DOI:10.13530/j.cnki.jlis.156008

基于图挖掘的文本主题识别方法研究综述*

郭红梅　张智雄

摘　要　本文通过文献调研分析,将基于图挖掘的文本主题识别方法总结为中心度方法、紧密关联子图查找和图聚类三种,后两者又细分为基于 clique 子团或类 clique 子团、基于图拓扑结构或结点属性聚类的方法。中心度方法通过对比文本网络中术语结点的重要度来实现文本主题的识别,紧密关联子图查找和图聚类方法则是根据文本图中术语结点和边的属性相似度来识别文本核心主题。基于语言文本网络自身特性,如何构建复杂文本关系图来同时揭示术语间的句法、共现和语义关系,如何基于术语关联和图拓扑结构识别其中的紧密关联子团,基于何种标准将紧密关联子团聚类以揭示文本核心主题,都是未来需要进一步深入研究的问题。表 1。参考文献 50。

关键词　文本主题识别　图挖掘　中心度　Clique 子团

分类号　G252.8

Methods of Text Theme Identification Based on Graph Mining

GUO Hongmei & ZHANG Zhixiong

ABSTRACT

With the development of the internet, electronic text is booming. These text resources, especially scientific journal papers, contain rich semantic and linked information. How to demonstrate the core topics quickly and accurately to assist researchers and improve research efficiency has been an urgent issue in text mining. Nodes and edges of graph can represent terms and their relations of texts, so many researchers tried to combine graph mining with natural language processing to identify text theme. This paper investigated and analyzed the studies and summarized their advantages and disadvantages in order to provide a reference for further research.

At present, the studies focus on textual representation of relation graph, theme identification based on centrality and subgraph detection or clustering. The method of theme identification based on cohesive subgraph detection mainly is to recognize clique or quasi-clique subgraph to represent the core content of the texts. Theme identification based on graph mining uses two methods: one is according to the graph topological structure, and the other considers graph topological structure and node attributes simultaneously. We mainly analyzed the clustering model, algorithm and evaluation criterion of clustering result. The methods of frequency statistics and external dictionary are relatively mature and often used as

* 本文系国家自然科学基金项目“基于语言网络的文本主题中心度计算方法研究”(编号:61075047)的研究成果之一。(This article is an outcome of the project “A study on textual topic identification by centrality algorithms based on language network” (No. 61075047) supported by National Natural Science Foundation of China.)

通信作者:张智雄,Email:zhangzhx@mail.las.ac.cn,ORCID:0000-0003-1596-7487 (Correspondence should be addressed to ZHANG Zhixiong, Email:zhangzhx@mail.las.ac.cn, ORCID:0000-0003-1596-7487)

benchmark. Centrality methods have been greatly improved, but the algorithm efficiency still needs to be improved. The methods based on graph mining have already shown advantages and are worth deeper exploration.

Language network of text has its unique characteristics. Various relations exist between terms, for example, co-occurrence relation, syntactic relation and semantic relation. How to construct complex text network which can reveal the relations of terms at the same time is one of the research directions in the future. Further studies need to address how to identify cohesive subgraph in complex text network according to relations between terms and topological structure of graph. In addition, the measure according to which these subgraphs are clustered to reveal core sub-themes and the relations of themes in texts also needs to be discussed. 1 tab. 50 refs.

KEY WORDS

Text theme identification. Graph mining. Centrality. Clique sub-group.

0 引言

随着网络技术的快速发展,电子化文本数量激增。越来越多的机构将研究成果以电子化文本形式呈现,使其成为信息传播的重要媒介。这些电子化文本,尤其是科技期刊文献中,蕴含着丰富的语义和主题信息,是重要的知识载体,同时也蕴含丰富的有情报价值的主题内容。很多学者尝试从不同角度对文本主题识别进行研究,以辅助科研人员快速把握文本主题,提高科研效率。

