软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(4):1790−1818 [doi: 10.13328/j.cnki.jos.006840]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

主题方面共享的领域主题层次模型*

万常选[1,3], 张奕韬[1,2,3], 刘德喜[1,3], 刘喜平[1,3], 廖国琼[1,3], 万齐智[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](江西省高校数据与知识工程重点实验室 (江西财经大学), 江西 南昌 330013)

通信作者: 张奕韬, E-mail: 25658497@qq.com

摘　要: 层次主题模型是构建主题层次的重要工具. 现有的层次主题模型大多通过在主题模型中引入 nCRP 构造方法, 为文档主题提供树形结构的先验分布, 但无法生成具有明确领域涵义的主题层次结构, 即领域主题层次. 同时, 领域主题不仅存在层次关系, 而且不同父主题下的子主题之间还存在子领域方面共享的关联关系, 在现有主题关系研究中没有合适的模型来生成这种领域主题层次. 为了从领域文本中自动、有效地挖掘出领域主题的层次关系和关联关系, 在 4 个方面进行创新研究. 首先, 通过主题共享机制改进 nCRP 构造方法, 提出 nCRP+层次构造方法, 为主题模型中的主题提供具有分层主题方面共享的树形先验分布; 其次, 结合 nCRP+和 HDP 模型构建重分层的 Dirichlet 过程, 提出 rHDP (reallocated hierarchical Dirichlet processes) 层次主题模型; 第三, 结合领域分类信息、词语语义和主题词的领域代表性, 定义领域知识, 包括基于投票机制的领域隶属度、词语与领域主题的语义相关度和层次化的主题-词语贡献度; 最后, 通过领域知识改进 rHDP 主题模型中领域主题和主题词的分配过程, 提出结合领域知识的层次主题模型 rHDP_DK (rHDP with domain knowledge), 并改进采样过程. 实验结果表明, 基于 nCRP+的层次主题模型在评价指标方面均优于基于 nCRP 的层次主题模型 (hLDA, nHDP) 和神经主题模型 (TSNTM); 通过 rHDP_DK 模型生成的主题层次结构具有领域主题层次清晰、关联子主题的主题词领域差异明确的特点. 此外, 该模型将为领域主题层次提供一个通用的自动挖掘框架.

关键词: 层次主题模型; 领域分类信息; 词语语义; 主题关联关系; 层次化的采样过程; 领域主题层次

中图法分类号: TP311

中文引用格式: 万常选, 张奕韬, 刘德喜, 刘喜平, 廖国琼, 万齐智. 主题方面共享的领域主题层次模型. 软件学报, 2024, 35(4): 1790–1818. http://www.jos.org.cn/1000-9825/6840.htm

英文引用格式: Wan CX, Zhang YT, Liu DX, Liu XP, Liao GQ, Wan QZ. Domain Topic Hierarchy Model for Topic Aspect Sharing. Ruan Jian Xue Bao/Journal of Software, 2024, 35(4): 1790–1818 (in Chinese). http://www.jos.org.cn/1000-9825/6840.htm

Domain Topic Hierarchy Model for Topic Aspect Sharing

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

[1](School of Information Managment, 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)

Abstract: The hierarchical topic model is an important tool to organize topic hierarchy. Most of the existing hierarchical topic models provide tree-structured prior distributions for document topics by introducing the nCRP construction method into the topic models, but they cannot acquire a topic hierarchy with clear domain meanings, referred to as domain topic hierarchy. Meanwhile, there are not only

* 基金项目: 国家自然科学基金 (61972184, 62272205, 62272206, 62076112)
收稿时间: 2022-03-09; 修改时间: 2022-06-28, 2022-09-29; 采用时间: 2022-11-30; jos 在线出版时间: 2023-07-28
CNKI 网络首发时间: 2023-08-01

© 中国科学院软件研究所 http://www.jos.org.cn

hierarchical relationships among domain topics but also sub-topic aspect sharing relationships under different parent topics. There is no appropriate model that yields such domain topic hierarchy in the current research on topic relationships. In order to automatically and effectively mine the hierarchical and correlated relationships of domain topics from domain texts, improvements are put forward as follows. Firstly, this study improves the nCRP construction method through the topic sharing mechanism and proposes the nCRP+ hierarchical construction method to provide a tree-structured prior distribution with hierarchical topic aspect sharing for topics generated from topic models. Then the reallocated hierarchical Dirichlet processes (rHDP) are developed based on nCRP+ and HDP models, and an rHDP model is proposed. By employing the domain taxonomy, word semantics, and domain representation of topic words, the study defines domain knowledge, including the domain membership degree based on the voting mechanism, the semantic relevance between words and domain topics, and the contribution degree of hierarchical topic words. Finally, domain knowledge is used to improve the allocation processes of domain topics and topic words in the rHDP model, and rHDP with domain knowledge (rHDP_DK) model is proposed to improve the sampling process. The experimental results show that hierarchical topic models based on nCRP+ are superior to those based on nCRP (hLDA and nHDP) and neural topic model (TSNTM) in terms of evaluation metrics. The topic hierarchy, built by the rHDP_DK model, is characterized by clear domain topic hierarchy and explicit domain differences among related sub-topics. Furthermore, the model will provide a general automatic mining framework for domain topic hierarchy.

Key words: hierarchical topic model; domain taxonomy; word semantics; correlated relationships of topics; hierarchical sampling process; domain topic hierarchy

互联网平台中存在大量与领域相关的文本数据 (简称领域文本), 这些文本数据不仅描述了领域在各个子领域的发展状况, 还刻画了领域融合发展的特点. 本文将反映领域、子领域和融合领域涵义的主题定义为领域主题, 领域主题关系则间接地反映了领域与子领域、领域与融合领域之间的内在关系. 因此, 领域主题关系的研究将为分析领域发展现状、洞察融合发展形态、辅助发展趋势预测等方面提供技术支撑.

以财经领域相关文本 (简称财经文本) 为例, 研究者按经济总量的产出方面将宏观经济分成 5 个一级经济指标[1], 国家统计部门进一步将其细分为 22 个子领域 (简称为二级经济指标). 这些一级经济指标与二级经济指标之间存在纵向的层次关系, 且不同一级经济指标下的二级经济指标之间还存在横向的关联关系. 将经济指标对应领域主题, 财经文本中领域主题的层次关系和关联关系可用图 1 所示领域主题层次表示.

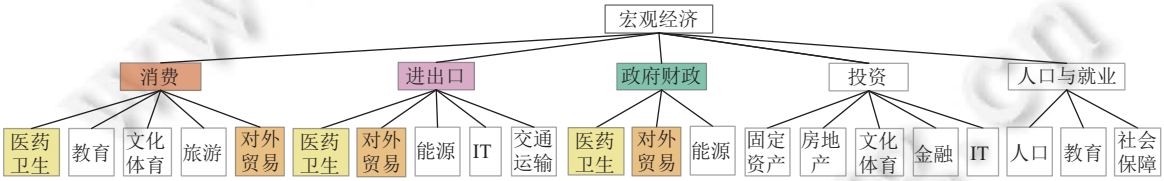


图 1 财经文本中领域主题的层次关系和关联关系

本文将同一层级中不同领域主题下共享的主题方面 (或子主题) 定义为关联子主题, 如图 1 中“医药卫生”“对外贸易”等子主题. 首先, 在消费、进出口、政府财政等一级经济指标 (主题) 下都有“医药卫生”二级经济指标 (子主题), 以反映“医药卫生”子领域在不同一级经济指标方面 (侧面) 的发展情况, 显然不同主题下这些共享的主题方面 (或子主题) 之间是存在关联关系的; 其次, 在分析领域主题词 (即反映主题涵义的词语) 时发现, 关联子主题之间还存在领域差异, 例如, “医药卫生”作为消费、进出口、政府财政等主题下的关联子主题, 分别体现了“医药卫生消费”“医药进出口”“医药卫生保障”的涵义, 即关联子主题的主题词应该是有差异的. 类似的现象在新闻文本 (例如, 20NewsGroup) 中也存在.

在构建主题层次关系和关联关系方面, 现有相关研究成果主要包括: ① 基于 nCRP (nested Chinese restaurant process) 的层次主题模型通过 nCRP 构造方法[2−4]为主题模型中的文档主题提供先验分布, 自动生成主题树形结构并抽取主题词; ② 通过种子词的树形关系引导机器学习方法生成主题树形结构[5,6], 描述主题层次关系, 该主题层次关系存在对固定的树形结构过分依赖的问题; ③ 结合深度模型中的多层 Gamma 信念网[7]、Dirichlet 信念网[8]或神经网络[9]为文档主题或主题词语提供层次化的先验分布, 生成树形结构的主题层次关系, 但是这类主题结构无法描述主题方面的关联关系; ④ 结合深度模型中的 Sigmoid 信念网[10]为文档主题提供先验分布, 生成主题关联

© 中国科学院软件研究所 http://www.jos.org.cn

图, 挖掘主题关联关系, 然而挖掘的关联主题只是主题网状结构中的公共节点, 导致其主题词没有差异性, 且主题层次隶属关系也不明确.

在共享主题分配方面, 研究者在主题模型的两层 DP (Dirichlet process) 结构中, 通过上层 DP 为下层若干 DP 过程提供共享的主题分布[11]. 在抽取领域主题和主题词方面, 研究者通过领域信息改进主题模型中主题和主题词分配机制, 明确领域主题的差异性[12].

本文将采用 nCRP 的主题层次构造原理实现主题层次自动构建. 然而, 在领域文本中使用基于 nCRP 的层次主题模型构建主题树, 该主题树无法明确地描述主题的领域属性以及主题间的领域关联性, 也就无法明确关联子主题的差异性. 主要原因包括, 首先, 主题树中同一层级的主题都是通过独立的 CRP (Chinese restaurant process, 中国餐馆经营过程) 生成, 导致同一层级中不同分支主题下的子主题之间缺少关联性; 其次, 模型通过词语在所有文档中的共现程度和出现频率确定词语的抽象程度, 进而构建主题的层次关系并抽取主题词, 导致不同层级的主题之间缺少领域隶属性, 主题词也无法明确地描述不同层级主题的领域涵义.

针对以上问题, 本文将结合主题模型中共享主题分配和领域主题抽取的思路, 分别从主题层次构造、主题分配、主题词选择和模型采样等方面改进基于 nCRP 的层次主题模型, 解决领域主题层次自动构建及其关联子主题挖掘的问题.

本文的主要贡献如下.

(1) 在 nCRP 中引入共享主题分配机制, 提出 nCRP+构造方法, 改变主题树形结构的先验分布; 基于 nCRP+构造方法和主题模型, 提出重分层的层次狄利克雷过程 (reallocated hierarchical Dirichlet processes, rHDP), 实现分层级的主题统一分配和主题方面共享.

(2) 在 rHDP 层次主题模型中, 通过领域类别信息定义基于投票机制的领域隶属度计算方法, 构建主题 (或子主题) 与领域 (或子领域) 之间的映射关系; 利用待分配词语与主题中已分配词语的语义相关性, 定义词语与领域主题的语义相关度, 提升主题词刻画领域主题涵义的能力; 通过词语与其所在分支主题的领域相关性, 定义层次化的主题-词语贡献度, 明确关联子主题中主题词的差异性. 因此, 本文的领域知识包括主题的领域信息、词语与主题的语义信息以及主题词的领域关联信息.

(3) 基于领域知识改进 rHDP 层次主题模型采样和后验概率推导过程, 提出一种领域主题层次自动生成算法.

本文第 1 节介绍主题层次构造的相关研究进展, 并分析这些模型在领域主题关系构建和主题词抽取中存在的问题. 第 2 节阐述 nCRP+层次构造方法和 rHDP 层次主题模型的基本思想. 第 3 节定义领域知识, 并提出结合领域知识的层次主题模型. 第 4 节为实验结果分析. 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 相关工作

根据主题层次构建原理和领域主题抽取方法, 本节将从 3 个方面分析领域主题层次构建的相关工作, 包括基于主题模型构建主题层次结构、基于机器学习方法构建主题层次结构和基于领域知识的主题模型或层次主题模型.

1.1 基于主题模型构建主题层次结构

为了明确主题层次结构、细化主题涵义, 研究者通过基于 HDP (hierarchical Dirichlet processes)[11]主题模型和基于嵌套 CRP (Chinese restaurant process)[2]的层次主题模型, 生成主题层次结构.

(1) 基于 HDP 的主题层次构建

Whye 等人[11]将 HDP 主题模型扩展成 3 层 DP 结构的主题模型, 构建领域文档的主题层次分布. 流式层次狄利克雷过程 FHDP (flow HDP)[13]通过在 HDP 中增加流动操作, 按主题生成的先后顺序定义主题的抽象程度和层次隶属关系, 存在主题层次隶属关系不清晰的问题.

Ma 等人[14]通过在 HDP 模型中加入标签区分全局或局部主题, 以及全局或局部词语; Ding 等人则结合“亚主题”和“层次映射”概念提出 nHDP (nested HDP)[15]和 mnHDP (mapped nHDP)[16]层次主题模型, 构建主题层次. 然而, 这类模型受限于主题层次扩展或主题方面共享.

© 中国科学院软件研究所 http://www.jos.org.cn

(2) 基于 nCRP 的层次主题模型

在构建主题树形结构方面, 研究者们采用 nCRP 构造方式生成主题层次的树形先验分布, 主题层次深度不受限制, 再结合不同的主题模型构建层次主题模型.

Blei 等人在 LDA (latent Dirichlet allocation) 主题模型中引入 nCRP 过程, 为文档主题提供树形结构的先验分布, 提出 hLDA (hierarchical LDA) 层次主题模型[2]. 也有研究者研究 hLDA 模型的采样效率[17−19]和时变主题的层次结构[20]. 然而, 由于 hLDA 主题模型生成文档-主题分布只限定在树中某一分支, 模型无法全面地分析主题分布.

为了解决文档主题分布路径单一的问题, 研究者将主题共享的思想应用于主题树形结构的共享, 实现从主题树分支中自由组合主题分布, 主题层次宽度也不受限制, 层次结构更加灵活[21].

nHDP (nested hierarchical Dirichlet processes) 层次主题模型[3]结合 nCRP 和 HDP 的思想, 为文档主题提供树形结构的基分布. 在 nHDP 模型中, 每篇文档主题不再局限于树中一条分支路径上的节点, 而是来源于整棵树中若干分支节点.

Ahmed 等人[4]把层次结构建模和文档建模分开, 通过改进的吉布斯采样方法, 提出 nCRF (nested Chinese restaurant franchise) 构造过程, 实现主题层次结构共享. Huang 等人[22]利用 nCRP 和多层 HDP 实现 BHMC (Bayesian hierarchical mixture clustering) 层次主题模型, 构建主题层次分布.

在实际应用中, 基于 nCRP 的层次主题模型在构建领域主题层次中存在以下问题: ① 主题与领域之间的映射不准确, 导致主题无法明确地描述对应的领域涵义; ② 主题关系无法同时体现领域主题之间的层次关系以及同一层级多个主题下主题方面的共享关系; ③ 生成的主题词既不能较好地描述父主题对其下子主题的概括性, 也不能较好地体现关联子主题之间的领域差异性.

1.2 基于机器学习方法构建主题层次结构

除了主题模型和层次主题模型, 通过机器学习方法生成主题层次结构也是研究的热点. 一种普遍的方法是结合矩阵分解的思想, 构建文档-主题和主题-词语分布. 研究者通过成对主题的关联矩阵构建 PAM (Pachinko allocation model) 模型[23,24], 将子主题的父主题定义为子主题的先验分布, 使得主题分布中既包含词语也包含其他主题, 构建主题层次关系. 研究者结合非负矩阵分解 (non-negative matrix factorization, NMF)[25,26], 提出基于正交约束的分层稀疏 NMF 主题模型, 构建主题层次.

在深度模型方面, 为了构建主题层次关系, 研究者通过多层伽马信念网络 (multi-layer Gamma belief network)、泊松因子分析 (Poisson factor analysis, PFA) 和词向量提出 WEDTM (word embeddings deep topic model) 模型[7], 或结合 Dirichlet 信念网[8]生成主题的先验分布, 构建主题的深层次表示; 为了构建主题关联关系, 研究者基于泊松因子分析, 通过 Sigmoid 信念网 (Sigmoid belief network, SBN), 生成主题特征矩阵的深层表示, 为文档主题提供先验分布, 生成主题的网状关联图[10], 其中包含了一些主题之间共同关联的主题, 但关联主题的主题词是没有差异的, 且主题的层次关系也不明确.

Isonuma 等人[9]基于自编码变分贝叶斯 (auto-encoding variational Bayes, AEVB), 构建主题的先验分布, 提出树形结构的神经主题模型 TSNTM (tree structured neural topic model), 生成主题树形结构. 也有研究者通过在模型中增加种子信息, 构建基于弱监督的、满足用户需求的主题层次构建模型[5,6,27].

然而, 以上研究存在如下问题: ① 主题与领域的对应关系不明确; ② 主要侧重于主题的层次关系构建, 忽略了相同层级主题下子主题间的关联关系; ③ 关联主题是主题网状结构中的公共节点, 导致其主题词没有差异性, 主题层次关系也不明确.

