融合锚词抽取的海量短文本主题层次挖掘

吕 品,计春雷,汪 鑫,罗宜元
(上海电机学院电子信息学院,上海 201306)

摘 要: 从短文本集中挖掘不同粒度的主题、构建主题的层次结构在舆情分析、视觉检测、语义挖掘和图谱构建等方面具有重要应用.围绕如何从短文本集中分层次地挖掘主题,在修改传统短语定义的基础上,提出了融合锚词抽取的海量短文本主题层次挖掘框架.提出的主题层次挖掘框架首先基于词共现图实现主题推断和锚词抽取;然后,应用关联规则挖掘频繁锚词短语;最后,采用排序方法量化锚词短语以寻找最具代表性的主题短语.与已有的基于词共现图构建主题层次的方法相比,融合了锚词抽取的词共现图分析方法更有利于构建层次更高的主题.在2个实际的中文短文本数据集上执行实验,结果表明提出的方法挖掘的短语能较好地解释主题和用于分类预测.

关键词: 短文本;词共现图;主题层次;锚词
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2018)05-1084-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.05.009

Mass of Short Texts Topical Hierarchy Mining Integrated Anchor Extraction

LÜ Pin,JI Chun-lei,WANG Xin,LUO Yi-yuan
(*School of Electronic and Information,Shanghai Dianji University,Shanghai* 201306,*China*)

Abstract: A topical hierarchy at different levels of granularity from short texts has many valuable applications in the areas of opinion analysis,vision detection,semantics mining and graph construction. Aiming at how to mine the hierarchy of topics from short texts,a topical hierarchy mining framework integrated anchor extraction is proposed based on the modification of the tradition phrase definition. Firstly,the topic inference and the anchor extraction are conducted in the proposed framework. Secondly,frequent anchor phrases are found by applying associate rule mining. Finally,a kind of rank method is used to quantity the criterion of anchor phrases in order to find the most representative topical phrase by ranking. Compared to the topic analysis method of the word co-occurrence graph,the word co-occurrence integrated into anchor is more beneficial to build the higher level of topics. Experiments with datasets from the two Chinese short texts are performed,and the results show that the proposed method can generate interpretably phrases and be used for classification prediction.

Key words: short texts;word co-occurrence graph;topical hierarchy;anchor word

1 引言

大数据时代文本数据迅猛增长,从这些海量文本中找到感兴趣的主题是很多应用的关键,如舆情分析、视觉检测[1]、语义挖掘[2]、图谱构建[3]等.电子科技文献作为一种文本,其标题浓缩了文献的核心研究内容.对文献标题构成的短文本集进行挖掘,从中抽取不同粒度的主题,并将其组织为层次结构,有利于科研人员快速发现关键科学问题.

已有研究表明短语是主题的表现形式[4].因此,本文研究的基本单元是短语.近年来,从文本中挖掘短语的方法主要有2种:图排序[5~8]和主题模型[9~12].图排序方法通过构建一个文本的词图,找到属于不同主题的词.通过设计词排序策略对词排序,将排序靠前的词组合成关键短语表示主题.由此可见,图排序方法的研究对象是单个文本,无法直接应用于短文本集.主题模型方法是基于LDA(Latent Dirichlet Allocation)模型的扩展,通过抽取连续的词构成短语表示主题.由连续

收稿日期:2016-11-17;修回日期:2017-11-14;责任编辑:郭游
基金项目:国家自然科学基金青年基金(No.61402280);上海电机学院计算机科学与技术优势学科(No.16YSXK04);上海市教育科学研究项目(No.C17014)

的词构成短语是短语的传统定义[13]. 然而,这种传统意义的短语定义在科技文献标题表示的短文本集中,会出现无法区分不同短语表示相同主题的情况,导致主题分类错误. 另外,短文本集具有稀疏性,如果采用主题模型中的文档-词分布表示主题,则无法挖掘子主题,难于形成主题的层次.