基于词频、拟文档词频、共现频次等统计方法,是最简单、使用最广泛的主题识别方法,主要依赖术语的频次和在文档集中的分布情况来识别文本主题[1],只是将文本模拟为所谓的语言包,并未考虑文本术语之间的关联,难以全面揭示文本中蕴含的丰富主题信息。近年来提出的各种文本挖掘和管理算法,如自动摘要、聚类、分类、标引或相似搜索等,大都是基于向量空间模型,仅将文本简单模拟为"词汇包",主要依赖词频及词在文本中的分布规律来对文本主题进行揭示,而对词序、句法及语义关系考虑较少[2]。以概率潜在语义索引模型(Latent Dirichlet Allocation,LDA)为代表的主题模型,利用软聚类方法来识别文本主题,依据概率值将术语分配到不同的主题中。但是该方法需要预先设定经验值,只能揭示术语之间的潜在语义关联。图是一种重要的可视化分析工具,利用网络或图中的结点和边可清晰反映网络中的对象及关系。电子化文本,尤其是科技文献是由大量相互关联的术语构成,蕴含着丰富的句法及语义关系信息。不少学者尝试将图挖掘技术与自然语言处理相结合,抽取文本中包含的概念或术语,将其作为结点,将它们之间的关系作为边,把文本表示为语言网络或术语关系图,借助社会网络或图挖掘方法对文本语言网络进行分析,查找文本关系网络中的核心结点和重要关联通路或紧密子团,从而识别文本中所蕴含的重要子主题和各子主题间的关联结构。

基于文本关系图进行主题挖掘的实质是,依据图中结点和边的属性识别图中核心的术语或关联子团,以揭示文本中的主旨内容。目前研究主要集中在两个方面:一是选取网络中的核心术语结点作为文本中的核心子主题,通常选取高中心度的点或在网络中起关键作用的桥结点;另一种是通过识别网络中的重要路径或紧密关联子团,识别文本中所蕴含的重要主题以及主题之间的演化。通过对国内外基于图挖掘的主题识别方法和技术进行广泛调研和深入分析,本文将其分为基于图中心度的主题识别方法、基于子图查找的方法和基于图聚类的方法。

1　文本关系图构建的相关研究

早期数据挖掘和知识发现研究，主要关注严格和明显结构化数据的处理方法。但随着网络技术的发展，产生大量结构复杂甚至非结构化的数据，如何对这些数据进行处理分析，挖掘其隐含的重要信息和模式，日益受到关注。很多学者对数据挖掘方法进行改进创新以提高性能，其中最重要的就是将结构复杂的数据利用图来表示，基于图挖掘技术从中识别重要的主题或知识。

文本关系图的构建是利用图或网络分析方法进行主题挖掘的基础，初期只是将文本中的术语及其之间的共现、句法或者语义关系表示为简单的语言网络，大多只是对文本中的一种关系进行分析。最常见的是基于句子共现关系，抽取文本中有意义的术语或概念，构建共现关系图，虽然易于构建，但只能体现词在位置或语境上的关联，并不能揭示隐含的语义关系。为全面表征文本中所蕴含的知识信息，越来越多的学者尝试将文本转化为具有更多结点和边属性的复杂关系图，以对文本主题进行详尽分析。