1.3 基于领域知识的主题模型/层次主题模型

为了抽取领域文本中的领域主题, Chen 等人[28]利用领域词语的频繁项集发现词语的不共现性, 以词语 cannot-link 集构建领域知识, 引导 LDA 主题模型中主题-主题词的分配过程. 在利用词语语义信息方面, 研究者在主题模型中通过同义词[29]或词向量[30]明确词语之间的语义相似性, 以词语 must-link 集构建领域知识, 改进 LDA 主题模型, 使得语义相近的词语尽量分配在相同主题中, 提高主题词对领域主题的描述力.

© 中国科学院软件研究所 http://www.jos.org.cn

在基于 HDP 主题模型的领域主题提取方面, Liu 等人[31]通过时间信息、用户信息和标签信息构建领域知识, 引导文档-主题的分配过程, 提取微博文本主题. Zhang 等人[12]通过文本的领域类别信息和词语语义信息生成领域知识, 改进文档-主题和主题-词语的分配过程, 提取领域主题和反映领域主题涵义的主题词.

为了生成领域主题层次结构, Alam 等人[32]结合预定义的领域类别层次和 HDP, 提出 ICB (imperial Chinese banquet) 构造方法和 THDP (tree-structure HDP) 主题模型, 构建了同一领域的主题层次关系, 该模型没有考虑领域主题之间的横向关联关系. 另外, Xu 等人[33]通过 hLDA 模型生成的主题和主题层次, 实现领域知识的自动获取, 在假设相近学术领域文本中主题或主题层次结构相同的前提下, 为相近领域文本构建领域主题层次. 然而, 对于领域涵义相差较大的领域文本, 这类模型无法准确地构建领域主题层次.

因此, 在层次主题模型中, 本文将借鉴领域知识在文档-主题和主题-词语分配过程中的影响机制, 改进领域主题和主题词的抽取效果.

2 主题方面共享的层次主题模型

本文首先提出主题方面共享的层次主题模型, 下面就其中涉及的层次构造方法和主题层次模型予以阐述.

2.1 nCRP+层次构造方法

中国餐馆过程 (CRP)[11]假设餐厅的每张餐桌都有一道菜肴与之对应, 通过为每个顾客分配餐桌间接实现该顾客的菜肴分配. 如果将顾客、菜肴和餐厅分别对应词语、主题和文档, 已有研究通过 CRP 过程描述 nCRP[2]的原理, 实现主题关系构建. nCRP 构造方法包括若干个 CRP 过程, 每个 CRP 过程对应一道菜肴分配, 嵌套调用若干个 CRP 过程则为每个顾客分配一系列体现层次关系的菜肴. 应用在主题模型中, 嵌套的 CRP 过程可用于构建主题的层次结构.

另外, 已有研究通过 CRP 过程描述 CRF (Chinese restaurant franchise) 构造方法[11]的原理, 实现主题统一分配. CRF 构造方法将 CRP 过程分为两种类型, 即顾客-餐桌 CRP 和餐桌-菜肴 CRP 过程, 分别实现所有顾客的餐桌分配和所有餐桌的菜肴分配. CRF 构造方法首先通过若干个顾客-餐桌 CRP 过程, 为每个餐厅中的每个顾客分配餐桌; 然后通过一个餐桌-菜肴 CRP 过程为所有的餐桌统一分配菜肴, 且每张餐桌只分配一道菜肴. 应用在主题模型中, CRF 分配过程通过统一的菜肴分配共享主题基分布, 实现文档中词语主题的统一分配.

在 nCRP 构造方法中, 由于主题树形结构中同一层级不同分支中的节点是通过若干独立的 CRP 过程生成, 导致同一层级不同分支节点的语义描述是相互独立的. CRF 构造方法通过在相同基分布中抽样出若干独立 CRP 过程之间的聚类分布, 使得聚类结果中蕴含这些 CRP 过程的共性, 体现在主题模型中就是实现文档之间的主题共享. 本文基于 CRF 的共享特性改变 nCRP 构造方法中相同层级 CRP 的分配过程, 即通过为每个层级中隶属于相同父节点的节点, 构建 CRF 分配过程, 设计分层级的主题共享机制, 从而实现这些节点之间语义涵义的共享, 并将这种基于 CRF 改进的 nCRP 构造方法称为 nCRP+层次构造方法.

为了实现层次化的共享, nCRP+构造方法定义了餐厅的菜肴风格和餐桌供应菜肴的菜肴类别, 并将它们分别对应为文档的领域属性和子领域属性. 同时, nCRP+方法假设每个餐厅拥有特定的菜肴风格 (如粤菜、川菜等), 且这些菜肴风格在所有餐厅中是共享的; 将各种菜肴风格中的所有菜肴划分成若干个菜肴类别 (如炒菜类、蒸菜类等), 且这些菜肴类别在不同菜肴风格的餐厅中是共享的. 基于这种假设, 相同菜肴风格的餐厅中可以供应不同菜肴类别的菜肴, 不同菜肴风格的餐厅中也可以供应相同菜肴类别的菜肴. 对应在层次主题模型中, 这种分配机制可实现同一层级中不同主题下子主题方面的共享.

在实现这种分层级的主题方面共享机制中, nCRP+构造方法将 nCRP 过程中每个父节点下的 CRP 过程扩展为 CRF 过程, 并且同一层级的 CRF 过程是共享基分布的.

以构建具有两层主题方面共享的主题层次为例, nCRP+层次构造方法的框架如图 2 所示, 包括两层 CRF 过程. 其中, 上层 CRF 包括一个 CRF 分配过程, 用于实现所有餐厅菜肴风格的统一分配, 目的是对具有相同菜肴风格的餐厅分配菜肴风格, 即首先明确餐厅的菜肴风格. 下层 CRF 由若干个 CRF 分配过程组成, 用于分别实现不同菜肴

© 中国科学院软件研究所 http://www.jos.org.cn

风格的餐厅中菜肴类别的统一分配, 即在餐厅菜肴风格明确的前提下, 进一步确定餐桌分配菜肴的菜肴类别, 用于在不同菜肴风格的餐厅中实现菜肴类别的共享. 为了描述这种分层次的共享机制, 需要在每个层级中增加相应的共享基分布, 其中上层 CRF 通过餐桌-菜肴 CRP 过程实现, 下层 CRF 则通过在其下的每一个餐桌-菜肴 CRP 中共享相同的基分布. 为了描述同一层级不同 CRF 中的共享机制, 图 2 中用灰色背景表示下层 CRF 的基分布共享区域.

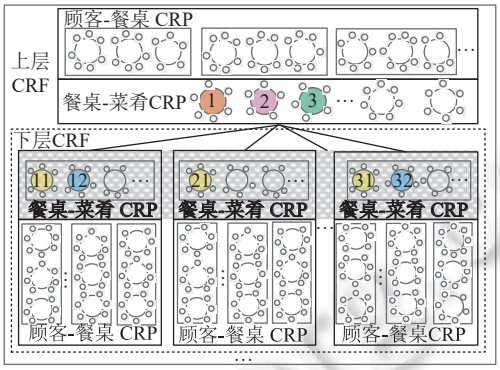


图 2 nCRP+层次构造方法的框架

利用上述 nCRP+层次构造方法生成的三层树形结构如图 3 所示, 其中, 树根节点 (黑色实心圆) 及其第 1 层节点 (节点 1, 2, 3) 是通过图 2 中上层 CRF 过程构造得到, 且这些节点分布是共享第 1 层基分布的. 在第 1 层节点分布的基础上, 通过图 2 中下层 CRF 生成第 2 层节点, 目的是为第 1 层的节点生成其在第 2 层的子节点, 且这些子节点是共享第 2 层基分布的. 因此, nCRP+层次构造方法包含了两方面关系的构建. 一方面, 利用上层节点分布, 通过嵌套调用 CRF 构造方法生成下层子节点分布, 形成节点之间的层次结构, 描述节点之间的隶属关系; 另一方面, 通过分层级的基分布共享, 实现在同一层级不同节点中共享下层子节点的分布共性, 明确同一层级不同节点在下层子节点之间的语义关联关系, 本文用相同颜色的圆表示主题方面共享的节点, 例如 1、2、3 节点下的 11、21、31 子节点, 以及 1、3 节点下的 12 和 32 子节点.

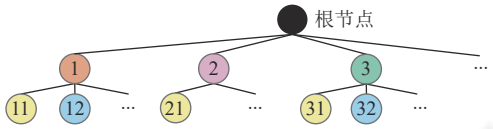


图 3 nCRP+层次构造方法对应的三层树形结构

2.2 rHDP 层次主题模型

nCRP+构造方法改变了主题层次结构的生成机制, 结合 HDP 主题模型的采样方法, 可重新定义主题与子主题的父子关系, 以及同一层级主题下子主题的关联关系, 实现对文档主题层次的重新分层. 因此, 本文将基于 nCRP+层次构造方法和 HDP 模型构建的层次主题模型称为重分层的层次狄利克雷过程 (reallocated hierarchical Dirichlet processes, rHDP), 模型中使用的符号说明见表 1.

将图 2 中的两层 CRF 结构扩展为多层 CRF, rHDP 模型可构建主题的多层结构. 将每个 CRF 过程对应成一个两层 DP 过程, 即上层 DP 过程和下层 DP 过程. 结合表 1 中的符号, 定义第 l 层 CRF 的抽样过程. 通过上层 DP 过程抽样产生第 l 层节点的全局随机概率测度, 表示为 G_0^l; 通过下层 DP 过程抽样产生词语集合 j 的随机概率测度, 表示为 G_j^l, 则第 l 层 CRF 对应的双层 DP 过程如公式 (1) 所示. 对于第 l 层的所有 CRF 过程, 基分布 H^l 是共享的.

$$\begin{cases} G_0^l \mid \gamma^l, H^l \sim DP(\gamma^l, H^l) \\ G_j^l \mid \alpha^l, G_0^l \sim DP(\alpha^l, G_0^l) \end{cases} \tag{1}$$

在 nCRP+层次构造方法中, 当 l=1 时, 通过公式 (1) 定义的基分布和超参数, 调用 1 次 CRF 分配过程生成第 0

© 中国科学院软件研究所 http://www.jos.org.cn

层和第 1 层节点, 即生成第 0 层和第 1 层主题, 其中第 1 层的每个主题中包括来自不同文档的词语分组中的词语; 为了满足下层 CRF 过程对输入文档的要求, 将划分到该主题的词语分组按照其所在文档进行合并, 得到若干个词语子集. 结合表 1 符号定义, 第 l–1 层的主题数可表示为 $\left|K^{l-1}\right|$, 并作为第 l 层 CRF 的输入参数. 因此, 当 l≥2 时, 第 l 层的 CRF 通过调用 $\left|K^{l-1}\right|$ 次两层 DP 过程生成第 l 层中每一个词语子集的主题分布.

表 1 rHDP 层次主题模型符号说明

符号	说明	符号	说明
l	符号所处的层次	t_{ji}	j 餐厅顾客 i 所坐的餐桌
G_0	全局随机概率测度	T_j	j 餐厅中已分配顾客的餐桌集合
G_j	文档 (词语子集) j 的随机概率测度	X_j^t	j 餐厅中就座餐桌 t 的顾客集合
H	上层DP的基分布	θ_{ji}	j 餐厅中顾客 x_{ji} 的餐桌分配参数
γ	上层DP的超参数	$\varphi_j^{t_{ji}}$	j 餐厅中餐桌 t_{ji} 的菜肴分配参数
α	下层DP的超参数	Φ	菜单中的菜肴集合
β	主题-词语分布的超参数	ϕ^k	菜肴 k 的分布参数
J	所有餐厅的集合	T^k	供应菜肴 k 的餐桌集合
j	单个餐厅编号	K	所有餐桌已供应的菜肴集合
x_{ji}	j 餐厅中的第 i 个顾客	$\delta_{\varphi_j^t}$	顾客就座餐厅 j 中餐桌 t 的概率分布参数
X	所有顾客集合	δ_{ϕ^k}	餐桌分配菜肴 k 的概率分布参数

通过每个层级中的两层 DP 过程生成文档 (或词语子集) 的主题分布. 假设 $\left\{\theta_{j1}^l,\theta_{j2}^l,\ldots,\theta_{ji}^l\right\}$ 是服从 G_j^l 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^l, 此时 θ_{ji}^l 对应词语 x_{ji}^l 的主题分布参数, $F(\theta_{ji}^l)$ 表示在给定参数 θ_{ji}^l 下词语 x_{ji}^l 的主题分布, 如公式 (2) 所示.

$$\begin{cases}\theta_{ji}^l \mid G_j^l\\ x_{ji}^l \mid \theta_{ji}^l \sim F(\theta_{ji}^l)\end{cases} \tag{2}$$

对于第 l 层的每个 CRF 分配过程, 统计餐厅 j 中餐桌 t^l 的顾客数和供应菜肴 k^l 的餐桌数, 分别记为 $\left|X_j^{t^l}\right|$ 和 $\left|T^{k^l}\right|$. 在新顾客分配餐桌时, 被分配到已有餐桌的概率与该餐桌已有顾客数成正比, 被分配到新餐桌的概率与该层参数 α^l 成正比, 顾客-餐桌的分配过程如公式 (3) 所示. 在新餐桌分配菜肴时, 被分配到已有菜肴的概率与该菜肴供应的餐桌数成正比, 被分配到新菜肴的概率与该层参数 γ^l 成正比, 餐桌-菜肴分配过程如公式 (4) 所示.

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

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

每层级 CRF 分配过程对应每层级节点的主题分配以及主题词的抽取; 多层 CRF 分配过程则对应多层主题关系的构建及其主题词的抽取. 通过采样和后验概率推导构造 rHDP 层次主题模型, 为每一篇文档生成层次化的主题分布以及主题-词语分布.

3 结合领域知识的 rHDP 层次主题模型

为了获取领域主题的层次关系、关联关系以及关联子主题的差异性, 基于领域知识改进 rHDP 层次主题模型, 构建结合领域知识的 rHDP 层次主题模型 (rHDP with domain knowledge, rHDP_DK).

以构建经济指标的 3 层主题树形结构为例, 如前文图 1 所示, 假设顾客对应词语、菜肴对应主题, 餐厅对应文档, 文档的表现形式是词语集. 将 nCRP+构造方法应用到财经文本中, 第 1 层 CRF 中的餐厅对应财经文本的词语集, 餐厅的菜肴风格对应财经文本的经济领域大类属性, 即对应一级经济指标的主题属性; 第 2 层 CRF 中的餐厅

© 中国科学院软件研究所 http://www.jos.org.cn

对应词语子集, 餐桌中供应菜肴的菜肴类别对应财经文本的经济子领域属性, 即对应二级经济指标的子主题属性.

3.1 基于投票机制的领域隶属度

在图 1 中, 一个一级经济指标下会有若干个二级经济指标, 且不同的一级经济指标下存在相同的二级经济指标, 反映多个一级经济指标 (主题) 下经济子领域 (子主题) 的交叉性. 对应在 nCRP+过程中, 通过每层 CRF 中统一的菜肴分配过程, 餐厅的菜肴风格在不同餐厅中是共享的, 餐桌供应菜肴的菜肴类别在不同菜肴风格的餐厅中也是共享的. 在给新餐桌分配菜肴之前, 餐厅的菜肴风格和餐桌供应菜肴的菜肴类别分别在上层 CRF 和下层 CRF 分配过程中确定, 目的是对具有相同菜肴风格的餐厅分配菜肴风格, 并进一步对相同或不同菜肴风格的餐厅中对菜肴类别有要求的餐桌分配菜肴的菜肴类别. 在领域文本中, 第 1 层、第 2 层 CRF 中的餐厅分别对应领域文本的词语集和重新划分主题后的词语子集, 通过在文档或词语子集的主题分配过程中共享领域划分信息, 分步明确文档的领域和子领域属性, 解决领域分类信息引入和子主题的共享问题.

在领域主题层次生成过程中, 已知的领域类别及其对应的代表性词语集 (简称领域词语集) 来源于专家定义. 在对文档或重新划分主题后的词语子集 (两者统称为词语集) 分配主题时, 首先, 通过计算词语集中每个词语与领域词语集中词语的语义相关性, 明确词语集中每个词语的领域属性; 然后, 分析词语集中所有词语的领域分布情况, 通过词语分布较多的领域类别明确该词语集的领域属性; 最后, 利用词语集的领域属性引导词语集的主题分配过程, 使得同一层级的主题之间具有明显的领域特性. 基于这种思想, 本文提出一种基于投票机制的领域隶属度计算方法, 计算流程如图 4 所示.