针对以上 2 种方法在本应用中的局限,提出了融合锚词抽取的主题层次挖掘框架. 该方法首先构建词共现图 G = < V,L > 推断短文本集中的主题分布;其次,估计每一主题的锚词,采用关联规则挖掘短语的频繁模式;最后,用锚词短语的不完整性标准修剪频繁模式,得到含锚词的频繁模式. 由于图 G 既能用原始短文本集中的词共现关系建立,也能用已得到的主题上的词共现关系建立. 因而,词共现关系能有效支持主题层次结构. 方法在真实的短文本集上进行了测评,结果表明,所提出的方法在发现子主题时,其性能比其它 2 种方法更好. 同时,该方法利用锚词短语,有效避免了主题聚类过程中频繁模式的丢失,提高了挖掘出的主题短语的质量.

2　融合锚词的主题层次挖掘框架

融合锚词的主题层次挖掘框架 THM(Topical Hierarchy Mining)给出了基于词共现图的主题层次挖掘流程,如图 1 所示. 该框架由 3 个层次构成:顶层主题、次顶层主题和最终的主题层次,体现了主题层次的递归构成过程. 主题层次用根为 r 的树 T 表示. 树中非根结点 t 称为主题,由信息对 $\{P_s^t, R(P_s^t)\}$ 表示,其中,P_s^t 表示属于主题 t 的锚词短语集,$R(P_s^t)$ 表示对主题 t 的锚词短语进行排序的排序函数. 一个锚词短语可同时属于多个主题,但其排序并不相同. 主题 t 的所有子主题由其孩子结点集 $Child^t = \{z \in T | par(z) = t\}$ 表示. 其中,$par(z)$ 表示 z 的双亲主题. 每一层次的主题均由构建词共现图和挖掘主题短语两个步骤组成.

2.1　构建词共现图

假设词 w_i 与 w_j 的每一次共现都归因于它们属于同一主题,主题 z 有 k 个子主题($z \in [1,k]$),$l_{ij} = \sum_{z=1}^{k} l_{ij}^z$ 表示 w_i 和 w_j 构成的边数量,则聚类含 k 个子主题的主题 z 只需要估计词共现图 G 中 l_{ij}^z 的分布. 于是,随机变量 l_{ij}^z 的生成模型服从泊松分布 $l_{ij}^z \sim Possion(\delta_z \alpha_i^z \alpha_j^z)$,其中,$\delta_z$ 表示事件“生成主题 z 的边”发生的次数,$\sum_{i=1}^{N} \alpha_i^z = 1$,$\sum_{i=1}^{N} \alpha_j^z = 1$,$\delta_z \geqslant 0$,$\alpha_i^z \geqslant 0$,$\alpha_j^z \geqslant 0$. 若用 l_{ij} 表示图 G 中所有边,则当 δ_z 较大时,由泊松分布的累加属性有 $l_{ij} = \sum_{z=1}^{k} l_{ij}^z \sim Poisson(\sum_{z=1}^{k} \delta_z \alpha_i^z \alpha_j^z)$. 于是,在模型参数 α, δ 已知时,可通过公式(1)观察图 G 中所有的边.

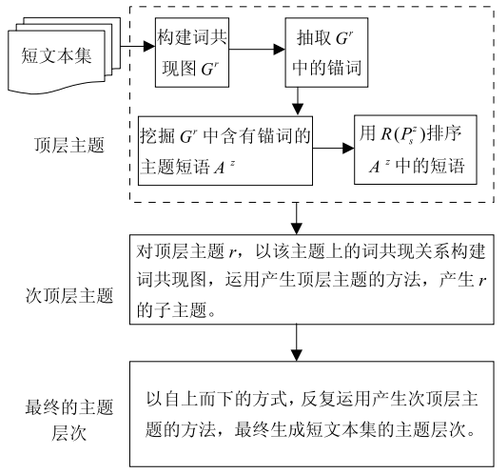

图 1　基于词共现图的主题层次挖掘流程

$$p(l_{ij} \mid \alpha, \delta) = \prod_{w_i, w_j \in W} p(l_{ij} \mid \alpha_i, \alpha_j, \delta) = \prod_{w_i, w_j \in W} \frac{(\sum_{z=1}^{k} \delta_z \alpha_i^z \alpha_j^z) \exp(-\sum_{z=1}^{k} \delta_z \alpha_i^z \alpha_j^z)}{l_{ij}!} \tag{1}$$