Martin 提出利用文本中抽取的术语来构建概念图，边代表术语之间的语义关系，克服了向量空间模型中关键词独立的缺陷，充分利用自然语言展示文本中潜在的概念关系，较之前基于特征和基于结构的知识发现有了更大的进步[3]。Chau 等提出了概念链图（concept-link graph），图中不仅包含文本内容，而且展示了概念之间的潜在语义结构，首先利用自组织方法对概念相关数据进行聚类得到概念集，然后将单篇文本内容与概念进行映射，利用奇异值分解方法根据概念发生的频次构建概念图，对文本语义结构内容进行可视化描述[4]。Popping 等利用知识图概念来表示文本实体关系，它是一种特殊的语义网络[5]。Aggarwal 提出距离图（distance graph），从句子粒度上对文本进行分析，图中保存了文本术语的相对词序和距离，根据它所保存的距离信息来定义等级变量；距离图模型可转化为向量空间模型的结构化视图，可使用目前存在的所有文本处理工具，并不需要开发新的算法和工具。此外，XML 格式的数据也可以直接利用该模型进行分析，较之前向量空间模型有更广泛的使用范围，但文中作者只是从理论上证明该方法的有效性，并未用具体实验进行验证[6-7]。Dmitry 等提出文本有意义循环路径（pathways for meaning circulation），将文本中数据关系利用图的形式进行可视化展示，基于网络分析查找图中重要概念组成循环路径，展示文本主题演化过程，揭示核心主题[8]。Malioutov 提出了图论框架模型，文本转化为无向加权图，结点表示句子，边量化这些关系，利用归一化切割标准对文本进行分割，通过考虑文本中更长范围词汇凝聚和分布，扩大了滑行窗口的局部凝聚力范围。根据这个标准，每个分割部分内部相似度最大，差异性最小[9]。Diesner 等抽取文本中潜在的社会和组织结构来进行文本主题识别，他首先将概念之间的关系表示为网络，将人等实体概念作为网络中的核心实体，分组到不同子概念网络中，以揭示文本社会结构[10]。有一些学者尝试对文本中的语义关系网络进行分析以更深层次揭示文本主题，Popping 等主张将文本语义关系表示为知识图，可更好理解文本数据，把握知识的动态变化[5]。

2　基于网络中心度的文本主题识别方法

随着信息抽取、网络分析以及中心度等方法的发展，学者不断提出新的思路和解决方法，其中最重要的一条思路是计算文本中各主题的中心度，根据中心度来区分主题的重要性，进而识别出核心主题以及主题之间的结构关系。目前常用的中心度测度方法有点度中心度、接近中心度、中介中心度和特征向量中心度。点度通常代表结点的频次，点度越高说明与网络中其他结点的联系越紧密，处于网络中的中心地位，在网络中起着关键的连接作用，代表文本中

的核心术语,可用于标识文本中的重要主题。接近中心度和中介中心度是基于网络连通性得到的,通过查找图或网络中的重要通路来揭示术语和主题之间的关联。特征向量中心度是网络中心势的一种标准化测度,用于查找网络中的核心结点。

基于网络中心度进行文本主题识别,首先将文本划分为不同文本单元,如句子、词或短语等,作为网络节点,各文本单元的关系作为网络的边,构建文本语言网络;而后基于构建的语言网络,运用中心度算法测度网络中各结点的重要度或重要通路,识别核心主题和主题的演化路径。Coursey 等通过所提出的中心度算法计算出外部维基百科概念图中的结点与所输入文本的相关度,从而进行文本主题的识别[11]。研究者发现,单一中心度方法并不能充分揭示文本核心主题,有学者尝试选取多种指标进行加权来计算主题重要度。吴思竹等通过对术语间多重关系进行修剪、融合,构建语言网络模型,提出了主题角色识别的多指标体系,从而识别文本中的重要主题和术语[12]。Zhao 等提出主题导向的子团识别方法,联合使用结点聚类与连接分析,识别网络中重要对象和群组[13]。Kas 等联合使用多重与概念网络结构有关的指标,如中心性、排他性、密度和聚类等,识别网络中的重要概念和主题或概念之间的重要关联[14]。实证结果表明,多指标评价体系较单一中心度方法更具优势。

3 基于紧密关联子图查找的主题识别方法研究

目前,子图挖掘方法多是基于图的拓扑结构或结点属性从文本关系网络中查找重要子图,构成子图结构的各结点联系紧密,可以反映文本中内容紧密关联的术语簇,通过对术语间各种关系的揭示来进行文本中核心主题的识别。关联子团识别是社会网络分析中的重要问题,不少学者将其应用在文本关系网络中,主要集中于对网络连接或拓扑结构的分析,对关联子团内部结构的分析有助于读者更深入理解文本内容,揭示核心主题。

Clique 子团是一种特殊的子图结构,要求每个结点都与子图中的其余结点直接相连,某种程度上可认为 clique 结点之间高度关联。这一特性使其成为查找网络或图中重要子团的非常重要的方法,它最早在社会网络分析中用于模拟密切关联群体,后来逐步扩展到图挖掘研究。