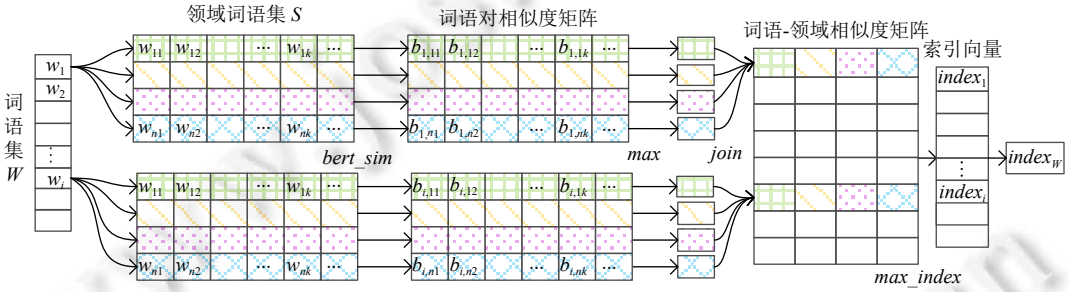


图 4 基于投票机制的领域隶属度计算流程

在图 4 中, 领域词语集记为 S, S_n 表示第 n 个领域词语集, $S_n=\{w_{nk}|0\leqslant n<|S|, 0\leqslant k<|S_n|\}$; 词语集记为 W, $W=\{w_i|0\leqslant i<|W|\}$; $bert\_sim(w_i, w_{nk})$ 表示词语对 w_i 与 w_{nk} 之间的 BERT (bidirectional encoder representations from Transformers) 语义相似度, 记为 $b_{i,nk}$; 根据词语对相似度矩阵, max 函数为词语集中每个词语筛选出其在每个领域中的语义相似度最大值, $join$ 函数将所有词语在各个领域中的语义相似度最大值拼接成一个词语-领域相似度矩阵, 描述每个词语与所有领域的相似度分布情况; max_index 函数为词语集中每个词语生成领域隶属索引值, 并为该词语集构建一个索引向量, 其中 $index_i$ 表示 w_i 的领域隶属索引值, 如公式 (5) 所示; 分析索引向量中领域的分布情况, 通过投票的方式生成词语集的领域隶属索引值, 用 $index_W$ 表示词语集 W 的领域隶属索引值, 如公式 (6) 所示; 根据与词语集 W 的领域类别相同的词语集个数, 定义词语集 W 隶属于该领域类别的程度, 即词语集 W 的领域隶属度, 记为 A_W, 如公式 (7) 所示.

$$index_i=\arg\max_{0\leqslant n<|S|}\left(\max_{0\leqslant k<|S_n|}(bert\_sim(w_i,w_k))\right) \tag{5}$$

$$index_W=\arg\max_{0\leqslant n<|S|}\sum_{0\leqslant j<|W|,i\neq j}I(index_i=index_j,index_i=n) \tag{6}$$

$$A_W=\sum_{W'\neq W}I(index_W=index_{W'}) \tag{7}$$

因此, 通过公式 (7) 可以明确文档或重新划分主题后的词语子集的领域隶属度, 改进每层级文档 (或词语子集)-主题的分配过程.

© 中国科学院软件研究所 http://www.jos.org.cn

3.2 词语与领域主题的语义相关度

在各层 CRF 顾客-餐桌分配过程中, 通过分析顾客之间对餐厅分配的菜肴风格或餐桌分配菜肴的菜肴类别要求的一致程度, 将要求相近的顾客分配在一起, 即将隶属于相同领域或子领域范畴的词语分配在相同的主题中, 改善同一领域词语的聚类效果. 本节通过词语的语义信息描述顾客之间对餐厅分配的菜肴风格或餐桌分配菜肴的菜肴类别要求的一致程度, 结合文献 [12], 定义词语与领域主题 (或词语与领域子主题) 的语义相关度. 通过统计待分配主题的词语 w_i 在主题 k 中语义相似的词语个数, 定义词语 w_i 与主题 k 的语义相关度, 记为 $A(w_i,k)$, 如公式 (8) 所示. 其中, *bert_sim* 表示词语之间的 BERT 语义相似度, ξ 表示语义相似度的阈值, X^k 表示主题 k 的词语集.

$$A(w_i,k)=\left|\left\{w_m \middle| bert\_sim(w_i,w_m)\geqslant \xi, w_m\in X^k, w_m\neq w_i\right\}\right| \tag{8}$$

3.3 层次化的主题-词语贡献度

在明确了主题领域属性、词语与主题的语义相关性的基础上, 需要进一步明确每层主题中主题词的领域代表性. 在构建领域主题层次时, 需要根据领域主题的层次关系和关联关系, 结合词语所在的层级和分支, 抽取每层级领域主题对应的主题词. 对应 nCRP+层次构造过程中, 需要精确区分顾客群体, 即根据顾客对餐厅菜肴风格和餐桌供应菜肴的菜肴类别的要求对顾客进行细分, 体现顾客对特定菜肴风格下特定菜肴类别的菜肴的“专一性”. 因此, 在每层主题-词语的分配过程中, 通过分析词语对各层级主题的代表性, 改进词语在主题中的分配概率, 定义层次化的主题-词语贡献度, 抽取符合各层领域主题涵义的主题词.

将刻画经济主题或子主题涵义的主题词称为经济要素词, 这些词语将通过层次主题模型从财经文本中自动抽取. 根节点对应主题树的第 0 层节点, 其主题是对所有财经文档的概括, 所以根节点主题中词语的贡献度由词语在第 1 层经济主题 (即一级经济指标) 中的出现概率和出现频次定义. 第 1 层节点主题是对其下所有词语子集的概括, 所以第 1 层节点主题中词语的贡献度由词语在第 2 层经济主题 (即二级经济指标) 中的出现概率和出现频次定义. 这里第 2 层经济主题是第 1 层经济主题下的经济子主题, 对应主题树中的第 2 层节点, 该层节点主题需要明确不同一级经济指标下相同二级经济指标的涵义. 所以, 第 2 层节点主题中词语的贡献度由词语与其所在分支中各级经济主题的关联关系共同定义. 通过逐层地计算词语在主题中的贡献度, 将能同时描述同一分支路径中经济主题和子主题的词语分配在对应的主题中, 提高经济要素词在关联子主题中的区分度.

通过在符号右上角加数字表示该符号所描述的层级, 定义层次化的主题-词语贡献度. $phi(w_i, k^1)$ 定义为词语 w_i 在第 1 层主题 k^1 中出现的概率, $stat(w_i, k^1)$ 定义为词语 w_i 在第 1 层主题 k^1 中出现的次数; 将词语 w_i 在第 2 层主题 k^2 中的逆主题频率记为 $itf^2_{w_i}$, 结合 $phi(w_i, k^2)$ 描述词语 w_i 对第 2 层主题 k^2 的代表性; 通过公式 (5) 中 *bert_sim* 函数计算 w_i 与其所在分支中第 1 层节点主题 k^1 中所有词语的语义相似值, 并用其平均值表示词语 w_i 与主题 k^1 的语义相似值, 记为 $bert\_sim(w_i, k^1)$, 用于描述词语 w_i 与第 1 层主题 k^1 的语义关系, 通过参数 λ 调节词语与其所在分支中第 1、2 层主题的相关性. 因此, 主题树中层次化的主题-词语贡献度记为 $C(w_i, k^l)$, 计算公式如公式 (9) 所示.

$$C(w_i,k^l)=\begin{cases} phi(w_i,k^1)\times stat(w_i,k^1), & l=0\\ phi(w_i,k^2)\times stat(w_i,k^2), & l=1\\ \lambda\times\left(phi(w_i,k^2)\times itf^2_{w_i}\right)+(1-\lambda)\times bert\_sim(w_i,k^1), & l=2 \end{cases} \tag{9}$$

3.4 结合领域知识的层次主题模型参数概率分布

本节介绍如何通过领域隶属度、词语与领域主题的语义相关度和层次化的主题-词语贡献度改进 rHDP 层次主题模型中不同层级的参数概率分布. 在第 l 层 CRF 过程中, 餐厅 j 对应财经文本的词语集或重新划分主题后的词语子集. 结合公式 (8), 当新顾客分配餐桌时, 新顾客 x^l_{ji} 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餐厅分配的菜肴风格或餐桌分配菜肴的菜肴类别要求的一致程度 $A(x^l_{ji}, k^{t^l}_j)$ 成正比, 其中 $k^{t^l}_j$ 表示餐桌 t^l 所在餐厅的菜肴风格或所分配菜肴的菜肴类别; 被分配到新餐桌的概率与超参数 α^l 成正比, 第 l 层 CRF 的顾客-餐桌分配过程的参数如公式 (10) 所示.

因此, 新顾客可以选择与其要求一致程度较高的餐桌就座, θ^l_{ji} 对应 $\varphi^{t^l}_j$, 且餐桌编号为 t^l; 也可以选择新餐桌就

© 中国科学院软件研究所 http://www.jos.org.cn

座, 并由 G_0^l 采样生成新餐桌.

$$\theta_{ji}^l \mid \theta_{j1}^l,\theta_{j2}^l,\ldots,\theta_{j(i-1)}^l,\alpha^l,G_0^l \sim \sum_{t^l\in T_j^l}\frac{A(x_{ji}^l,k_j^{t^l})}{\sum_{t^l\in T_j^l}A(x_{ji}^l,k_j^{t^l})+\alpha^l}\delta_{\varphi_j^{t^l}}+\frac{\alpha^l}{\sum_{t^l\in T_j^l}A(x_{ji}^l,k_j^{t^l})+\alpha^l}G_0^l \tag{10}$$

如果新顾客选择新餐桌就座, 则需要为新餐桌分配菜肴. 新餐桌可被分配到其所在餐厅中的已有菜肴 k^l, 其分配概率或与相同菜肴风格的餐厅中供应该已有菜肴的餐桌数成正比, 或与供应该相同菜肴类别的已有菜肴的餐桌数成正比; 结合公式 (7) 计算词语集的领域隶属度, 改进餐桌-菜肴分配过程, 即计算相同菜肴风格的餐厅中供应该已有菜肴 k^l 的餐桌数, 或供应该相同菜肴类别的已有菜肴 k^l 的餐桌数, 记为 $A_j^{k^l}$, 如公式 (11) 所示; 被分配新菜肴的概率与超参数 γ^l 成正比, 第 l 层 CRF 的餐桌-菜肴分配过程的参数如公式 (12) 所示.

$$A_j^{k^l}=\sum_{\substack{j'\neq j,t^l\in T_j^l,t^{l'}\in T_{j'}^l\\ \text{s.t. } k_{jt^l}^l=k^l\wedge k_{j't^{l'}}^l=k^l}} I(index_j=index_{j'}) \tag{11}$$

$$\varphi_j^{t^l}\mid \varphi_1^1,\ldots,\varphi_1^{t^l},\varphi_2^1,\ldots,\varphi_2^{t^l},\ldots,\varphi_j^1,\ldots,\varphi_j^{t^l-1},\gamma^l,H^l \sim \sum_{k^l\in K^l}\frac{A_j^{k^l}}{\sum_{k^l\in K^l}A_j^{k^l}+\gamma^l}\delta_{\phi^{k^l}}+\frac{\gamma^l}{\sum_{k^l\in K^l}A_j^{k^l}+\gamma^l}H^l \tag{12}$$

因此, 新餐桌可以分配与其所在餐厅的菜肴风格相同的餐厅中的已有菜肴, 或分配与已有餐桌供应菜肴的菜肴类别相同的已有菜肴, 也可通过 H^l 分配新菜肴.

通过领域隶属度、词语与领域主题的语义相关度分别明确了主题领域属性、词语与主题的语义相关性, 结合公式 (9) 计算每层主题中主题词的领域代表性, 生成每层级领域主题的主题词分布.

3.5 模型的层次化采样

结合文档词语、主题的先验分布和 nCRP+分配过程, 改进各层级参数的 Gibbs 采样过程, 生成各层级参数的后验概率分布. 由于各层级参数采样过程需要明确参数或其涉及变量对应的层级, 通过上标 l 表示参数或变量所在的层级. 本节通过采样第 l 层的索引变量 t_{ji}^l 和 $k_j^{t^l}$, 生成第 l 层参数 θ_{ji}^l 和 $\varphi_j^{t^l}$.

(1) 计算变量 x_{ji}^l 和 $X_j^{t^l}$ 的条件概率

在层次化采样过程中, 通过每层级中的单个词语 x_{ji}^l 和单组词语集 $X_j^{t^l}$ 的采样实现索引变量 t_{ji}^l 和 $k_j^{t^l}$ 的采样. 结合文献 [12] 定义 x_{ji}^l 和 $X_j^{t^l}$ 的条件概率, 如公式 (13) 和公式 (14) 所示.

$$f_{k^l}^{-x_{ji}^l}(x_{ji}^l)=\begin{cases}\dfrac{n\_klv[v]}{|X^{k^l}|}, & k^l\in K^l\\ \dfrac{1}{|X^l|}, & k^l=k_{\text{new}}^l\end{cases} \tag{13}$$

$$f_{k^l}^{-X_j^{t^l}}(X_j^{t^l})=\begin{cases}\dfrac{\Gamma(|X^{k^l}|)}{\Gamma(|X^{k^l}|+|X_j^{t^l}|)}\dfrac{\prod_v\Gamma(n\_klv[v]+n\_jtlv[v])}{\prod_v\Gamma(n\_klv[v])}, & k^l\in K^l\\ \dfrac{\Gamma(|X^l|\beta^l)}{\Gamma(|X^l|\beta^l+|X_j^{t^l}|)}\dfrac{\prod_v\Gamma(\beta^l+n\_jtlv[v])}{\prod_v\Gamma(\beta^l)}, & k^l=k_{\text{new}}^l\end{cases} \tag{14}$$

其中, j 表示文档对应词语集 (或重新划分主题后的词语子集) 的编号, i 表示其中的词语编号, x_{ji}^l 表示词语集 (或词语子集) j 中第 i 个词语, 即餐厅 j 的第 i 个顾客. t^l 表示第 l 层的第 t 组词语, X^l 表示第 l 层中所有词语集合, X^{k^l} 表示第 l 层主题 k^l 的词语集合. v 是采样过程中词语 x_{ji}^l 对应的索引变量, $n\_klv[v]$ 表示第 l 层主题 k^l 中索引值为 v 的词语数, $n\_jtlv[v]$ 表示词语集 $X_j^{t^l}$ 中索引值为 v 的词语数, β^l 表示第 l 层主题分布参数.

(2) 采样 t_{ji}^l

在第 l 层的顾客-餐桌分配过程中, 索引变量 t_{ji}^l 对应分配参数 θ_{ji}^l, $A(x_{ji}^l,k_j^{t^l})$ 表示顾客选择已有餐桌的先验概率, $f_{k_j^{t^l}}^{-x_{ji}^l}(x_{ji}^l)$ 表示顾客的条件概率; 顾客选择新餐桌的先验概率为 α^l, 该顾客的条件概率结合公式 (12) 可表示为 $p(x_{ji}^l \mid t_{ji}^l=$

© 中国科学院软件研究所 http://www.jos.org.cn

t^l_{new}, $t^{l,-ji}$, k^l), 如公式 (15) 所示. t^l_{ji} 的后验概率如公式 (16) 所示.

$$p(x^l_{ji} \mid t^l_{ji}=t^l_{\text{new}}, t^{l,-ji}, k^l)=\sum_{k^l\in K^l}\frac{A^{k^l}_j}{\sum_{k^l\in K^l}A^{k^l}_j+\gamma^l}f^{-x^l_{ji}}_{k^l_j}(x^l_{ji})+\frac{\gamma^l}{\sum_{k^l\in K^l}A^{k^l}_j+\gamma^l}f^{-x^l_{ji}}_{k^l_{\text{new}}}(x^l_{ji}) \tag{15}$$

$$p(t^l_{ji}=t^l \mid t^{l,-ji}, k^l, X^l)\propto\begin{cases}A(x^l_{ji},k^{t^l}_j)f^{-x^l_{ji}}_{k^{t^l}_j}(x^l_{ji}), & t^l\in T^l_j\\ \alpha^l p(x^l_{ji}|t^l_{ji}=t^l_{\text{new}}, t^{l,-ji}, k^l), & t^l=t^l_{\text{new}}\end{cases} \tag{16}$$

当顾客选择就座于新餐桌时, 根据 nCRP+层次构造方法的餐桌-菜肴分配过程, 该餐桌分配到已有菜肴的概率与相同菜肴风格的餐厅中供应该已有菜肴 k^l 的餐桌数成正比, 或与餐桌中供应相同菜肴类别的该已有菜肴 k^l 的餐桌数成正比, 分配到新菜肴的概率与超参数 γ^l 成正比, 该新餐桌分配菜肴的概率如公式 (17) 所示.

$$p(k^{t^l_{\text{new}}}_j=k^l \mid t^l, k^{l,-t^l_{\text{new}}}_j)\propto\begin{cases}A^{k^l}_j f^{-x^l_{ji}}_{k^l}(x^l_{ji}), & k^l\in K^l\\ \gamma^l f^{-x^l_{ji}}_{k^l_{\text{new}}}(x^l_{ji}), & k^l=k^l_{\text{new}}\end{cases} \tag{17}$$

(3) 采样 $k^{t^l_j}$

索引变量 $k^{t^l_j}$ 对应第 l 层餐桌-菜肴分配参数 $\varphi^{t^l_j}$, 为了适应 t^l_{ji} 更新对菜肴分配的影响, 结合 $X^{t^l_j}$ 的条件概率, 计算 $k^{t^l_j}$ 的后验概率. 在第 l 层餐桌-菜肴分配过程中, 餐桌分配已有菜肴的先验概率为 $A^{k^l}_j$, $X^{t^l_j}$ 的条件概率表示为 $f^{-X^{t^l_j}}_{k^l}(X^{t^l_j})$; 餐桌分配新菜肴的先验概率为 γ^l, 此时 $X^{t^l_j}$ 的条件概率表示为 $f^{-X^{t^l_j}}_{k^l_{\text{new}}}(X^{t^l_j})$, $k^{t^l_j}$ 的后验概率如公式 (18) 所示.

$$p(k^{t^l_j}=k^l \mid t^l, k^{l,-t^l_j})\propto\begin{cases}A^{k^l}_j f^{-X^{t^l_j}}_{k^l}(X^{t^l_j}), & k^l\in K^l\\ \gamma^l f^{-X^{t^l_j}}_{k^l_{\text{new}}}(X^{t^l_j}), & k^l=k^l_{\text{new}}\end{cases} \tag{18}$$

3.6 领域主题层次自动挖掘算法

结合领域知识和层次化采样构建每层 CRF 中文档-主题 (或词语子集-主题) 和主题-词语分布参数 $\varphi^{t^l_j}$ 和 θ^l_{ji}, 利用这些参数自动构建领域主题层次和抽取领域主题词. 基于这种主题层次生成思想, 本文提出一种领域主题层次自动挖掘算法, 如算法 1 所示.