采用期望最大化算法对公式(1)中的参数进行估计. 由公式(2)实现 E 步的更新,由公式(3)和(4)实现 M 步的更新.

$$\gamma_{ij}^z = l_{ij} \frac{\delta_z \alpha_i^z \alpha_j^z}{\sum_{z=1}^{k} \delta_z \alpha_i^z \alpha_j^z} \tag{2}$$

$$\delta_z = \sum_{i,j} \gamma_{ij}^z \tag{3}$$

$$\alpha_i^z = \frac{\sum_{i} \gamma_{ij}^z}{\delta_z} \tag{4}$$

2.2　挖掘主题短语

由于锚词有较好的主题解释性[14],其推断也以词共现为统计基础. 因此,为改善挖掘的短语质量,利用学习到的参数 γ_{ij}^z,δ_z 和 α_i^z 进一步抽取主题 z 中锚词. 假定每一主题至少包含一个锚词,s 表示主题 z 已有的锚词集合 $s = \{s_1, \cdots, s_n\}$,$O_{i,\neg}$ 表示已知 w_i,属于主题 z 的边中含词 w_j 的概率,则通过归约参数 γ_{ij}^z 中的词 w_j 计算出 $O_{i,\neg}$ 后,主题 z 能用锚词 s_i 表示. 因此,锚词抽取的关键是重构,即以线性方式组合锚词表示非锚词. 若用 C 表示重构系数矩阵,C_{iz} 表示词 w_i,非锚词属于主题 z 的概率,则 $C_{iz} = \sum_j \gamma_{ij}^z / \delta_z$,$O_{i,\neg} \approx \sum_{s_i \in s} C_{iz} O_{s_i,\neg}$. 通过锚词抽取方法找到表示主题 z 的锚词 s 后,主题短语挖掘转换为通过主题 z 中的锚词,获得主题频率大于给定最小支

持度阈值且包含锚词的主题频繁模式. 为此,先用关联规则挖掘算法 FP-growth 挖掘频繁模式集 A^z,其次用锚词短语完整性标准修剪 A^z,最后对照已得到的锚词,删除 A^z 中不含锚词的主题短语. 由于同一主题可用多个锚词短语描述,为了选择最合适的锚词短语,一方面,本文把锚词短语定义为不同锚词的任意序列,并借鉴文献[4]的思想对其进行排序.

3 实验结果与分析

3.1 实验设置

文章基于知网采用跨库检索,收集了 2000 年到 2015 年之间,与数据库、数据挖掘、信息检索、机器学习和自然语言处理等关键词相关的文献标题. 采用 ICTCLAS2016 分词包对标题分词,哈工大的停用词表等预处理技术,最终得到包含 143,2567 个短文本,185,986 个不同词的短文本集作为第 1 个实验数据集 DataSet_ZW. 以建筑、文学、自动化,计算机,法学和商学共 6 个类别的书籍标题作为第 2 个实验数据集,该数据集 DataSet_BL 共包含 23,2176 个短文本,25,6563 个不同词.

本文从主题层次质量分析、主题层次对主题短语质量的影响以及主题层次对分类的影响三个方面对提出的方法进行了评测. 参与比较的方法分别是 CATHY[4] (Constructing A Topical HierarchY) 和 hPAM[15] (hierarchical Pachinko Allocation Model),它们在构建主题层次时也对短语施加了排序的策略,与本文提出的 *THM* 框架具有可比性.

3.2 主题层次质量分析

主题层次质量分析主要评估 THM 构建主题层次的能力. 由于 hPAM 只能构建 3 级的主题层次,所以 3 种方法在数据集 DataSet_ZW 上只构建 3 级主题层次. 设置根结点的子主题数量为 5,非根结点的主题数量为 4. 表 1 给出了 3 种方法构造主题为“大数据分析挖掘”的主题层次中第二层非根结点的子集. 观察表 2 发现:无论是从子主题的角度,还是从锚词短语或短语的角度,THM 与 CATHY 构建的主题层次都能表示“大数据分析挖掘”的某一领域和其子领域,能较清晰地表达双亲孩子结点关系. hPAM 方法构造的主题层次不能直观反映双亲孩子结点关系,只能结合其输出的词才能表达一个主题. 因此,THM 与 CATHY 构造的主题层次中的短语质量优于 hPAM.