3.1 基于 clique 子团查找的方法

Clique 子团确保各顶点之间的完美可达性,是研究聚类和紧密子图的基础,目前已应用到很多学科领域。社会网络分析中常基于 clique 来研究企业间的发展网络、学校之间的友情网络、协作团队之间的合作网络等,在生物信息学领域,常基于 clique 子团识别蛋白结构,如从蛋白与蛋白反应网络中发现蛋白复合体,还有基于最大 clique 子团进行图像识别的研究[15]。Zheng 等基于最大 clique 子团启发式方法来识别昆虫基因图中的常见基因子团网络,利用子图异构方法对多物种基因网络进行分析,较之前的单物种的加权关联网络分析、贝叶斯网络、自动回归模型、图高斯模型等具有更好的识别结果[16]。组织间网络结构是提高人类组织协调和合作的方式,Ngamassi 等研究 clique 结构在人道主题组织信息交互网络中所发挥的作用,结果表明,人道救援领域与公共卫生服务领域相似,都可利用网络集成和网络派系来解释网络的有效性[17]。这些研究通过查找网络或图中的 clique 来实现对核心团体、核心子团、核心对象的识别。不少学者尝试将 clique 子团识别方法应用到文本挖掘领域中,通过查找文本关系图中的紧密 clique 子团结构揭示文本中所蕴含的核心子主题。

3.2 基于类 clique 子团查找的方法

Clique 是图论中的经典模型,它要求所有结点必须两两相连,其所保证的理想凝聚特性限

制了它的应用。实际网络中有些非常紧密的结构可能在数据收集出现差错时导致某些边的缺失;一些很重要的具有很多结点的大型网络,由于边的稀疏在 clique 聚类过程中会产生大量无意义的类;图中最大 clique 的查找和最小 clique 的分类在计算上一直很具有挑战性。鉴于以上原因,在不少应用领域如社会网络分析以及计算生物学中,研究者提出了 clique 弛豫模型(clique relaxation models),常见的有 s-plexes、s-clubs、g-quasi-cliques、k-community 等[18-19]。其中,比较有代表性的是 k-clique 渗透元组(k-Clique Percolated Components, k-PCs)[20-21],它是一些至少包含 k 个顶点 clique 子团的集合,这些 clique 子团之间可通过一系列邻接的 clique 子团相连通,即数据集中的每个 k-clique 都可以通过其余重叠的 k-1 个顶点相连,即每个顶点都可通过一组紧密联系的结点相连(k-clique)。Quiniou 等基于 k-clique 渗透元组对大型文本主题进行识别,通过设定 k、α、γ来限定主题中所包含顶点的个数以及各顶点间的关联度,其中 k 为 k-PCs 中所包含的 clique 中至少包含的顶点数,α为同一主题中各顶点最小共享的属性数,γ为主题中所包含的 k-PCs 的个数[20]。Boginski 等将弛豫 clique 模型应用在市场中,结合加权市场图模型和相关标准提出选择投资组合的新框架来识别股份集群,基于弛豫 clique 模型进行聚类得到候选股票群组[22]。Palla 等将 k-clique 渗透元组应用在 E-R 图中来发现大型网络中的重叠群体[23]。Fan 等研究了在网络演化中 clique 凝聚子群的相互渗透情况[24]。目前,clique 查找算法大多需要很大的计算负载和内存资源,Gregori 等从并行优化的角度出发,研究如何提高 clique 渗透元组查找的效率,提出了 FLIP-CPM 并行探测方法[25]。Clique 弛豫模型既继承了 clique 凝聚子团的优点,又没有像 clique 那样严格的条件,适用范围更为广泛。

4　基于图聚类的主题识别方法研究

图聚类与传统的关联数据聚类的最大区别在于,它是基于连通性和结构相似度来测度结点的连接性(如两节点之间可能的路径数),而关联数据聚类主要基于属性相似度来测度距离(如两属性结点之间的欧氏距离)。对于规模较大的图来说,里面可能含有多个紧密关联子团,如何进一步将其分组来提取图中几个重要的部分,也是文本挖掘常面临的问题。早期紧密关联子团的查找大多依赖图的拓扑属性(如直径、聚类系数等)和局部模式规律来完成。近年来,随着聚类算法和稠密子图挖掘方法的提出,为保证所抽取出的子团在结构和内容上都紧密相关,不少学者尝试将文本关系图中结点和拓扑属性结合起来,使所识别出的子主题在结构上关联,在内容上同质。

