11	人员;社会;男子;政府;企业;公司;微博;警方;学生;政经;经济;市场;记者;媒体;情况;时间;决赛;女子;事件;消息;官员;项目;钱;政策;比赛;城市;奥运;个人;孩子;新闻;领导;选手;价格;官方;信息;队;行为;收入;图片;改革;历史;手机;网络;地方;原因;投资;老人;银行;足球;事故;员工;冠军;学校;活动;金牌;关系;数据;制度;董事长;工资;专家;发展;游客;影响;房价;单位;机构;产品;人士;干部;管理;网站;司机;方式;方面;报告;法院;生活;人民币;服务;居民;儿子;医院;标准;法律;民警;地震;措施;主席;负责人;房地产;住房;地区;组织;政治;行业;世界杯;公务员;裁判;父母

附录 1 HDP 主题模型的生成结果(续 2)

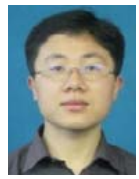
主题	主题词
Topic12	人员;警方;男子;社会;公司;政府;事件;学生;企业;市场;政经;情况;微博;记者;事故;经济;孩子;消息;媒体;官员;时间;女子;钱;信息;新闻;银行;官方;老人;手机;医院;收入;价格;政策;专家;行为;城市;原因;项目;村民;图片;个人;司机;民警;爆炸;网络;案件;地方;家属;领导;数据;游客;结果;影响;市民;改革;警察;法院;人民币;乘客;员工;学校;人士;医生;产品;局长;儿子;措施;活动;管理;投资;董事长;地震;关系;居民;女儿;服务;女孩;报告;用户;救援;机构;方面;过程;方式;工资;生活;飞机;父母;总统;资金;全部;组织;风险;制度;女生;房价;书记;群众;发展;老师

张奕韬(1984－),女,江西进贤人,博士生,主要研究领域为 Web 数据管理,数据挖掘,自然语言处理.

万常选(1962－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为 Web 数据管理,数据挖掘,情感分析,信息检索.

刘喜平(1981－),男,博士,副教授,CCF 专业会员,主要研究领域为数据库,数据挖掘,信息检索.

江腾蛟(1976－),女,博士,副教授,主要研究领域为数据挖掘,情感分析,Web 数据管理.



刘德喜(1975－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为社会媒体处理,信息检索,自然语言处理.

廖国琼(1969－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,数据挖掘,社会网络.