算法 1. 领域主题层次自动挖掘算法 *DomainTopicHierarchy*.

输入: 领域文档集 X, 模型各层级超参数 α^l、β^l 和 γ^l, 参数 λ 和语义相似度阈值 ξ, 领域类别及其代表性词语集 S, 主题层级 L;

输出: 领域主题层次及其主题词.

1. l=1;
2. WHILE (l<L) / * 采样第 l 层的索引变量 t^l_{ji} 和 $k^{t^l_j}$, 生成第 l 层的分布参数 θ^l_{ji} 和 $\varphi^{t^l_j}$ * /
3. FOR 每个文档或词语子集 $X^{t^l_j}\in X^{t^l}$ DO
4. FOR 每个词语 $x^l_{ji}\in X^{t^l_j}$ DO
5. 按公式 (8) 计算词语与主题的语义相似度 $A(x^l_{ji},k^{t^l_j})$, 明确顾客-餐桌分配先验;
6. 结合 $A(x^l_{ji},k^{t^l_j})$ 和公式 (16) 为 x^l_{ji} 分配餐桌, 索引为 t^l_{ji};
7. IF $t^l_{ji}=t^l_{\text{new}}$ THEN
8. 按公式 (11) 计算文档或词语子集 j 的领域隶属度 $A^{k^l}_j$, 明确餐桌-菜肴分配先验;
9. 结合 $A^{k^l}_j$ 和公式 (17) 为新餐桌分配菜肴, 索引为 $k^{t^l_{\text{new}}}_j$;
10. IF $k^{t^l_{\text{new}}}_j=k^l_{\text{new}}$ THEN
11. 结合 γ^l 和公式 (13) 为餐桌分配新菜肴;

© 中国科学院软件研究所 http://www.jos.org.cn

12. ELSE
13. 结合 $A_j^{k^l}$ 为餐桌分配已有菜肴;
14. END IF
15. ELSE
16. 为顾客分配已有餐桌;
17. END IF
18. END FOR
19. FOR 每张餐桌 $t_{ji}^l \in T_j^l$ DO
20. 结合公式 (14) 和公式 (18) 计算餐桌 t_{ji}^l 的菜肴分布参数, 索引为 $k_j^{t^l}$, 明确所有餐厅中餐桌的菜肴分布;
21. END FOR
22. 利用索引变量 t_{ji}^l 和 $k_j^{t^l}$ 生成第 l 层分布参数 θ_{ji}^l 和 $\varphi_j^{t^l}$;
23. 按公式 (9) 计算主题-词语贡献度, 生成主题-词语分布;
24. l=l+1;
25. END FOR
26. END WHILE

在给定领域类别和领域代表性词语的前提下, 通过领域隶属度、词语与领域主题的语义相关度和层次化的主题-词语贡献度, 描述领域知识对文档-主题 (或词语子集-主题) 和主题-词语概率分布影响, 结合算法 1, 本文给出了一种领域主题层次自动挖掘的通用框架.

4 实验分析

实验将从 2 个方面比较模型对领域主题层次的挖掘效果: 首先, 通过主题模型常用的评价标准和模型挖掘的主题在文本分类中的效果, 采用困惑度、复杂度、主题多样性和分类指标等 4 个评价指标进行定量评价; 然后, 通过构建的主题层次模型抽取主题词, 对模型挖掘效果进行定性评价.

4.1 评价指标

(1) 困惑度

困惑度可用于度量概率模型预测样本好坏的能力, 这里通过困惑度评估层次主题模型的表现. 通过训练数据集构建层次主题模型, 利用该模型生成测试数据集中每个数据的似然值, 通过这些似然值的对数定义困惑度 (perplexity, *perp*), 计算公式如式 (19) 所示. 困惑度越低说明层次主题模型的预测效果越好.

$$perp = \exp\left(-\frac{1}{|X_{\text{test}}|}\sum_{l\in L,k^l\in K^l}\sum_{x_{ji}^l\in X_{\text{test}},x_{ji}^l\in X^{k^l},j<|X_{\text{test}}|}\log(phi(x_{ji}^l|k^l)theta(k^l,j))\right) \tag{19}$$

其中, X_{test} 表示测试数据集, x_{ji}^l 表示其中的一个词语, X^{k^l} 表示主题索引值为 k^l 的词语集, *phi* 和 *theta* 分别表示基于训练数据的层次主题模型生成测试数据第 l 层主题索引值为 k^l 的词语概率分布和文档 (或词语子集) j 的主题概率分布, 通过 *phi* 和 *theta* 可以计算出测试数据集中每个词语的似然值.

(2) 复杂度

当主题模型的后验概率是通过 MCMC (Markov chain Monte Carlo) 采样方法推导生成时, 模型的复杂度 (complexity, *comp*) 被定义为层次主题模型每层级主题个数与所有文档中不同主题个数之和, 计算如公式 (20) 所示. 复杂度越低说明层次主题模型效果越好.

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

© 中国科学院软件研究所 http://www.jos.org.cn

其中, K^l 表示第 l 层的主题集合, T_j^l 表示文档 (或词语子集) j 在第 l 层中已分配的主题集合, k_j^l 和 k^l 表示主题在第 l 层餐桌-菜肴分配过程的索引值.

(3) 主题多样性

在主题模型中, 通过主题之间的 KL 散度值评价主题多样性, KL 散度值越小说明主题之间的差异性越小, 主题多样性越差; KL 散度值越大说明主题之间的差异性越大, 主题多样性越好. 在层次主题模型中, 计算父主题下任意一对子主题的 KL 散度值可分析这两个子主题之间的差异性, 同一父主题下所有子主题对的 KL 散度值则综合反映了父主题下子主题的多样性. 由于每个父节点主题下的子主题数是不相同的, 因此, 本文将层次主题模型的主题多样性定义为每个父节点主题下所有子主题对的 KL 散度值的平均值, 记为 *mean_KL*, 计算如公式 (21) 所示, *mean_KL* 值越大说明父节点主题下子主题之间的差异性越大, 子主题多样性越好.

$$mean\_KL=\sum_{i\neq j}\sum_{x_{ji}^l\in k_i^l\cap k_j^l} phi(x_{ji}^l|k_i^l)\log\frac{phi(x_{ji}^l|k_i^l)}{phi(x_{ji}^l|k_j^l)}\Bigg/\sum_{i\neq j}1 \tag{21}$$

其中, k_i^l 和 k_j^l 表示隶属于同一父节点主题的、第 l 层任意两个主题的索引值, *phi* 表示主题-词语概率分布.

(4) 分类指标

将层次主题模型生成的文档-主题分布作为逻辑回归分类器的输入, 采用分类度量指标, 精确率 (precision, P)、召回率 (recall, R) 和 $F1$ 值 ($F1$) 评估模型在文档子领域分类中的表现. 计算如公式 (22) 所示, $F1$ 值越大, 说明模型整体分类效果越好.

$$P=\frac{|M\cap N|}{|M|},\ R=\frac{|M\cap N|}{|N|},\ F1=\frac{P\times R\times 2}{P+R} \tag{22}$$

其中, M 表示被预测为属于该子领域类别的文档构成的集合, N 表示被标注为属于该子领域类别的文档构成的集合, $M\cap N$ 表示同时被标注和预测为属于该子领域类别的文档构成的集合.

4.2 数据集和领域知识

由于官方微博相较于个人微博或其他非官方微博有较高的权威性、代表性, 且更正式、可信度更高, 因此本文选择财经网微博 (https://weibo.com/caijing) 文本为实验数据, 时间跨度为 6 年. 已有研究从财经微博文本中提取描述一级经济指标的领域主题及其主题词[12], 本实验将从中提取经济指标对应的领域主题层次及其主题词, 并通过一个公开的英文数据集 20NewsGroup (http://qwone.com/~jason/20Newsgroups/) 对比验证模型的有效性.

在数据预处理方面, 首先, 通过百度自然语言处理工具 (https://ai.baidu.com/tech/nlp_basic) 对财经微博文本进行分词和词性标注; 然后, 通过词性分析、频数统计、去除低频词 (出现次数低于 10 的词语) 和高频词 (出现次数高于 7000 的词语) 等方法实现数据预处理; 对于 20NewsGroup 文本, 通过 NLTK 工具包 (https://www.nltk.org/) 实现预处理, 并将文档中低频词 (出现次数低于 10 的词语) 或高频词 (出现次数高于 5000 的词语) 删除. 由于财经微博文本中描述经济指标的代表性词语大多表现为名词词性, 本文挖掘的主题词是经济要素词, 预处理后只保留词性标注为普通名词、动名词和专有名词的词语. 实验数据的详细描述如表 2 所示, 其中, 财经微博文本中的所有词语、普通名词、动名词和专有名词分别列出计重复和不计重复的词语数量; 20NewsGroup 只列出计重复和不计重复的词语数量, 不区分词语词性.

表 2 数据详细描述

数据集名称	文本数量	每篇文档平均字数	词语总量	普通名词数量	动名词数量	专有名词数量	一级指标/领域代表性词语数量	二级指标/领域代表性词语数量
财经微博文本	92 747	88	8 167 165 267 722	1 159 093 47 185	352 961 12 574	173 635 35 454	85	873
20NewsGroup	11 296	105.7	53 619 30 222	— —	— —	— —	70	1 173

© 中国科学院软件研究所 http://www.jos.org.cn

在领域知识方面, 领域类别信息包括官方定义的领域类别, 以及预处理后各个领域的代表性词语. 针对财经微博文本, 文献 [1,12] 给出了 5 个一级经济指标, 包括投资、进出口、消费、政府财政和人口就业, 及其代表性词语; 本文根据国家统计局官网 (http://www.stats.gov.cn/) 对经济子领域的划分, 选择了 22 个二级经济指标, 包括人口、固定资产投资 (侧重房地产方面)、对外经贸、能源、财政、人民生活、城市基础设施、资源环境、农业、工业、建筑业、运输邮电、信息技术、批发零售、旅游、金融业、教育、科技、医药卫生、社会服务、文化体育、公共管理. 结合官网定义的领域代表性词语, 通过词语之间的点互信息 (pointwise mutual information, PMI) 确定了各个领域的代表性词语, 如附录 A 所示.

20NewsGroup 数据集划分了 7 个一级领域, 包括 alt (无神论方面的主题)、comp (计算机软硬件方面的主题)、misc (无法明确分类的其余主题, 这里主要是与二手市场有关的主题)、rec (recreation, 休闲娱乐方面的主题)、sci (科学研究与应用方面的主题)、soc (社会科学方面的主题) 和 talk (辩论或人们长期争辩的主题); 同时也定义了 15 个二级领域, 包括 atheism (无神论)、graphics (图形)、os (操作系统)、sys (系统, 偏硬件)、windows (系统, 偏软件)、forsale (二手交易)、motorcycles (摩托车)、auto (汽车)、sport (运动)、crypt (密码学)、electronics (电子学)、med (医学)、space (太空学)、religion (宗教信念) 和 politics (政治). 由于英文数据集中没有专家给定的代表性词语, 本文根据这些领域名称与文本中其他词语的 PMI 值选择对应的领域代表性词语, 如附录 A 所示. 由于领域数据集只包括领域和子领域的定义, 因此, 实验只分析具有两层主题方面共享的主题层次.

4.3 对比模型与参数设置

(1) 对比模型

本文的层次主题模型主要考虑了两个方面: ① nCRP+层次构造方法是在 nCRP 的基础上考虑每个层级内的主题方面共享, 改变了主题树形结构的先验分布, 构建 rHDP 层次主题模型; ② 结合领域知识的层次主题模型改进了每层级主题和主题词的分配过程. 因此, 实验内容需要验证两个方面: ① 对比基于 nCRP 和神经网络的层次主题模型, 基于 nCRP+的层次主题模型对主题层次构建的影响; ② 对比各方面领域知识对领域主题层次构建和主题词抽取的影响.

本文的对比模型包括以下 9 种.

① hLDA[2]: 基于 nCRP 构造方法和 LDA 的层次主题模型.

② nHDP[3]: 基于 nCRP 构造方法和 HDP 的层次主题模型.

③ TSNTM[9]: 基于自编码变分贝叶斯和双递归神经网络的主题层次构建模型.

④ WEDTM[7]: 基于多层信念网络、泊松因子分析和词向量的主题层次构建模型.

⑤ rHDP: 基于 nCRP+构造方法和 HDP 的层次主题模型.

⑥ rHDP_1: 单独考虑了领域隶属度的 rHDP 模型.

⑦ rHDP_2: 单独考虑了词语-主题语义相关度的 rHDP 模型.

⑧ rHDP_3: 单独考虑了层次化主题-词语贡献度的 rHDP 模型.

⑨ rHDP_DK: 综合考虑了领域隶属度、词语-主题语义相关度和层次化主题-词语贡献度的 rHDP 模型.

(2) 参数设置

在层次主题模型中, CRF 分配过程中的下层 CRP 超参数 α 值越小, 文档中的主题数越少; 上层 CRP 超参数 γ 值越大, 生成新主题的概率越大. 主题-词语分布超参数 β 值将会影响词语的主题分布, 其值越小, 词语属于一个主题的概率越大. 通过分析和调试模型生成的主题及其词语分布, 本文将层次主题模型 rHDP、rHDP_1、rHDP_2、rHDP_3 以及 rHDP_DK 中上层 CRF 分配过程的参数 α、β 和 γ 分别设置为 1.5、0.5 和 0.01, 下层 CRF 对应参数分别设置 1.5、0.3 和 0.1. 词语之间的语义相似性用于计算待分配词语与该词语所在层级主题中词语的语义相似性, 便于进一步分析词语与该主题的语义相关度, 用于改善主题-词语分配过程. 语义相似度阈值 ξ 将影响模型中主题的词语分布, 阈值越高, 主题中词语的语义更接近, 影响了主题中词语的多样性. 考虑到领域主题涵义的明确性和每层级主题词的代表性, 同时兼顾主题中词语的多样性, 经调试, 上层和下层 CRF 分配过程中的语义相似性

© 中国科学院软件研究所 http://www.jos.org.cn

阈值 ξ 分别设置为 0.3 和 0.4.

为了提高 rHDP_1 和 rHDP_DK 层次主题模型中各层次领域隶属度的计算效率, 在计算词语的领域隶属索引值时, 只对文档中出现次数在上述规定范围内的词语计算基于 BERT 的语义相似度. 因此, 在计算文档 (或词语子集) 的领域隶属索引值时, 如果文档 (或词语子集) 中只包含高频词或超低频词, 该文档 (或词语子集) 的领域隶属索引值将设为二级领域数量+1, 区别于已有的领域 (或子领域) 类别索引值.

在 rHDP_3 和 rHDP_DK 层次主题模型中, 层次化的主题-词语贡献度不仅反映了每层级词语对领域主题的代表性, 而且体现了关联子主题中主题词的差异性. 根节点和第 1 层主题中词语的代表性主要体现为对领域和子领域的概括性, 计算过程中通过词语在对应主题中的概率值及其词语集中的频次共同决定. 第 2 层主题中词语的代表性需要同时反映该词语所在分支中的领域主题和子主题的涵义, 凸显关联子主题的差异性. 因此, 在生成领域子主题的主题词时综合了两个方面的代表性, 一方面通过词语与该领域子主题的逆主题频率和其在主题中的概率值, 描述词语对领域子主题的代表性; 另一方面, 通过词语与其所在分支领域主题的语义相关性描述词语对领域主题的代表性; 参数 λ 为这两部分的权值. 为了凸显关联子主题在主题词方面的差异性, 第 2 层主题中的主题词需更侧重其对子主题涵义的描述, 在设置词语对领域子主题的代表性和词语与领域主题的语义相关性之间的比例时, 前者的权值更大. 通过分析和调试主题中词语的贡献度变化情况, 将 λ 值设置为 0.6. 对于困惑度的计算, 实验中训练数据集和测试数据集的设置统一参照文献 [9].