4.1　基于图拓扑结构的聚类方法

如何尽可能将图中同性质的结点或边聚到一组是图分析研究中最常遇到的问题。Polanco 等应用图聚类方法对文本共现关系图进行分析,以实现文本知识发现,该聚类方法并不局限于同种类,更多强调强关联路径中的不同种类[26]。Zhang 等基于 clique 聚类生成生物医学文本中的重要摘要,文中选取心脏病、帕金森病等 11 个主题文本集,利用 semrep 抽取生物医学文本中的语义关系,利用概念表结点、概念之间的关系表边构建文本关系图,利用 ucinet 工具抽取图中所包含的 clique,利用 k-mean 对 clique 进行聚类得到文本集的重要主题,并根据各主题中高频词汇和关系对主题进行标记,利用可视化的方式展示文本中重要的主题结构[27]。Zubcsek 等利用 clique 重叠模型来识别封闭网络内部关系较强的信息子团,其效果优于一般的网络模型[28]。Ah-Pine 等提出了基于 clique 的聚类方法以标注命名实体[29]。Wang 等提出通过 k-clique 聚类来发现、识别和评价网络舆论领导社区[30]。Clique 的大小由其中所包含的顶点数决定,有些学者提出利用约束模型限制类中 clique 的数量以及 clique 子团中所包含的顶点数。Ji 等利用分支和成本框架解决每个类中的

顶点个数问题,其中成本问题是利用整数规划法解决的[31]。Jaehn 等利用分支界定法解决 clique 分区问题,他基于三角约束提出上限值,利用约束性条件传递方法得到很多固定的边,然后再利用二分法得到新的分支框架[32]。Mougel 等提出利用最大同质 clique 集(Maximal Homogeneous Clique Sets,MHCSs)概念识别图中同性质的关联子团[33]。Krems 等研究了在超链接影院网络中 clique 子团大小与网络特征的关系[34]。Gjoka 等从概率角度研究 clique 的分布,结果表明它与单个 clique 规模大小以及结点属性有关[15]。

仅基于图拓扑结构进行核心结点或子团查找的方法更多强调结点之间在结构上的关联,并未揭示结点在实质内容上是否相同或相似,因而对文本主题揭示不充分。比如在生物医学领域,疾病和药物之间既可以是治疗关系,也可以是副作用的关系,如果仅基于结点之间的关联进行聚类,很有可能将表示治疗和副作用的内容聚类到同一主题中,造成聚类误差。

4.2 基于图拓扑结构和结点属性的聚类方法

目前,存在的图聚类方法大多是基于图的拓扑结构,并未考虑结点之间的多种属性。在很多实际应用中,图拓扑结构和结点属性都是很重要的,如在社会网络中,结点属性描述一个人的角色,而拓扑结构表示不同组别中人与人之间的关系。理想上,图聚类后的结果应该是高凝聚度的类中的各结点具有较多相同或相似的属性,在结构和属性相似度上要具有均衡性,即将结点属性与紧密子团结构建立关联。国内外很多学者在同时基于拓扑结构和结点属性进行图聚类方面进行了很多研究,并提出了值得借鉴的模型和算法。

4.2.1 聚类模型构建的相关研究

目前,基于 clique 子团聚类识别文本主题的研究通常仅将基于 clique 子团间共享结点的情况进行主题分析,并未从共享结点属性是否同质的角度,分析聚类到同一主题的 clique 子团是否在内容上一致。鉴于上述不足,不少学者尝试对聚类模型进行修正,以兼顾结点在内容和结构上的关联。