此外,观察 THM 和 CATHY 挖掘的主题发现:(1) THM 发现的主题是对 CATHY 发现的主题的进一步凝练. 例如:“数据分析”主题中,THM 得到的子主题之一“数据处理”,能包括 CATHY 方法中“大数据分析”的任一子主题. 可能的原因是:THM 发现表示主题的短语时,先找表示主题的锚词,然后运用锚词对频繁短语修剪,因此,删除了主题中更详细的研究点. (2) THM 发现的主题是 CATHY 发现的主题中要使用的技术或方法. 例如:THM 发现的子主题“机器学习”或“数据挖掘”是 CATHY 发现的前 3 个子主题“大数据分析、大数据系统、移动计算”中必须要使用的技术. 这种现象说明先找到主题的锚词,可以为研究者提供大数据分析挖掘使用的具体技术. (3) THM 方法挖掘出的主题并不都能准确的找到锚词. 例如:“机器学习”主题中的 4 个短语中,组成短语的所有词都不是该主题的锚词. 这表明假设每一主题中至少存在一个锚词在实际应用中可能存在偏差.

表 1 “大数据分析挖掘”的某一主题层次子集

		子主题			
		数据分析	数据挖掘	机器学习	数据查询
THM	锚词短语	数据处理 数据检索 数据存储 数据抽取	主题挖掘 信息挖掘 图挖掘 分布式挖掘	特征提取 深度学习 主动学习 神经网络	位置查询 轨迹查询 SQL 语句 Top-k
		子主题			
		大数据分析	大数据系统	移动计算	数据查询
CATHY	短语	可视化分析 社交媒体分析 众包计算 时空轨迹数据	Spark Hadoop 分布式系统 数据库系统	无线传感网络 可穿戴 位置查询 隐私保护	复杂空间查询 skyline 查询 Top-k 数据清洗
		子主题			
		信息	数据	知识	查询
hPAM	词	数据 应用 系统 案例	网络 新闻 数字 媒体	基于 系统 推理 表达	检索 评估 相关 文本

3.3 主题层次对主题短语质量的影响

为了分析层次高度对主题短语质量的影响,将提出的方法与 CATHY 进行了比较. 评估的性能标准是公式(5)表示的准确率(P)、查全率(R)和 F 值.

$$P=\frac{N_{correct}}{N_{extracted}},R=\frac{N_{correct}}{N_{standard}},F=\frac{2PR}{P+R} \tag{5}$$

其中,$N_{correct}$ 表示由 THM 或 *CATHY* 抽取的正确的主题短语数量,$N_{extracted}$ 表示由 PDLDA[10] (Phrase-Discovering Latent Dirichlet Allocation) 模型抽取的主题短语数量,$N_{standard}$ 表示人工标注的主题短语数量. 选择 PDLDA 模型输出的主题数量作为短文本集中所有主题数量的原因:(1) PDLDA 模型输出的短语由连续的词构成,符合传统意义上短语的定义,所以 PDLDA 模型发现的短语肯定比 THM 和 CATHY 方法挖掘的短语要多;(2)

PDLDA 模型用短语表示主题,对主题有较好的解释性.

实验过程中先获取 PDLDA 模型挖掘出的主题和属于这些主题的短语;然后,分别按照不同的主题层次,执行 THM 或 CATHY 方法,并将其获得的所有子主题与 PDLDA 模型挖掘出的主题进行准确率、查全率和 F 值的比较分析. 表 2 列出了不同的层次高度对主题短语抽取质量的影响. 观察表 3 可知:(1)随着主题层次的增加,准确率、查全率和 F 值都呈现增加趋势;(2)在主题层次为 3 时,CATHY 方法优于 THM 方法;但在主题层次为 4 和 5 时,THM 方法优于 CATHY 方法. 原因是 THM 方法采用了锚词来寻找最具代表性的主题短语后,在递归构建层次更高的主题后能发现更详细的子主题.