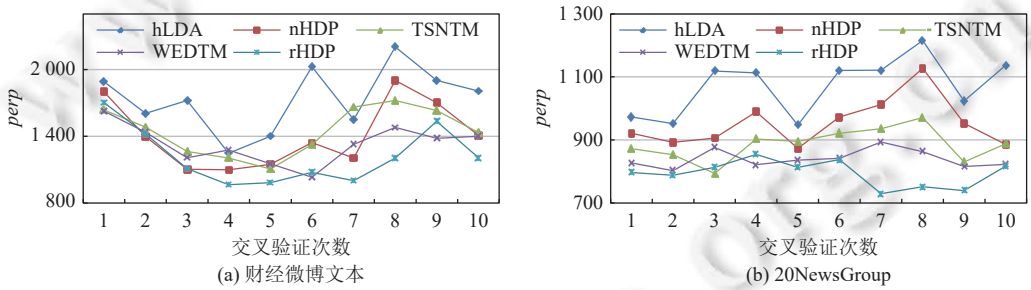
实验采用的分类模型是逻辑回归分类器. 由于 20NewsGroup 数据集中每篇文档都有子领域划分, 因此实验中基于这些子领域划分进行文档类别标注, 共分为 15 个类别. 按照 3:1 的比例将每个类别的文档集分成训练数据集和测试数据集. 利用模型生成的文档-主题概率分布对文档进行特征描述, 作为分类器的输入向量.

4.4 对比实验结果与分析

对比实验结果分析包括前 5 个对比模型在 2 个数据集和各评价指标上的定量分析.

(1) 困惑度

5 个对比模型在 2 个数据集的困惑度如图 5(a) 和图 5(b) 所示.



(a) 财经微博文本　　(b) 20NewsGroup

图 5　对比模型的困惑度

在图 5 中, 横坐标表示十折交叉验证方法中的验证次数, 纵坐标表示每次验证生成的困惑度. 从图 5 可看出, rHDP 层次主题模型的困惑度总体上低于 hLDA 和 nHDP 模型, 说明基于 nCRP+的层次主题模型的预测能力比基于 nCRP 的层次主题模型要好一些; rHDP 层次主题模型的困惑度也总体上低于 TSNTM 和 WEDTM 模型, 说明基于 nCRP+的层次主题模型在预测方面比基于神经网络的构造模型要好一些. 导致 rHDP 层次主题模型困惑度较低的主要原因是, 通过分层主题方面共享, 模型能更好地明确领域文档和关联子主题中词语的主题属性, 进而提高词语的生成概率, 降低词语的不确定程度.

(2) 复杂度

由于 nHDP 层次主题模型的采样方法是变分推导, TSNTM 模型采用变分自动编码方法, 本文的复杂度以 MCMC 采样方法为计算依据, 所以, nHDP 和 TSNTM 模型不参与模型复杂度比较. 其他 3 个模型在 2 个数据集的复杂度如图 6(a) 和图 6(b) 所示, 横坐标表示模型的迭代次数, 纵坐标表示每次迭代模型的复杂度. 由于第 1 层、

© 中国科学院软件研究所 http://www.jos.org.cn

第 2 层节点的迭代过程是顺序执行的, 而在迭代生成第 1 层节点的过程中, 每篇文档的主题分布是无法确定的, 因此在比较模型复杂度时, 模型的迭代次数描述的是每个模型第 2 层节点生成的迭代过程, 每次迭代的复杂度也是第 2 层各分支节点在该迭代次数中复杂度的平均值.

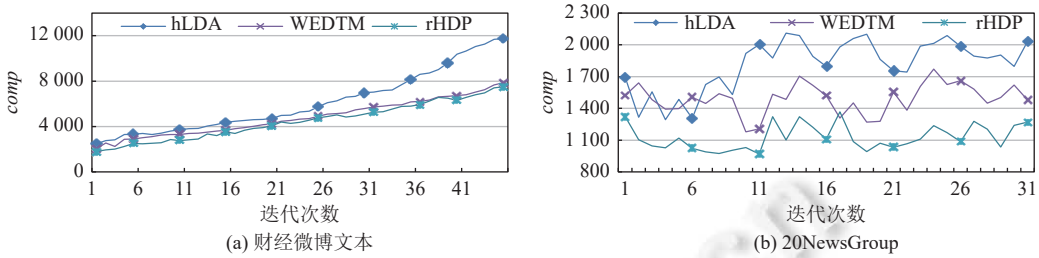


图 6 对比模型的复杂度

从图 6 可看出, 与 hLDA 模型相比较, rHDP 模型的复杂度要低很多; 主要原因是基于 nCRP+方法的层次主题模型生成的低层级 ($l \geq 2$) 主题数目低于 hLDA 模型, 使其在模型复杂度方面表现较好; 这在一定程度上反映出, 相较于 hLDA 模型, rHDP 模型生成的主题更加简洁. 与 WEDTM 模型相比较, rHDP 模型的复杂度低一些, 主要原因是 rHDP 模型中的文档主题分布相较于 WEDTM 模型要简单一些.

(3) 主题多样性

5 个对比模型在 2 个数据集中高层级 ($l \leq 1$) 节点的 (子) 主题多样性如图 7(a) 和图 7(b) 所示, 横坐标表示节点编号, 纵坐标是该节点的 (子) 主题多样性 *mean_KL* 值.

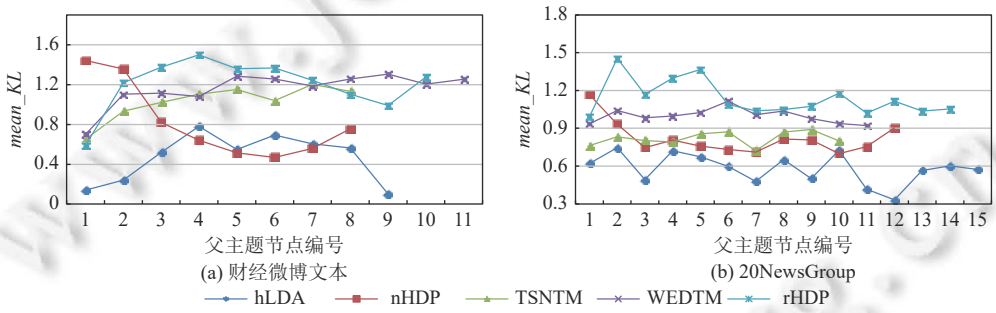


图 7 对比模型的主题多样性

在图 7 中, 节点编号为 1 表示第 0 层节点即根节点, 其余表示第 1 层节点. 由于 hLDA、nHDP、TSNTM、WEDTM 和 rHDP 模型在财经微博文本数据集中生成的第 1 层节点数分别是 8、7、7、10 和 9, 因此 5 个模型中第 1 层节点的编号取值范围分别为 [2, 9]、[2, 8]、[2, 8]、[2, 11] 和 [2, 10]; 这 5 个模型在 20NewsGroup 数据集中生成的第 1 层节点数分别是 14、11、9、10 和 13, 因此各模型中第 1 层节点的节点编号取值范围分别为 [2, 15]、[2, 12]、[2, 10] 、[2, 11] 和 [2, 14].

从图 7 可看出, rHDP 模型的 *mean_KL* 值整体上都要显著高于 hLDA 模型, 说明基于 nCRP+的层次构造方法显著丰富了子主题的涵义; 对于根节点 (节点编号为 1), nHDP 模型的 *mean_KL* 值比其他模型都要大, 说明 nHDP 模型根节点的 (子) 主题多样性较好; 但对于第 1 层节点, rHDP 模型的 *mean_KL* 值基本上都显著大于 nHDP 模型 (只有财经微博文本数据集中节点编号为 2 的节点例外), 说明 rHDP 模型第 1 层节点的 (子) 主题多样性整体上显著好于 nHDP 模型. rHDP、TSNTM 和 WEDTM 模型在根节点的 *mean_KL* 值相差不大, 说明 3 个模型在根节点的主题多样性都较好, 但 rHDP 模型第 1 层节点的主题多样性总体上要好于 TSNTM 和 WEDTM 模型.

(4) 分类效果

针对模型对 20NewsGroup 数据集的分类效果进行评价. 通过 5 个主题层次模型生成训练集中文档的主题概

© 中国科学院软件研究所 http://www.jos.org.cn

率, 作为分类器的特征向量. 限于篇幅, 实验中只给出了 graphics 和 crypt 类别的分类结果, 如表 3 所示.

从表 3 可发现, rHDP 模型的分类指标值总体上均高于 hLDA、nHDP、TSNTM 和 WEDTM 模型, 主要原因是 rHDP 模型通过分层次的主题共享生成主题分布, 拓宽了每个领域下子领域的范围, 模型将使用更加丰富的主题特征, 提高文档类别预测的效果. 其中, nHDP 模型的 R 值比 rHDP 模型高, 说明多分支的主题分布可以明确更多文档的类别.

表 3 对比模型的分类指标值

模型	graphics			crypt		
	P	R	$F1$	P	R	$F1$
hLDA	0.7761	0.8272	0.8008	0.9102	0.9395	0.9246
nHDP	0.8023	0.8683	0.8340	0.9336	0.9637	0.9484
TSNTM	0.8294	0.8601	0.8444	0.9393	0.9355	0.9374
WEDTM	0.8322	0.8621	0.8469	0.9403	0.9478	0.9440
rHDP	0.8502	0.8642	0.8571	0.9442	0.9556	0.9499

4.5 领域知识实验结果与分析

领域知识实验结果分析包括 rHDP 模型及其 4 个改进模型在 2 个数据集和各评价指标上的定量分析, 以及 rHDP_DK 模型生成的领域主题中主题词分布的定性分析.

(1) 定量分析

模型的定量评价指标包括困惑度、复杂度、主题多样性和分类指标, 其中 5 个层次主题模型在 2 个数据集的困惑度和复杂度分别如图 8 和图 9 所示.

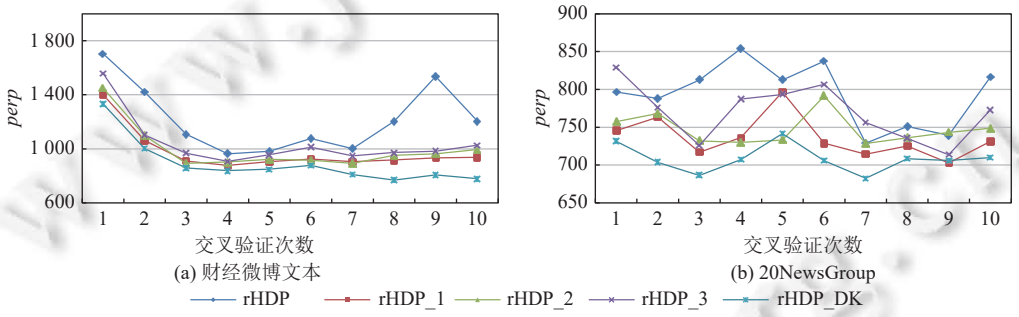


图 8 层次主题模型的困惑度

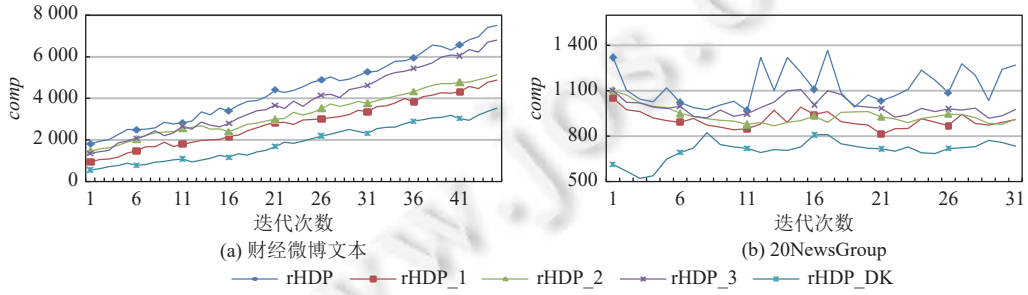


图 9 层次主题模型的复杂度

从图 8 可看出, rHDP_1、rHDP_2、rHDP_3 和 rHDP_DK 模型的困惑度总体上低于 rHDP 模型, 说明增加领域隶属度、词语与主题的语义相关度、层次化主题-词语贡献度可以改善模型在数据预测方面的效果. 其中, rHDP_1 和 rHDP_2 模型的困惑度低于 rHDP_3 模型, 说明领域隶属度和语义信息在模型的数据预测方面的效果

© 中国科学院软件研究所 http://www.jos.org.cn

要好于层次化主题-词语贡献度; rHDP_DK 模型的困惑度总体上低于其他 4 个模型, 说明同时在每层级文档-主题和主题-词语分配过程中添加领域知识的引导, 可以较好地改善层次主题模型的整体预测力.

从图 9 可看出, rHDP_1、rHDP_2、rHDP_3 和 rHDP_DK 模型的复杂度均低于 rHDP 模型, 说明增加领域隶属度、词语与主题的语义相关度和层次化主题-词语贡献度可以改善模型的复杂度. 其中, rHDP_1、rHDP_2 模型的复杂度低于 rHDP_3 模型, 说明领域隶属度和词语与主题的语义相关度对改善模型复杂度的效果要好于层次化主题-词语贡献度; rHDP_DK 模型的复杂度低于其他 4 个模型, 说明通过领域知识改进各层级文档-主题和主题-词语分配过程可以较好地降低层次主题模型的复杂度.

5 个层次主题模型在 2 个数据集中高层级 ($l \leqslant 1$) 节点的 (子) 主题多样性如图 10(a) 和图 10(b) 所示. 其中, rHDP_1、rHDP_2、rHDP_3 和 rHDP_DK 层次主题模型在财经微博文本数据集中生成的第 1 层节点数分别是 7、9、8 和 7, 因此 4 个模型中第 1 层节点的编号取值范围分别为 [2, 8]、[2, 10]、[2, 9] 和 [2, 8]; 这 4 个层次主题模型在 20NewsGroup 数据集中生成的第 1 层节点数分别是 15、17、16 和 14, 因此各模型中第 1 层节点的节点编号取值范围分别为 [2, 16]、[2, 18]、[2, 17] 和 [2, 15].

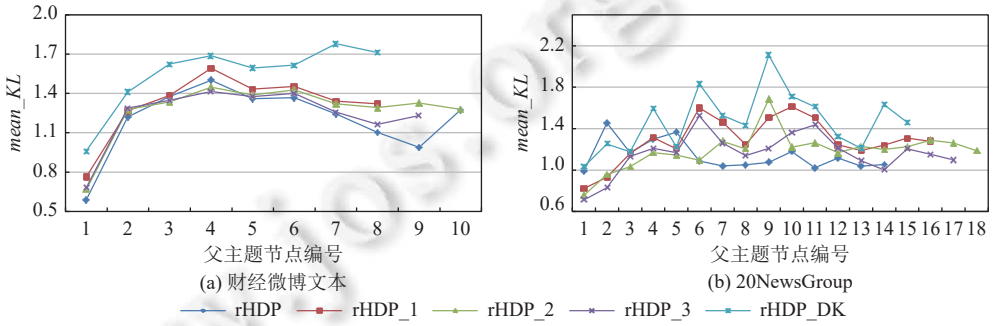


图 10 层次主题模型的主题多样性

从图 10 可看出, rHDP_3 模型的 *mean_KL* 值比 rHDP 模型整体上要高一些, 说明增加词语在主题中的贡献度可以明确词语对主题的代表性, 进而明确了子主题在领域涵义方面的差异性, 改善了子主题的多样性. 相较于 rHDP_3 模型, rHDP_1、rHDP_2 模型的 *mean_KL* 值高一些, 说明主题的领域特性和词语的语义信息在改善子主题多样性方面具有更好的效果. 主要原因是: 一方面, 在层次主题模型中引入领域 (或子领域) 类别特性, 对应于菜肴风格 (或菜肴类别) 概念, 并通过领域 (或子领域) 类别改进各层级文档-主题分配过程, 实现将主题 (或子主题) 涵义映射到领域 (或子领域) 类别, 明确的领域 (或子领域) 类别信息将丰富主题 (或子主题) 的领域涵义; 另一方面, 在分层次的主题共享机制下通过词语与主题的语义信息改进主题词的聚类过程, 明确主题涵义, 进而改善主题下子主题的多样性. 在 5 个层次主题模型中, rHDP_DK 模型在子主题多样性方面表现最好, 说明在考虑了 3 方面领域知识之后, 不仅从全局上明确了主题的领域涵义, 而且从局部上明确了词语在主题中的代表性, 凸显了关联子主题之间的领域差异性, 间接提高了主题多样性.

5 个层次主题模型在 graphics 和 crypt 中的分类结果, 如表 4 所示.

表 4 层次主题模型的分类指标值

模型	graphics			crypt		
	P	*R*	*F*1	*P*	*R*	*F*1
rHDP	0.8502	0.8642	0.8571	0.9442	0.9556	0.9499
rHDP_1	0.8740	0.8848	0.8793	0.9522	0.9637	0.9579
rHDP_2	0.8699	0.8807	0.8753	0.9555	0.9516	0.9535
rHDP_3	0.8785	0.8930	0.8857	0.9676	0.9637	0.9657
rHDP_DK	**0.8871**	**0.9053**	**0.8961**	**0.9797**	**0.9718**	**0.9757**

© 中国科学院软件研究所 http://www.jos.org.cn

从表 4 可发现, rHDP_1、rHDP_2 和 rHDP_3 模型的分类指标值比 rHDP 模型高, 说明增加领域知识对模型的分类效果有帮助. 其中, 层次化主题-词语贡献度明确了关联子主题的区分度, 提高了分类效果; 相较于语义相关度, 领域隶属度也有较好的预测效果. rHDP_DK 模型的分类指标值是最好的, 主要原因是通过领域知识改进了主题的领域特征, 明确了文档的领域类别.