Moser 等利用图中边结构和结点特征向量信息的互补来抽取共享多个特征向量的子图结构,凝聚模式是满足属性约束、密度约束和连通性约束的子图结构[35]。Silva 等提出,利用结构关联模式挖掘(structural correlation pattern mining)方法来抽取图中具有高频属性集的类 clique 凝聚子团[36],属性集的结构关联度和结构关联模式之间相互补充,作者定义了结点属性集关联度函数,从结点规模和图密度两方面评价结构关联模式的优势,并给出了具体算法的实现过程,选取学术合作、音乐创作以及引文三个实际的关系属性图来验证该方法的可行性和算法的有效性。Ge 等提出了聚类模型 Connected k-Center(CkC),该模型同时考虑属性和关系两种数据,通过常数因子近似算法降低问题的复杂性和自然语言处理的难度,通过实验证明层次聚类算法在大规模数据处理过程中的有效性和可行性[37]。Miyoshi 等对凝聚图做进一步扩展,加入了对定量属性的考虑[38]。Berlingerio 等在以往属性图中子图模式挖掘的基础上又加入了时间属性,挖掘能表征图演变规律的子团模式,但该方法并没有定义如何在同一模式下获取子图结构[39]。Crémilleux 等定义了一个在多种约束条件下抽取子模式的通用框架,通过挖掘子图模式的微阵列数据探索数据源之间的交互[40]。Nowicki 等提出的随机 Blockstructures 模型基于关系数据对发现网络中潜在的结构、群组或类,这些关系对可以有向也可以无向[41]。Wang 提出 Group-Topic(GT)模型,不仅考虑到对象之间的关系,也考虑到关系的属性,基于这些属性可更准确识别文本中的主题,它实际上是 Blockstructures 模型的扩展[42]。McCallum 提出 Author-Recipient-Topic(ART)模型,它建立在 LDA 和 Author-Topic(AT)模型之上,将关键属性加入每个主题中,基于网络结构来识别核心子团,但这些方法只能

反映连接强度，并不能充分揭示结点之间的语义关系[43]。学者在基于图聚类进行文本主题识别的研究中，依据文本关系图中的不同属性设计具体的聚类模型，为基于文本图进行主题识别提出了新思路。

4.2.2　聚类算法设计的相关研究

在属性图中通过局部模式挖掘将同质的结点进行聚类，主要有两种方式：一种要求模式是独立的紧密子图，如 clique 子团或者类 clique 子团；另一种要求模式是由多个紧密的子图组成的集合，集合中的结点共享相同的属性集。不同的研究者给出了不同的要求，如 Mougel 等对子图结构的限制非常严格，要求结点集必须是 clique[33]，Sese 等对子组的要求比较宽松，仅简单要求结点来自连接子图[44]。Cheng 等提出了 SA-Cluster 聚类算法，通过统一距离测度指标将结构和属性相似度结合起来，通过矩阵相乘来计算图中结点的随机行走距离，将具有属性的大图划分到 k 个类中，每个类与一个紧密子图相对应，该子图中的所有结点具有相同的属性值，在后期研究中为提高 SA-Cluster 算法的效率和可扩展性，又进一步尝试利用边权重递增来更新随机行走距离，提出了 Inc-Cluster 算法，并进一步探索紧密子图的挖掘和图匹配的相关研究[45]。Mougel 等提出在布尔属性图中将同质性 clique 渗透元组聚类的方法（Collection of Homogeneous k-clique Percolated components，CoHoP），并给出算法的具体实现过程，通过科学合作网络和基因交互网络证明该方法能抽取出有意义的结构模式[18]。Dorndorf 等提出发射链启发式方法和分支界线法[46]。Brusco 等基于嵌入式定位算法提出邻接搜索启发式方法[47]。随着越来越多的领域利用图结构来模拟数据对象之间的关系，如网页、社会网络、传感器网络、生物网络和交互网络等，研究者提出了适用于各网络的聚类算法，如基于归一化切分的聚类，基于模块化、结构密度或流的聚类和基于结点属性相似度的图分解方法[48-49]。

图聚类算法的提出使一些理论上合理的聚类模型得以实现，是对基于图聚类方法进行文本主题识别研究的进一步探索。在算法实现的过程中，可以对以往所提出的文本聚类模型进行修正，提高基于图挖掘进行文本主题识别的效果。