表 2　主题层次对主题短语质量的影响

方法	主题层次	准确率(P)	查全率(R)	F 值
THM	3	0.268	0.330	0.296
	4	0.282	0.348	0.312
	5	0.333	0.350	0.313
CATHY	3	0.276	0.340	0.304
	4	0.280	0.345	0.309
	5	0.284	0.173	0.227

3.4　主题层次用于分类预测

为评估构建的主题层次对分类预测准确性的影响,为数据集 DataSet_BL 设计了一个分类预测实验. 分类预测的准确性采用第 k 个位置上的正确率 $P@k$ 作为评价标准. 因此,基于锚词短语的覆盖面标准构建了一个 5 层的主题层次,用于度量顶层分支在 k 处(k 表示前 k 个短语的位置)的覆盖与分类 $P@k$. 为计算 k 处的覆盖 $P@k$,事先需要对每一主题中的前 k 个短语进行标注,并检查是否有短语出现在书籍标题中. 若用 $p(t,c)$ 表示锚词短语出现在某一类书籍中的联合概率,$p(t)$ 表示锚词短语出现在主题 t 中的概率;$p(c)$ 表示书籍属于类别的概率,则可用公式(6)计算 k 处的覆盖与分类 $P@k$,其中,t 表示锚词短语属于主题 t,c 表示书籍所属类别.

$$P@k = \sum_{t,c\wp} (t,c)\log_2 \frac{p(t,c)}{p(t)p(c)} \qquad (6)$$

图 2 给出了每一种方法的 $P@k(k\in[1,60])$. 由于 $P@k$ 既考虑了锚词短语的覆盖面,也考虑了它与一个分类的联合概率,所以 $P@k$ 的值随着 k 的增大而变大. 从 3 条 $P@k$ 曲线可知,THM 与 CATHY 的分类的区分度优于 hPAMk. hPAM 的 $P@k$ 曲线几乎立即渐近为一条直线. 这是因为 hPAM 用词表示短语,覆盖面高. 另外,THM 的性能在约为 20 个短语之前的性能低于 CATHY,之后的性能优于 CATHY. 其原因是前 20 个锚词短语处于 THM 方法构建的主题层次的较低层,而其后的短语处于主题层次的较高层,进一步体现了锚词在主题挖掘中具有重要作用.

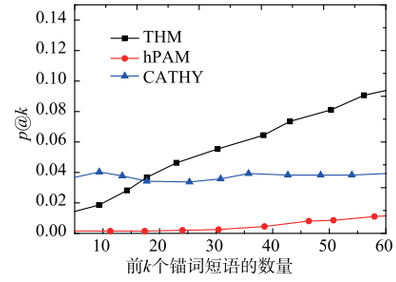

图 2　前 k 个锚词短语对 $p@k$ 的影响

4　结语

通过词共现图技术与锚词抽取技术的理论研究,提出了一种融合锚词抽取的海量短文本主题层次挖掘框架,用于提高主题挖掘的质量和改进基于主题的预测分类准确度. 在 2 种不同的短文本集上评估了提出的方法,实验结果表明,相比于已有的 hPAM 和 CATH,该方法构建的主题层次、挖掘得到的主题短语在质量方面均得到了改善;该方法构建的主题层次用于分类预测时的准确性也有明显的提升.

参考文献

[1] 周炫余,刘娟,卢笑,邵鹏,罗飞. 一种联合文本和图像信息的行人检测方法[J]. 电子学报,2017,45(1):140-146.
ZHOU Xuan-yu, LIU Juan, LU Xiao, SHAO Peng, LUO Fei. A method for predestrian detection by combining textual and visual information[J]. Acta Electronica Sinica, 2017,45(1):140-146. (in Chinese)

[2] 廖律超,蒋新华,邹复民,贺文武,邱淮. 一种支持轨迹大数据潜在语义相关性挖掘的谱聚类方法[J]. 电子学报,2015,43(5):956-964.
LIAO Lu-chao, JIANG Xin-huo, ZOU Fu-min, HE Wen-wu, QIU Huai. A Spectral clustering method for big trajectory data mining with latent semantic correlation[J]. Acta Electronica Sinica, 2015,43(5):956-964. (in Chinese)