(2) 定性分析

以财经文本为例, 分析基于 rHDP_DK 模型的、具有两层主题方面共享的主题层次. 结合词语集 W 的领域隶属索引值, 如式 (6) 所示, 通过计算主题词与经济指标中代表性词语之间的语义相似度, 分别计算主题与一级指标/领域、子主题与二级指标/领域之间的语义相关性, 将主题、子主题分别映射到语义最相关的一级指标/领域、二级指标/领域, 主题分布结果如表 5 所示.

表 5 一级经济指标、二级经济指标分别与主题、子主题的对应情况

二级经济指标	一级经济指标				
	投资 (topic1)	进出口 (topic3)	政府财政 (topic6)	消费 (topic5)	人口与就业 (topic4)
人口	subtopic1	subtopic2	subtopic10	subtopic8	subtopic11
固定资产投资	—	—	subtopic16	subtopic16	subtopic12
对外经贸	subtopic11	subtopic1	—	—	subtopic20
能源	subtopic4	subtopic4	subtopic6	subtopic10	subtopic3
财政	subtopic3	subtopic12	subtopic3	subtopic21	subtopic9
人民生活	subtopic8	subtopic9	subtopic4	subtopic2	subtopic6
城市基础设施	subtopic19	subtopic19	subtopic19	subtopic3	subtopic22
资源环境	subtopic18	subtopic16	subtopic12	subtopic17	subtopic2
农业	subtopic17	subtopic20	subtopic21	subtopic12	subtopic24
工业	subtopic9	subtopic8	subtopic5	subtopic1	subtopic8
建筑业	—	—	—	—	—
运输邮电	subtopic14	subtopic13	subtopic7	subtopic14	subtopic18
信息技术	subtopic13	subtopic3	subtopic14	subtopic11	subtopic16
批发零售	—	—	subtopic18	subtopic22	—
旅游	subtopic23	subtopic15	subtopic17	subtopic7	subtopic19
金融业	subtopic6	subtopic10	subtopic13	subtopic13	subtopic1
教育	subtopic10	subtopic6	subtopic1	subtopic4	subtopic5
科技	—	subtopic5	subtopic9	subtopic6	subtopic7
医药卫生	subtopic21	subtopic18	subtopic8	subtopic9	subtopic13
社会服务	subtopic20	subtopic21	subtopic20	subtopic20	subtopic14
文化体育	subtopic22	subtopic11	subtopic15	subtopic19	subtopic10
公共管理	subtopic7	subtopic7	subtopic11	subtopic5	subtopic4

表 5 中, 表头括号中的内容表示一级经济指标在第 1 层主题中对应的主题编号, 如一级经济指标“投资”对应第 1 层主题中的第 1 个主题; 表体中的每一行表示该二级经济指标分别在哪些一级经济指标所对应的主题下生成了子主题, 分量为对应主题下的子主题编号, 如二级经济指标“人口”分别在一级经济指标“投资”“进出口”“政府财政”“消费”“人口与就业”下生成了相互关联的 (共享的) 子主题, 在对应主题中的子主题编号分别为 1、2、10、8 和 11.

每个主题下与二级指标/领域语义相关的子主题分布的统计结果如表 6 和表 7 所示. 其中, 对于每个主题, 在第 1 行中有两个值“sb_1/sb_2”, 分别表示该主题下在合并前/合并后与对应二级指标/领域语义相关的子主题数量; 在第 3 行中是语义相关子主题覆盖率, 表示合并前的语义相关子主题数与第 2 行中的所有子主题数的比值.

© 中国科学院软件研究所 http://www.jos.org.cn

表 6 每个主题下与二级经济指标语义相关的子主题分布 (财经微博文本)

主题 (一级经济指标)	投资 (topic1)	进出口 (topic3)	政府财政 (topic6)	消费 (topic5)	人口与就业 (topic4)
语义相关的子主题数	23/18	21/19	20/20	21/20	22/20
全部子主题数	24	22	20	23	22
语义相关子主题覆盖率 (%)	95.8	95.5	100.0	91.3	100.0

表 7 每个主题下与二级经济指标语义相关的子主题分布 (20NewsGroup)

主题 (一级领域)	alt (topic1)	comp (topic2)	misc (topic3)	rec (topic4)	sci (topic5)	soc (topic7)	talk (topic8)
语义相关的子主题数	14/4	15/8	14/9	16/7	15/12	16/4	15/7
全部子主题数	14	15	14	18	16	16	16
语义相关子主题覆盖率 (%)	100.0	100.0	100.0	88.9	93.8	100.0	93.8

进一步说明:

① 表 5 中会出现子主题编号超出子主题总数的现象 (各个主题下子主题总数如表 6 第 2 行所示), 原因是某些子主题中的主题词概率值太低, 导致该子主题中词语分布为空. 例如, 在实验结果中主题 topic4 下的子主题最大编号为 24, 但子主题 21、23 为空, 因此有效的子主题总数仅为 22 个. 另外, 从国家统计局官网中选择的“建筑业”代表性词语较少; 与此同时, 实验数据集中关于这个领域的文本也相对较少, 导致实验结果中描述该二级经济指标的对应子主题为空.

② 根据第 3.1 节中的领域隶属索引值计算方法可知, 主题 (或子主题) 的领域隶属结果可能存在 3 种情况.

i. 每个主题 (或子主题) 隶属于不同的一级经济指标 (或二级经济指标);

ii. 多个主题 (或子主题) 隶属于同一个一级经济指标 (或二级经济指标);

iii. 主题 (或子主题) 的领域隶属索引值不在已知的一级经济指标 (或二级经济指标) 索引范围内, 即该主题 (或子主题) 不能映射到任何一个一级经济指标 (或二级经济指标).

③ 在实验结果中, 主题 (或子主题) 的领域隶属结果主要以第 i 种情况出现. 第 ii 种情况只在第 2 层子主题的领域隶属分布中出现, 此时我们采取子主题合并的方法进行处理, 即将领域隶属于同一个二级经济指标的多个子主题合并成一个子主题, 其中, 主题 topic1 下合并后的子主题有 subtopic1 (子主题 1 和 2 合并)、subtopic11 (子主题 11 和 15 合并)、subtopic4 (子主题 4 和 5 合并)、subtopic13 (子主题 13 和 16 合并) 和 subtopic10 (子主题 10 和 12 合并); 主题 topic3 下合并后的子主题是 subtopic7 (子主题 7、14 和 17 合并); 主题 topic4 下合并后子主题是 subtopic8 (子主题 8、15 和 17 合并); 主题 topic5 下合并后的子主题是 subtopic6 (子主题 6 和 18 合并).

④ 第 iii 种情况仅出现在 topic1、topic3 和 topic5 中, 包括 topic1 中的 subtopic24、topic3 中的 subtopic22、topic5 中的 subtopic15 和 subtopic23.

⑤ 在表 7 中, 由于 alt、soc 和 talk 与其他 4 个领域的领域涵义差异性较大, 所以它们之间的子领域交叉性较弱.

根据表 6 和表 7 中主题与领域中词语之间的语义相关分析结果可发现, rHDP_DK 层次主题模型生成的子主题可以较全面地反映二级领域主题的领域涵义.

为了更直观地呈现层次主题模型中关联子主题在不同父主题中共享主题方面的有效性, 以及关联子主题之间的差异性, 根据 rHDP_DK 模型生成的主题词语分布, 抽取出反映不同经济领域主题下关联子主题在“对外经贸”和“人民生活”方面的主题-词语分布, 结果分别如表 8 和表 9 所示.

从表 8 中发现, “进出口”领域中的“对外经贸”子领域主要描述我国进出口贸易的发展方面, 如反倾销调查、中俄原油管道、服务出口、丝绸之路经济带、一带一路等方面; “政府领域”中的“对外经贸”子领域主要体现为政府在对外经贸方面的税收政策, 如空置税、零关税、增值税暂行条例等方面; “消费”领域中的“对外经贸”子领域主要体现人们的外贸行为及其涉及的产品, 如进口博览会、中国茶、药品等方面.

从表 9 中发现, “投资”领域中的“人民生活”子领域主要描述与人民生活相关方面的投资, 如纪念币、茅台酒、纪念钞、理财产品等方面; “进出口”领域中的“人民生活”子领域主要描述人们日常生活中的进出口产品的词语,

© 中国科学院软件研究所 http://www.jos.org.cn

如咖啡、IQOS (烟草巨头菲利普莫里斯公司生产的加热式电子烟)、大切诺基 (进口 Jeep) 等方面; “消费”领域中的“人民生活”子领域主要描述人们日常生活中的消费产品, 如牛肉、龙虾、扇贝等方面.

表 8 共享“对外经贸”方面的子主题的主题词

进出口	政府财政	消费
离岸人民币; 中间价; 反补贴; 报关单; 反倾销调查; 中俄原油管道; 服务出口; CNH; CNY; 直销; 寡头; 纸币; 离岸; 降准; 贸易帐; 自由化; 免税品; 资源性产品; 博览会; 集成电路产业; 先行指数; 出口订单指数; FDI; 周一欧市; 免税购物额; 货币错配; 应税; 衰退性顺差; 丝绸之路经济带; 自由贸易园区; 一带一路; 滑准税; 税目; 自由贸易区; 贸易顺差	空置税; 零关税; 免税店; 计税毛利率; 增值税暂行条例; 地方主体税种; 成品油消费税; 中韩; 包工; 反垄断; 经营权; 要债; 热潮; 回暖; 关税; 贸易战; 单边主义; 贸易保护主义; 乐视; 税收; 关口; SDR(特别提款权); 免税款; 无纸化; 政策性金融机构; 关税配额管理	外贸; 进口博览会; 中国茶; 阿斯匹林; 马息岭; 美元; 护照; 海关; 药; 茶叶; 药品; 假药; 假冒; 价格; 假货

表 9 共享“人民生活”方面的子主题的主题词

投资	进出口	消费
纪念币; 茅台酒; 老酒; 金块; 赌球; 石油币; 高铁币; 航天纪念钞; 投资理财产品; 月租; 金银币; B站; 充电; 供暖; 单身公寓; 四万亿计划; 募资; 快递公司; 购物中心; 主题公园; 纪念邮票; 证券投资a者保护基金; 乐视汽车; 克强指数; 菜鸟驿站; 虾米音乐; 收藏行为; 航天纪念币纪念钞; 菜篮子; 迅雷看看电影院; 冤枉钱; 金银纪念币; 采暖; 新能源车	咖啡; 中成药; 民族主义; 乌木; 广交会; 啤酒; 老干妈; 楚菜; 导盲犬; IQOS; 大切诺基; 中药; 房车; 黄唇鱼; 四国; 赤霞珠; 山寨食品; 排他性条款; 美国葡萄酒; 路虎神行者2; 国产版; 雷克萨斯; 可乐; 揽胜; 原研药; 星巴克; 平行进口车; 甜味剂; Coco Cola Life; Pepsi; 金丝楠阴沉木; 费列罗巧克力; 智利红酒; 护肤品; 纸尿裤; 真空保温杯; 太阳镜	牛肉; 龙虾; 扇贝; 大闸蟹; 死猪; 宠物; 床单; 泡面; 冷冻; 奶茶; 狗肉; 兔子; 西红柿; 蛋壳; 车位; 餐具; 饲料; 灭火器; 虹鳟; 三文鱼; 大米; CPI; 胚胎; 洋奶粉; 番茄炒蛋; 象牙; 盒饭; 蜗牛; 恩格尔系数; 松鼠; 女装; 生蚝; 鲤鱼; 散热片; 酸奶; 肉类; 汉堡; 白领; 枪手; 鸭子; 导盲犬; 猫咪危房; 鸭肉; 炒饭; 鸟类; 炸鸡

在 20NewsGroup 数据中抽取出反映不同领域主题下的关联子主题在“sys”方面的主题-词语分布, 结果如表 10 所示.

表 10 共享“sys”方面的子主题的主题词

comp	rec	sci
stylewriter; selling; macintosh; canon; uhf; cellular; mobile; equip; surround; eprom; dpl; micor; camcorder; delivered; converter; deskjet; license; combo; receiver; amd; usm; scanjet; prologic; adaptec	det; buf; tor; har; min; chus; ott; edm; nyy; mil; matchup; transmit; swap; countersteer; scma; listserv; moog; replica; lojack; eeprom; westfield; zeppelin	rfus; odometer; starter; hood; hump; powering; efficiency; parachute; faa; runway; qualcomm; cdma; mileage; pace; suing; panda; communicant; incrementally; personel; ebcdic; rado; thermic; gasturbine; thermoelement; gaz; capacity; modulation; imprement

从表 10 中发现, “comp (计算机软硬件)”中的“sys (系统, 偏硬件)”主要描述计算机硬件, 如 macintosh (苹果电脑)、eprom (只读存储器)、camcorder (摄影机)、deskjet (打印机)、combo (光驱)、amd (主板) 等方面; “rec (休闲娱乐)”中的“sys”主要描述与汽车系统相关的硬件, 如 det (结构设计鉴定)、ott (互联网电视)、mil (故障指示灯)、listserv (服务器)、moog (音响合成器)、eeprom (电可擦除只读存储器) 等方面; “sci (科学研究方面)”中的“sys”主要描述航天系统中的硬件, 如 rfus (光度计上检测荧光强度)、powering (动力估计)、qualcomm (高通公司, 通信技术)、cdma (电信版)、gasturbine (燃气轮机)、thermoelement (热元件).

实验结果表明, 结合领域知识和 nCRP+层次构造方法的 rHDP_DK 层次主题模型可以有效地挖掘出不同主题下的关联子主题, 明确关联主题中主题词的差异性.

5 总 结

领域主题之间不仅包含纵向的领域层次关系而且还包括横向的子领域方面共享关系, 且不同领域下关联子领域的词语也具有领域差异性. 因此, 通过机器学习方法、深度模型思想和基于 nCRP 的层次主题模型均无法挖掘

© 中国科学院软件研究所 http://www.jos.org.cn

出具有明确层次关系和关联关系的主题层次结构.

本文通过分层次的主题方面共享机制改变 nCRP 构造方法中主题树形结构生成过程, 提出 nCRP+层次构造方法和 rHDP 层次主题模型, 挖掘不同主题下的关联子主题. 结合领域类别信息定义基于投票机制的领域隶属度计算方法, 目的是引导每层级词语集的主题分配过程, 明确主题与领域之间的映射关系, 构建领域主题之间的层次关系. 通过词语与领域主题的语义相关度引导主题-词语分配过程, 目的是将语义相近的词语分配在相同主题中, 凝聚领域主题涵义. 同时, 通过词语与其所在主题树分支中主题的领域相关性, 定义层次化的主题-词语贡献度, 明确关联子主题在主题词上的领域差异性.

结合基于投票机制的领域隶属度、词语与领域主题的语义相关度和层次化的主题-词语贡献度, 设计领域知识的形式化描述, 改进层次化的采样过程, 提出一种通用的、结合领域知识的层次主题模型 rHDP_DK, 实现领域主题层次关系和关联子主题共享关系的构建, 以及领域主题词的提取.

下一步工作将研究基于时变信息的领域主题层次结构, 便于分析各领域主题下的子主题及其主题词在不同时期的变化规律.

References:

[1] Liu TX, Xu MF. Can internet search behavior help to forecast the macro economy? Economic Research Journal, 2015, 50(12): 68–83 (in Chinese with English abstract).

[2] Blei DM, Griffiths TL, Jordan MI. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. Advances in Neural Information Processing Systems, 2007, 16(2): 17–24.

[3] Paisley J, Wang C, Blei DM, Jordan MI. Nested hierarchical Dirichlet processes. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015, 37(2): 256–270. [doi: 10.1109/TPAMI.2014.2318728]

[4] Ahmed A, Hong LJ, Smola AJ. Nested Chinese restaurant franchise processes: Applications to user tracking and document modeling. In: Proc. of the 30th Int'l Conf. on Machine Learning. Atlanta: JMLR.org, 2013. 1426–1434.

[5] Meng Y, Zhang YY, Huang JX, Zhang Y, Zhang C, Han JW. Hierarchical topic mining via joint spherical tree and text embedding. In: Proc. of the 26th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. ACM, 2020. 1908–1917. [doi: 10.1145/3394486.3403242]

[6] Huang JX, Xie YQ, Meng Y, Zhang YY, Han JW. CoRel: Seed-guided topical taxonomy construction by concept learning and relation transferring. In: Proc. of the 26th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. ACM, 2020. 1928–1936. [doi: 10.1145/3394486.3403244]

[7] Zhao H, Du L, Buntine W, Zhou MY. Inter and intra topic structure learning with word embeddings. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stroudsburg: PMLR, 2018. 5892–5901.

[8] Zhao H, Du L, Buntine W, Zhou MY. Dirichlet belief networks for topic structure learning. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 7966–7977.

[9] Isonuma M, Mori J, Bollegala D, Sakata I. Tree-structured neural topic model. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 800–806. [doi: 10.18653/v1/2020.acl-main.73]

[10] Gan Z, Chen CY, Henao R, Carlson D, Carin L. Scalable deep Poisson factor analysis for topic modeling. In: Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: JMLR.org, 2015. 1823–1832.

[11] The YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 2006, 101(476): 1566–1581. [doi: 10.1198/016214506000000302]

[12] 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 with English abstract). http://www.jos.org.cn/1000-9825/5898.htm [doi: 10.13328/j.cnki.jos.005898]

[13] Han ZM, Zhang MM, Li MQ, Duan DG, Chen Y. Flow hierarchical Dirichlet process for complex topic modeling. Chinese Journal of Computers, 2019, 42(7): 1539–1552 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2019.01539]

[14] Ma TF, Sato I, Nakagawa H. The hybrid nested/hierarchical Dirichlet process and its application to topic modeling with word differentiation. In: Proc. of the 29th AAAI Conf. on Artificial Intelligence. AAAI, 2015. 2835–2841.

[15] Ding YQ, Li SP, Zhang Z, Shen B. Hierarchical topic modeling with nested hierarchical Dirichlet process. Journal of Zhejiang University-SCIENCE A, 2009, 10(6): 858–867. [doi: 10.1631/jzus.A0820796]

© 中国科学院软件研究所 http://www.jos.org.cn

[16] Ding YQ. Probabilistic generative models-based topic modeling of text and its applications [Ph.D. Thesis]. Hangzhou: Zhejiang University, 2010 (in Chinese with English abstract).

[17] Wang C, Blei DM. Variational inference for the nested Chinese restaurant process. In: Proc. of the 22nd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2009. 1990–1998.

[18] Chen JF, Zhu J, Lu J, Liu SX. Scalable inference for nested Chinese restaurant process topic models. In: Proc. of the 23rd ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. Halifax, 2017. 1–9.

[19] Chen JF, Zhu J, Lu J, Liu SX. Scalable training of hierarchical topic models. Proc. of the VLDB Endowment, 2018, 11(7): 826–839. [doi: 10.14778/3192965.3192972]

[20] Song J, Huang Y, Qi X, Li YH, Li F, Fu K, Huang TL. Discovering hierarchical topic evolution in time-stamped documents. Journal of the Association for Information Science and Technology, 2016, 67(4): 915–927. [doi: 10.1002/asi.23439]

[21] Xuan JY, Lu J, Zhang GQ. A survey on bayesian nonparametric learning. ACM Computing Surveys, 2019, 52(1): 13. [doi: 10.1145/3291044]

[22] Huang WP, Laitonjam N, Piao GY, Hurley NJ. Inferring hierarchical mixture structures: A Bayesian nonparametric approach. In: Proc. of the 25th Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Delhi: Springer, 2021. 206–218. [doi: 10.1007/978-3-030-75768-7_17]

[23] Li W, McCallum A. Pachinko allocation: DAG-structured mixture models of topic correlations. In: Proc. of the 23rd Int'l Conf. on Machine Learning. Pittsburgh: ACM, 2006. 577–584. [doi: 10.1145/1143844.1143917]

[24] Hida R, Takeishi N, Yairi T, Hori K. Dynamic and static topic model for analyzing time-series document collections. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 516–520. [doi: 10.18653/v1/P18-2082]

[25] Liu R, Wang XG, Wang DQ, Zuo Y, Zhang H, Zheng XZ. Topic splitting: A hierarchical topic model based on non-negative matrix factorization. Journal of Systems Science and Systems Engineering, 2018, 27(4): 479–496. [doi: 10.1007/s11518-018-5375-7]

[26] Kim H, Drake B, Endert A, Park H. ArchiText: Interactive hierarchical topic modeling. IEEE Trans. on Visualization and Computer Graphics, 2021, 27(9): 3644–3655. [doi: 10.1109/TVCG.2020.2981456]

[27] Yang SH, Kolcz A, Schlaikjer A, Gupta P. Large-scale high-precision topic modeling on twitter. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM, 2014. 1907–1916. [doi: 10.1145/2623330.2623336]

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

[29] Chen ZY, Liu B. Mining topics in documents: Standing on the shoulders of big data. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM, 2014. 1116–1125. [doi: 10.1145/2623330.2623622]

[30] Li CL, Wang HR, Zhang ZQ, Sun AX, Ma ZY. Topic modeling for short texts with auxiliary word embeddings. In: Proc. of the 39th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Pisa: ACM, 2016. 165–174. [doi: 10.1145/2911451.2911499]

[31] 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). [doi: 10.11897/SP.J.1016.2015.01408]

[32] Alam H, Peltonen J, Nummenmaa J, Järvelin K. Tree-structured hierarchical dirichlet process. In: Proc. of the 15th Int'l Symp. on Distributed Computing and Artificial Intelligence. Toledo: Springer, 2018. 291–299. [doi: 10.1007/978-3-319-99608-0_33]

[33] Xu YS, Yin JW, Huang JB, Yin YY. Hierarchical topic modeling with automatic knowledge mining. Expert Systems with Applications, 2018, 103: 106–117. [doi: 10.1016/j.eswa.2018.03.008]

附中文参考文献:

[1] 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, 50(12): 68–83.

[12] 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 [doi: 10.13328/j.cnki.jos.005898]

[13] 韩忠明, 张梦玫, 李梦琪, 段大高, 陈谊. 面向复杂主题建模的流式层次狄里克雷过程. 计算机学报, 2019, 42(7): 1539–1552. [doi: 10.11897/SP.J.1016.2019.01539]

[16] 丁轶群. 基于概率生成模型的文本主题建模及其应用 [博士学位论文]. 杭州: 浙江大学, 2010.

[31] 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. [doi: 10.11897/SP.J.1016.2015.01408]

© 中国科学院软件研究所 http://www.jos.org.cn

附录 A

表 A1 数据集的领域分类及其代表性词语

数据集	领域类别	领域名称	领域代表性词语
财经微博文本	一级	投资	华夏基金; 国债; 利率; 融资; 期权; 收藏品; 贵金属; 外汇交易; 股市; 炒股; 证券; 创业; 邮票; 投资; 房地产; 保险; 贷款; 银行; 基金; 理财产品
	一级	进出口	运费; 出口; 进口; 进出口; 出口许可证; 成品油; 汽车零部件; 出口退税; 铁矿石; 石油; 外汇; 期货; 进口化妆品; 进口葡萄酒; 出口信贷; 中国进出口贸易网; 进出口代理; 海关编码查询; 来料加工; PMI; 海运费; 关税查询; 原油; 进出口银行; 钢材; 集装箱; 大豆
	一级	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
	一级	消费	户外; 首饰; 消费; 装饰; 电器; 促销; 书; 茶; 酒店; 自助; 出国; 音响; 餐巾纸; 享受; 时装; 演唱会; 维修; 健身房; 玩具; 团购; 减肥; 美容
	一级	人口与就业	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
	二级	人口	人口; 出生率; 死亡率; 人口年龄结构; 抚养比; 寿命; 人口普查; 人口抽样; 经济活动人口; 就业; 工资; 平均工资; 失业; 失业率
	二级	固定资产投资(侧重房地产方面)	固定资产投资; 自筹; 贷款; 建筑安装工程; 设备工器具购置; 中央项目; 地方项目; 新建固定资产; 投资; 扩建固定资产投资; 改建固定资产投资; 固定资产交付使用率; 投资; 国家预算资金; 房屋施工面积; 房屋竣工面积; 施工项目; 新开工项目; 建成投产项目; 项目建成投产率; 建设规模; 施工规模; 新开工规模; 累计新增生产能力; 新增生产能力; 农村农户建房; 投资额; 建筑面积; 房屋造价; 住宅造价; 房地产开发企业; 房地产开发企业从业人数; 开发土地面积; 购置土地面积; 土地成交价款; 土地购置费用; 新开工房屋; 住宅; 别墅; 高档公寓; 办公楼; 商业营业用房; 商品房; 预收款; 个人按揭; 房地产企业资产负债; 累计折旧; 所有者权益; 资产负债率; 土地转让收入; 商品房销售收入; 商品房销售; 房屋出租; 投资额; 营业利润; 竣工
	二级	对外经贸	货物进口总额; 出口货物; 进出口总额; 进口货物; 海关; 外商投资企业; 外资; 合同利用外资项目; 外商投资项目; 外商投资额; 外资项目; 注册资本; 对外经济合作; 对外承包工程; 对外劳务合作; 人民币汇率
	二级	能源	能源生产弹性系数; 能源消费弹性系数; 能源加工转换效率; 发电; 电站供热; 炼焦; 炼油; 能源生产量; 消费量; 煤炭; 焦炭; 原油; 燃料油; 汽油; 煤油; 柴油; 天然气; 液化石油气; 热力; 电力; 人工煤气; 高耗能产品; 钢材; 铜; 铜合金; 铝; 铝合金; 肥料; 纸浆; 纺织用合成纤维; 水泥; 平板玻璃; 铜材; 铝材; 锌; 锌合金; 纸; 纸板; 能源工业; 煤炭; 石油; 天然气; 电力; 蒸汽; 热水; 石油加工; 炼焦业; 煤气; 能源生产; 原煤; 原油; 天然气; 水电; 核电; 风电; 焦炭; 汽油; 煤油; 柴油; 燃料油; 天然气; 发电量; 水力发电; 火力发电; 供热; 蒸汽供热; 热水供热
	二级	财政	国家财政; 中央财政; 地方财政; 政府收入; 政府支出; 财政收入; 财政支出; 财政收入比重; 财政支出比重; 财政收入项目; 增值税; 消费税; 营业税; 关税; 企业所得税; 个人所得税; 资源税; 城市维护建设税; 房产税; 印花税; 证券交易印花税; 土地使用税; 土地增值税; 车船税; 船舶吨税; 车辆购置税; 关税; 耕地占用税; 契税; 烟叶税; 非税收入; 专项收入; 行政事业性收费; 罚没收入; 国有资本经营收入; 国有资源有偿使用收入; 财政支出项目; 公共服务支出; 财政外交支出; 对外援助支出; 国防支出; 公共安全支出; 武装警察支出; 教育支出; 科学技术支出; 文化体育支出; 传媒支出; 社会保障支出; 就业支出; 医疗卫生支出; 环境保护支出; 城乡社区事务支出; 农林水事务支出; 交通运输支出; 车辆购置税支出; 地震灾后恢复重建支出; 国内债务; 国外债务; 预算外资金收入; 支出; 外债
	二级	人民生活	居民收支; 居民人均收入与支出; 家庭人均收入; 恩格尔系数; 人民币储蓄存款; 消费; 现金; 食品; 粮食; 肉禽; 蛋类; 水产品; 奶制品; 衣着; 服装; 居住; 住房; 家庭设备和用品; 耐用消费品; 交通通信; 文教娱乐; 文化娱乐用品; 医疗保健; 购买主要商品; 粮食; 鲜菜; 食用植物油; 猪肉; 牛羊肉; 禽类; 鲜蛋; 水产品; 鲜奶; 鲜瓜果; 酒; 煤炭; 耐用消费品; 洗衣机; 电冰箱; 彩色电视机; 组合音响; 照相机; 空调; 淋浴热水器; 计算机; 摄像机; 微波炉; 健身器材; 移动电话; 固定电话; 家用汽车
	二级	城市基础设施	城市建设; 城市供水; 城市燃气; 供气; 用气; 城市集中供热; 城市市政设施; 道路; 桥梁; 排水管道; 污水处理; 道路照明; 城市公共交通; 公共交通车辆; 公共汽电车; 轨道交通; 出租汽车; 城市绿地; 园林; 绿地面积; 公园; 绿化覆盖率; 城市市容环境卫生; 道路清扫保洁; 生活垃圾清运; 粪便清运; 市容环卫; 公共厕所; 城市设施水平; 公共交通车辆; 道路面积; 公园绿地面积; 公共厕所

© 中国科学院软件研究所 http://www.jos.org.cn

表 A1 数据集的领域分类及其代表性词语 (续)

数据集	领域类别	领域名称	领域代表性词语
财经微博文本	二级	资源环境	矿产; 石油; 天然气; 煤炭; 铁矿; 锰矿; 铬矿; 钒矿; 原生钛铁矿; 铜矿; 铅矿; 锌矿; 铝土矿; 镍矿; 钨矿; 锡矿; 钼矿; 锑矿; 金矿; 银矿; 稀土矿; 菱镁矿; 普通萤石; 硫铁矿; 磷矿; 钾盐; 盐矿; 芒硝; 重晶石; 玻璃硅质原料; 石墨; 滑石; 高岭土; 水; 地表水; 地下水; 废水; 污染物; 废气; 二氧化硫; 氮氧化物; 烟尘; 粉尘; 生活垃圾; 森林; 造林; 林业重点工程造林; 天然林保护工程; 退耕还林工程; 三北及长江流域防护林建设工程; 京津风沙源治理工程; 速生丰产用材基地建设工程; 造林面积; 草原建设利用; 种草面积; 治理面积; 受害面积; 自然保护区; 自然灾害; 旱灾; 洪涝; 山体滑坡; 泥石流; 台风; 风雹; 低温; 雪灾; 地质灾害; 地面坍塌; 崩塌; 森林火灾; 森林病虫鼠害防治; 地震灾害; 环境污染; 基础设施建设; 园林绿化建设; 市容环境卫生建设; 工业污染源治理; 工业污染; 废水; 废气; 固体废物; 噪声
	二级	农业	农村基层组织; 乡村人口; 农业器械; 有效灌溉面积; 农用化肥; 农村水电建设; 灌溉; 水库; 除涝; 治水; 农用柴油; 农药; 农用塑料薄膜; 农作物; 播种; 果园; 瓜果; 茶园; 茶叶; 水果; 林产品; 牲畜饲养; 畜产品; 水产品; 国有农场
	二级	工业	企业; 国有控股工业企业; 私营工业企业; 外商及港澳台商投资工业企业; 大中型工业企业; 工业产品; 原盐; 精制食用植物油; 成品糖; 罐头; 啤酒; 卷烟; 纱; 布; 机制纸; 纸板; 汽油; 柴油; 焦炭; 硫酸; 烧碱; 纯碱; 乙烯; 合成氨; 农用; 氮; 磷; 钾; 化肥; 化学农药原药; 塑料; 橡胶; 合成洗涤剂; 中成药; 化学纤维; 水泥; 玻璃; 生铁; 粗钢; 钢材; 重轨; 钢筋; 有色金属; 铜; 铁路客车; 货车; 轿车; 自行车; 摩托车; 家用电器; 火电发电; 水电发电
	二级	建筑业	建筑业企业; 承包; 劳务分包; 勘察设计机构; 工程招标代理机构; 建设工程监理企业; 工程监理; 工程招标代理; 工程招标咨询; 建设工程监理
	二级	运输邮电	交通运输; 铁路; 公路; 城市公共交通; 水上运输; 航空运输; 管道运输; 装卸搬运; 邮政业; 电信; 信息传输服务业; 高速等级路公路; 一级等级公路; 等外公路; 内河航道; 定期航班航线; 国际航线; 管道输油; 民用航空; 地方铁路; 合资铁路; 远洋货运; 管道货运; 邮电通信; 运输线路; 客运量; 旅客周转量; 货运量; 货物周转量; 旅客运输平均运距; 货物运输; 国家营业铁路; 铁路机车; 铁路客车; 铁路货车; 干线; 旅客发送量; 货物发送量; 民用汽车; 私有汽车; 新注册民用汽车; 公路营运汽车; 民用运输船舶; 港口分货类吞吐量; 货物吞吐量; 港口码头泊位数; 民航航线; 民航飞机; 快递; 邮政业网点
	二级	信息技术	互联网; 电信; 邮政; 企业信息化; 电子商务; 万吨级泊位; 运输量; 包裹; 报刊; 汇票; 集邮; 移动电话; 固定电话; 通话时长; 3G; 公用电话; 快递业务; 信筒; 信箱; 投递路线; 邮路; 交换机; 万路端; 光缆; 电话; 互联网; 宽带; 通邮; 营业网点; 上网人数; 域名; 网站; 网页; 带宽; 接入端口; 拨号; 移动互联网; 软件; 软件产品; 信息技术服务; 信息安全; 嵌入式系统; 软件业务; 计算机
	二级	批发零售	批发企业; 零售企业; 连锁零售企业; 市场摊位; 连锁
	二级	旅游	住宿; 餐饮; 连锁餐饮; 客房数; 床位数; 国内旅游; 国际旅游; 外国入境; 游客
	二级	金融业	金融机构; 人民币信贷资金; 单位存款; 个人存款; 财政性存款; 临时性存款; 委托存款; 金融债券; 流通中货币; 金融机构负债; 贷款; 融资租赁; 票据融资; 垫款; 有价证券; 股权; 黄金占款; 外汇占款; 货币供应量; 货币供应量; M1; 准货币供应量; M2; 活期存款; 定期存款; 流通中现金供应量; M0; 黄金储备; 外汇储备; 债权; 货币当局资产负债表; 存款新公司资产负债表; 委托贷款; 外币贷款; 信托贷款; 未贴现银行承兑汇票; 企业债券; 债券; 证券; 期货; 外资银行; 资产负债表; 社会融资规模; 证券市场; 上市公司; 股票发行量; A股; B股; H股; 股票筹资额; 股票交易; 流通股本; 流通市值; 成交金额; 上证综合指数; 深证综合指数; 收盘; 保险系统; 保险公司; 保费; 赔款; 给付
	二级	教育	学校; 教职工; 专任教师; 学历教育; 非学历教育; 学生; 招生; 在校学生; 毕业生; 研究生; 留学生; 高等教育学校; 成人本科; 成人专科; 普通本科; 普通专科; 网络本科; 网络专科; 中等职业学校; 职业技术培训机构; 普通高中; 初中学校; 普通小学; 技工学校; 学龄儿童; 入学; 高等学校; 生师比; 教育经费; 博士; 硕士; 研究生; 升学率; 入学率; 授予学位; 正高级; 副高级; 中级; 初级; 无职称; 行政人员; 教辅人员; 工勤人员; 职业资格证书; 应届; 校外教师; 民办学校; 社会捐赠; 事业收入; 学杂费
	二级	科技	科技活动; 科学研究; 开发机构; 科技活动; 新产品开发; 专利; 研究; 试验发展; 经费; 项目; 课题; 研究开发机构; 产品创新; 工艺创新; 高技术产业; 申请受理; 授权; 发明; 实用新型专利; 科技论文; 检索工具; 高技术产品; 工业制成品; 技术人员; 测绘地理信息部门; 测绘资料; 地震监测; 气象业务站点; 产品质量; 出入境货物检验检疫; 中国科协系统科技活动; 科技著作; 申请受理专利; 专利申请授权; 国内技术; 国外技术; 技术改造
	二级	医药卫生	医疗卫生机构; 卫生人员; 卫生室; 床位数; 门诊; 住院; 医疗服务; 床位; 乡镇卫生院; 社区卫生服务中心; 传染病; 疾病; 死亡率; 死因; 儿童; 孕产妇; 新兴农村合作医疗