[3] 国琳,左万利. 基于兴趣图谱的用户兴趣分布分析及专家发现[J]. 电子学报,2015,43(8):1561-1567.
GUO Lin, ZUO Wan-li. Analysis of user interest distribution and expert finding based on interest graphs[J]. Acta Electronica Sinica, 2015, 43(58):1561-1567. (in Chinese)

[4] Chi Wang, Marina Danilevsky, Nihit Desai, et al. A phrase mining framework for recursive construction of a topical hierarchy[A]. The 19th ACM SIGKDD International Confer-

ence on Knowledge Discovery and Data Mining[C]. Chicago:ACM,2013. 437 – 435.
[5] Zhao Wayne Xin, Jiang Jing, He Jing, et al. Topical keyphrase extraction from twitter[A]. The 49th Annual Meeting of the Association for Computational Linguistics[C]. Portland:ACL,2011:379 – 388.
[6] Grineva M, Grinev M, Lizorkin D. Extracting key terms from noisy and multitheme documents[A]. Proceedings of the 18th International Conference on World Wide Web [C]. 2009. 661 – 670.
[7] Tomokiyo T, Hurst M. A language model approach to keyphrase extraction[A]. In Proc. ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18[C]. Sapporo:ACL,2003. 33 – 40.
[8] Liu Z, Huang W, Zheng Y, et al. Automatic keyphrase extraction via topic decomposition[A]. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing [C]. Massachusetts: ACL, 2010. 366 – 376.
[9] Blei D M, Laerty J D. Visualizing topics with multi-word expressions[J]. Statistics arXiv:0907. 1013vl[stat. ML], 2009:1 – 12.
[10] Robert V. Lindsey, William Headden, Michael Stipicevic. A phrase-discovering topic model using hierarchical pitman-yor processes[A]. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning [C]. Jeju:ACL,2012. 214 – 222.
[11] Wallach H M. Topic modeling:beyond bag-of-words[A]. Machine Learning, Proceedings of the Twenty-Third International Conference [C]. Pittsburgh: ACM, 2006. 977 – 984.
[12] Wang X, McCallum A, Wei X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval[A]. Proceedings of the 7th IEEE International Conference on Data Mining[C]. Omaha:IEEE,2007. 697 – 702.
[13] Kim S N, Kan M-Y. Re-examining automatic keyphrase extraction approaches in scientific articles[A]. Proc Workshop on Multiword Expressions:Identification, Interpretation, Disambiguation and Applications[C]. Singapore:ACL,2009:9 – 16.
[14] Thang Nguyen, Yuening Hu, Jordan Boyd-Graber. Anchors Regularized:Adding Robustness and Extensibility to Scalable Topic-Modeling Algorithms[A]. Proc of the 52nd Annual Meeting of the Association for Computational Linguistics[J]. Baltimore:ACL,2014. 359 – 369.
[15] D. Mimno, W. Li, A. McCallum. Mixtures of hierarchical topics with pachinko allocation[A]. The 24th Annual International Conference on Machine Learning[C]. Oregon: ACM,2007. 633 – 640.

作者简介

吕 品 女,1973 年 3 月出生,湖北鄂州人,现为上海电机学院副教授、博士,研究方向为数据挖掘、观点挖掘与情感分析.
E-mail:lvp@ sdju. edu. cn

计春雷 男,1964 年 1 月出生,上海人,现为上海电机学院教授、博士、硕士生导师,研究方向为大数据、数据挖掘.
E-mail:jicl@ sdju. edu. cn

汪 鑫 男,1978 年 3 月出生,安徽黟县人,现为上海电机学院讲师、硕士,研究方向为数据挖掘、云计算.
E-mail:wangx@ sdju. edu. cn

罗宜元 男,1986 年 9 月出生,河南信阳人,现为上海电机学院副教授、博士,研究方向为密码学与计算机安全.
E-mail:luoyy@ sdju. edu. cn