© 中国科学院软件研究所 http://www.jos.org.cn

表 A1 数据集的领域分类及其代表性词语 (续)

数据集	领域类别	领域名称	领域代表性词语
财经微博文本	二级	社会服务	社会服务机构; 社会工作师; 社会福利企业; 孤儿; 收养; 社会救助; 医疗救助; 优抚安置; 福利彩票; 社会捐赠; 社区服务机构; 婚姻服务; 殡葬服务; 社会组织; 残疾人事业
	二级	文化体育	文化文物机构; 艺术表演团体; 艺术表演场馆; 公共图书馆; 群众文化机构; 文物; 博物馆; 广播电视; 节目; 节目制作; 播出; 有线广播电视; 传输干线网络; 广播电视技术; 电视节目; 电影; 图书出版; 期刊; 报纸; 少年儿童读物; 课本; 录音制品; 电子出版物; 音像; 版权合同登记; 引进和输出版权; 出版印刷; 发行机构; 档案馆; 体育系统; 运动员; 教练员; 冠军
	二级	公共管理	全国人民代表大会代表; 全国政治协商会议委员; 妇联干部; 工会组织; 劳动争议; 律师; 公正; 调解; 公证文书; 民间纠纷; 在押服刑; 公安机关; 立案; 刑事案件; 治安案件; 交通事故; 火灾事故; 人民检察院; 侦察; 审查批准; 逮捕; 犯罪嫌疑人; 申诉; 出庭; 公诉; 刑事; 抗诉; 举报; 控告; 纠正违法; 一审; 审理; 收结案; 犯罪; 婚姻家庭; 继承; 合同纠纷; 权属; 侵权纠纷; 民事; 行政一审案件; 社会保险基金; 基本养老保险; 社会保险; 失业保险; 基本医疗保险; 工伤保险; 生育保险; 新型农村社会养老保险; 城乡居民社会养老保险
20News Group	一级	alt	atheism; god; believe; atheist; bible; moral
	一级	comp	graphic; image; os; program; directory; mode; microsoft; hardware; monitor; window; server; display; computer
	一级	misc	sale; shipping; sell; price; contact; pay; reply
	一级	rec	auto; power; vehicle; motorcycle; bike; moto; sport; game; team; player; hockey; baseball
	一级	sci	encryption; secure; electronic; circuit; chip; radio; voltage; signal; medical; doctor; patient; medicine; health; space; earth; nasa
	一级	soc	religion; christian; god; jesus; christ; church; faith; lord
	一级	talk	politics; gun; war; weapon; force; crime; attack; military
	二级	atheism	uncompromising; hippo; gloating; encompassing; profer; theism; swearing; hindered; solipsism; fakeness; gurer; callously; anyday; ozguven; precept; Sadducee; punishement; theological; atheist; god; nonsolution; sackcloth; dawned; stallion; vaudeville; repellant; observation; whill; atheist; blasphemer; lob; reinscribed; unwantingly; polly; gipu; deemphasizing; insincere; falwell; burnet; galvanizing; Kierkegaard; caramelizing; rodolf; believe; freethought; unscientific; unassailable; eternality; bible; identically; empathize; aquatic; stemmed; moral; sabrus; jvagre; agapate; heus; diseased; grandeur; determnined; darby; masculine; freethinker; prupose
	二级	graphics	Wysiwig; bpus; university; iraf; graphic; attlet; foti; nln; moschettus; vernam; uninhabited; ugle; nobly; scivus; subtraction; iname; image; color; quek; wollongong; bassel; forsman; redrawing; nnn; yuv; Isabel; merchantability; sucessfuly; lgtgah; chongo; bof; Irenaeus; attentive; line; mftobdf; lenth; unipress; rewrote; myterm; Erivan; chasis; brazen; tenfold; tiled; congressionally; invaluable; lezghistan; fputc; huygen; Disneyland; tangy; nodi; rewriting; elevation; knu; stdarg; physic; scf; Confucianism; isalnum; condor; petruccus; isascius; ellipsis; software; wisc; ioccc; siam; mcelwane; algorithm; bit; feg; ncoa; framegrabber; visualizer; statbuf
	二级	os	jaffer;obispo; implantation; vcpus; sft; troff; setenv; chilicothe; submittion; seamlessly; lzh; sculpt; iraf; stadt; roelof; submitter; clipbychildren; supervga; dcw; lub; disgustingly; program; mactcp; soundcard; bcopy; card; ikph; allegro; ffff; lbhefry; Microsoft; wytiwig; forsman; biscuit; gobbling; checkling; multipart; imdisp; chatted; hian; watchit; outsell; morphological; lnq; supposed; paintshop; orcad; rasterizing; roell; topographic; vef; raveling; heterogeneity; chasis; prodesigner; uncomplete; idl; memory; announcement; stasis; hua; niftp; armegedon; ultrastore; mode; dragondictate; proprinter; will; evaded; fstobdf; nedde; bitblit; pctcp; bbn; radious; transport; version; os; disk; untarred; mimicry; feg; kelder; yourk; reconfiguring; directory; bbs; erratum
	二级	sys	cda; cfg; spaceward; trigonometry; argumentum; photograde; rokne; unplugging; pamelum; cryptologium; blinn; itselft; kinesis; voynich; ogus; ugle; rium; apple; qwerty; macromind; cyen; mackermit; modularity; gatekeeper; maariv; incompatibable; jaffer; vender; catone; panned; macpaint; hardware; dhh; machine; mac; inconsiant; rivero; valus; uncomplete; tiled; denton; openl; board; lostweekender; idl; assy; nearside; yih; nodi; will; cramped; lvc; crystalline; funneled; Iverson; oid; monitor; vigenere; sumitomo; spacify; ofvlb; zambonus; ecr; referring; compute; succeded; kantrowitz; irbm; unsorted; hollyman; nistnew; isspace; techmo; seanmac; datahand; tecmo; heinbokel

© 中国科学院软件研究所 http://www.jos.org.cn

表 A1 数据集的领域分类及其代表性词语 (续)

数据集	领域类别	领域名称	领域代表性词语
20News Group	二级	windows	remailing; rasterizing; pfuetzner; vimage; astrod; commonview; kisseberth; sharable; myterm; unizh; unghost; eece; iraf; shurik; complile; kaleidagraph; buffington; ariston; will; aster; galvanizing; gambarian; stdarg; nikolao; bosco; rivlin; forsman; visualid; Iverson; unintelligible; kafan; display; wind; ow; previewing; lightwave; server; amann; eqn; mftobdf; cyberware; visibilitynotify; miff; bib; visualizer; package
	二级	forsale	northland; stephanopoulo; britamrica; indirecly; foregone; ibmpc; dutifully; dispel; sating; mcgowan; duster; marr; bwsmith; interopen; swirling; keown; biodegradable; outliner; disencourage; matic; naif; biasing; sell; bushbaby; deadpool; price; arghhh; commercialization; sluice; goblin; davar; pucker; clrview; sale; stopgap; stuffit; glavcosmo; prakash; pftb; tactile; layne; myterm; unizh; reply; magil; mcfarlane; originator; suke; hobgoblin; liefeld; infiltrating; dachsel; geordy; basalt; tist; sabretooth; railing; pay; veo; qjn; plasce; uww; chvrtd; shipping; heus; danto; lugar; slr; lulling; contact; cuperman; blurt; giger; ashkar; toa
	二级	motorcycles	waterskus; peeking; delaminate; moisten; shelly; yamaha; pigeonhole; fringed; urshan; carport; kotdohl; thinkage; motorcycle; cotterpin; rosander; nakada; helmet; rodale; guzzi; bowels; monza; telepathic; goggleless; dodland; coterill; locksmithing; kqspt; predisposing; nondamaged; bike; tweeked; bieffe; sturge; testor; fant; preciated; hemorroid; wheeling; moto; cheerfully; handson; vij; heus; pinkie; hogssc; throiugh; weathered; sievert; ican; kotrb; lezgin; nascar; ride; linered; kotwitdodl; yuka; jerked; cousinetc; lotto; hopalonga
	二级	auto	outperform; roell; accellerator; linclon; matejka; arfa; idb; hereabout; millet; adsorbent; characteristically; congregational; amc; stitched; synchro; carburated; guesstimate; nubar; vehicle; engine; srinagesh; busiest; rosander; offramp; cushioned; car; gruffness; switchover; arcing; impreza;speed; adsorption; auto; affectionado; sung; lusky; mro; stricly; layton; flatpack; reingold; thumbing; drriver; oversteer; tripmeter; merican; patrolman; couldnt; cranston; power; draftee; gonorrhea; autotrace; ayoob; rrh; canine; concurrence; nrao; bedload; jit; affordably; aflame; synchronization; deform
	二级	sport	nepotism; nettle; revelant; gordie; kcr; mightier; apd; game; keikaku; bongo; reipa; sportscenter; booo; bip; kea; arbour; sisler; eerily; offen; beltway; liiga; matikainen; kamloop; rium; stuppid; ojalum; spry; armchair; team; kemppainen; harrus; sport; sincerest; linare; stargell; lgtgah; unsung; soderstron; swe; ciau; catullus; cringing; sively; damphouse; homesick; newfound; jyvaskylum; cfl; dandridge; hit; afc; hovering; porus; hockey; horvath; restless; restructure; elitserien; ecac; buckner; ynyus; backhand; stelter; minuteman; hoth; likened; fan; oquendo; henley; superpro; rosander; homeboy; linseman; skated; brodcaster; markoff; greania; playoff; bandwaggoner; player; rname; maniacal; earring; bronson; underachiever; mattel; pravin; acclimated; kovat; wheeling; lenarduzzus; dehydrator; league; essary; laud; siz; lafleur; abotu; laurilum; overhyped; amarillo; danno; baseball; graig
	二级	crypt	dolgo; khafre; vogt; strengthening; carman; britamrica; agitate; vertiginous; egaltarian; vladimirov; suall; kaliskus; encryption; cryptologium; lub; weinberger; bamfd; secure; subacute; cryptography; wullenweber; ftprf; mundelein; secret; koelln; scivus; helsingius; knu; iname; criminality; audoban; ree; coder; lpf; stdarg; overwriting; modularity; keyspace; ecpa; physik; effector; wel; pgp; merchantability; keysize; ftpsf; unicity; scfvjspj; chongo; algorithm; unsolved; scif; feg; prus; stuffit; steganographic; pseudonymity; visualizer; statbuf
	二级	electronics	wysiwig; netzer; gfci; cusation; roboticist; keikaku; nowady; modernizing; peltier; wazing; tulin; chonglab; bjt; hanauma; dbm; tekprstbl; chip; plonk; signal; electronic; dectector; bitstream; cruptology; anderton; reallocated; fumigant; ungrounded; epd; offed; electrical; selonoid; cometary; stryker; decus; radio; mamishev; chongo; rrh; leben; momentary; pseudonymity; phenomenologist; multimeter; multidimensional; noiseless; lovall; draught; sbl; peddling; meholum; paramus; vica; meida; circuit; selonid; suncoast; easel; electrican; satelly; busbar; naf; unencrypted; amp; elisp; shiney; bulldada; voltage; jelloman; irled; ibuild; ription; rectified; input; penet; milliwatt; bullwinkle; charproc; cicruit; inseperably; nrao; intermittant; galvanic; dunning; icl; paleness; snugly; skd; ckt

© 中国科学院软件研究所 http://www.jos.org.cn

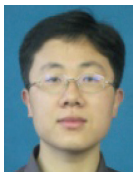
表 A1 数据集的领域分类及其代表性词语 (续)

数据集	领域类别	领域名称	领域代表性词语
20News Group	二级	med	cupping; psoriasis; sewer; hypha; chilicothe; kale; esplanade; genentech; haldol; iraf; muscled; medicine; shereware; microgenesy;s carcer; rmit; effctive; subacute; coli; rodale; panetta; occidental; litigate; sheryl; gephardt; rivlin; hicnet; inhibitor; verbably; undyed; socialized; hemorroid; throiugh; privatizing; melanoma; indefensible; health; eugenics; bolden; schwartzenegger; heteropathic; accosted; morbidity; medical; unwarented; treatment; unrepetent; dodell; sadikov; banschbach; hysteric; ovulation; shingle; effect; disease; meniere; fright; dce; offloaded; physiotherapist; understoood; iocomm; predisposing; innoculate; spirochetal; rayaz; peptide; patient; quarantine; stenson; twiddler; aspartate; danto; doctor; lugar; turmeric; teaspoon; deteriorated; subterranean; bigshot; hazarded; humanoid; wavefront; chiron
	二级	space	gto; upa; magnetosphere; bowery;l arrison; hypha; elv; llnl; trignometry; satellite; yeian; tether; keplerian; nfoti; rjl; datacube; standdown; asuka; atraction; nasa; keybd; eugenia; prospector; usno; clipert; carumba; prnt; ppb; rejuvenated; space; astrophysical; cometary; commercialization; ntt; wavefront; raz; goblin; bof; passageway; earth; jsut; salyut; qna; jbi; dozer; hemoraging; argonaut; nizoral; topographic; konstantin; uubuild; taur; shrrchr; flight; mclaughlin; crust; tonne; thoughtless; billionaire; congressionally; dataset; moon; reassignment; rocket; stesis; haggle; spacelink; bettadpur; satelly; topography; yeasteryear; liefeld; erratum; wullenweber; unneccessary; lunar; basalt; miya; memmedov; costar; trotman; sameach; blimp; puppet; hic; layton; rio; lodging; willmann; shuttle; kinky; scintillation; waivered; launch; uncorrected; soyuzkarta; schuerkamp; angstrom; jahn; raker; intelsat; firstep; cate; teleoperated; trig
	二级	religion	astronomically; millet; hargreave; prego; nebuchanezzar; maccabee; sahl; precept; cling; christian; whill; preferring; faith; assyrian; thanksgiving; impostor; grafted; christology; everlasting; dardanelle; desciple; falwell; burnet; clime; collosion; clipert; tare; bce; kierkegaard; crossword; hornet; god; harangue; delcaration; liane; filmmaker; eternality; assimilated; richness; aquatic; stemmed; lord; huebner; crucification; psalmist; fervent; indwell; arrogantly; disobeyed; contorting; jotted; hippo; popupwindow; donpaul; swearing; hindered; tiered; babtized; tional; impule; indescribably; cheerleader; unfaithful; disobedient; presiden; rationalism; fifteenth; freethought; cometh; cannibal; religion; cohesion; patriarchal; jvagre; suspecting; assyrium; schoep; falty; masculine; freethinker; munn; jregzhyyre; christ; church; maiming; jesus
	二级	politics	attrocity; narloch; blight; military; adjourned; armied; patriarchate; foulston; siyasus; saderak; wncus; techy; humanbeing; war; consular; tractinsky; condusive; weapon; dardanelle; antiochus; force; menachem; medicare; shalhevet; neccity; stepanakert; iksa; ealyan; akman; concealibility; disvalidate; sharif; voluntarilly; weitz; endeth; heghinian; stemmed; notoriety; sabrus; politics; ffl; sabetta; authenticator; genitals; stalemate; ratification; polybagged; purposeful; preyed; uludamar; rumanium; countersigned; ihtilalus; moustache; synthesize; terrritory; nejm; apocalypso; sinaus; sadikov; barrelled; sepulchre; victimization; theyu; menahem; belgrade; ihsan; concideration; unmet; armoured; yuval; damsel; comfirm; administrivium; heracleous; belligerent; noncombatant; israel; sheduled; nagasakus; godiva; origination; gun; thatcher; azer; northener; implimenting; leclerc; cation; heftier; intermittant; crime; klan; mecli; hitlerism; awalus; attack

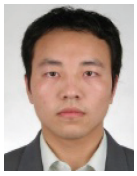
© 中国科学院软件研究所 http://www.jos.org.cn

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.

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



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 Web 数据管理, 信息检索, 自然语言处理.



刘喜平(1981－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为信息检索, 数据挖掘.

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.

万齐智(1988－), 男, 博士, 讲师, CCF 专业会员, 主要研究领域为自然语言处理, 数据挖掘.

© 中国科学院软件研究所 http://www.jos.org.cn