4.3　图聚类结果的选择和评价

由于潜在高度复杂的数据集在不同参数或不同聚类方法下得到的结果存在较大差异，最优聚类结果的选择还没有一个公认的标准。目前常用的方法有三种：一是基于信息熵或其他测度信息，判断两个类之间的交互信息；二是设置对比，如将来自第一类聚类结果的每类与第二聚类中的最相似类进行映射，计算召回率、准确率或其他指标；三是基于对象对分析，通过可视化分析对象对的差距。

虽然文献中记载了不少计量评价指标，但它们并不能给出全面准确的评价结果，而且也不能确定哪种方法效果更好。可视化方法虽能给出较多有用的判断信息，但目前的相关研究较少。Achtert 等提出环形片段可视化（circle segments visualization），可支持对聚类结果的比较，并且对所选用的聚类方法进行度量评价。环形片段可视化是基于不同方法下对象对的比较，既可对聚类过程有更深入的了解，又可与通常的聚类评价指标进行互补。文中的可视化是对聚类结果差异的定量可视化，因为它的复杂性要依赖类的差异和数量。它不仅适用于对大型、高纬度的数据集的可视化，还可对类的合并和分裂进行分析[50]。环形布局增加对聚类结果细节的说明，因为越外围的环，其碎片数量越多；同时该方法也支持对聚类细节的进一步探索，如类的合并或分组等，尤其是那些在一种聚类方案下丢失，而在另一种方案下出现的成对分割。Zhang 等通过训练数据集，从仅有三个单独类的结果行开始，根据 clique 子团的语义标签，在层次聚类冰柱图中，分析下一行的结果中按语义标签是否有合并的类，直至不再有可合并的标签结束[27]。但这种方法最终的聚类数是

一个固定的阈值,并不适用各种聚类分析。

5 现有研究方法的特点及未来研究思路

5.1 现有方法的特点

在对国内外文本主题识别方法充分调研的基础上,本文将其总结归纳为五大类:基于频次统计的方法,基于外部词典的方法,基于潜在语义索引的方法,基于中心度的方法和基于图挖掘的方法,并对各方法的优势、缺点以及未来的发展趋势进行分析,具体如表1所示。

表1 目前存在的文本主题识别方法对比分析

	代表性方法	优势	缺点	发展趋势
频次统计方法	词频统计 词间关系统计等	简单易行	未考虑词间关联,对重要的低频词揭示不足	发展成熟,常作为基础对比方法
外部词典	基于 WordNet 词表 基于 MESH 词表	较以往统计方法有很大改进,能很好地反映概念在词典的映射关系	脱离文本内容,易造成新词丢失	相对成熟,较统计方法有很大改进,目前常用于基础数据集的构建
潜在语义索引方法	LDA 模型	根据文档出现某词的概率进行主题识别	仅基于文本和词的潜在语义计算关联概率,很难将文本与所识别主题建立联系,参数的经验值难以获得	有待进一步提高
中心度方法	点度中心度 中介中心度 接近中心度	从网络的角度对文本进行分析,考虑了词间的多种关系	很多算法只适用于规模较小的无向图,对于大规模的复杂网络效率很低,甚至不能实现。对语义关系揭示得不充分	相对成熟,较统计和外部词典方法有很大改进,很多学者尝试对中心度算法进行改进来提高文本主题识别的效率
子图挖掘方法	紧密关联子图抽取 子图聚类	利用图结构可清晰反映术语间的关系,揭示文本主题间的关联	对子图规模和所识别主题的数量难以确定	已显优势,是一种可待深入探索的新的研究思路

在现有研究中,研究者利用词频、词间关系统计方法对文本主题识别的研究相对成熟,取得了较好的效果,但其孤立地考虑词的特征而忽略了其在文本中的相互影响,并且对低频但表示主题的词无法揭示。基于词间关系统计的方法虽然考虑了词间的联系,但是仅基于词序、共现关系,过于粗糙和模糊,并不能准确和全面揭示文本内容。虽然利用外部词典拓展了词间关系,能较好地反映文本中的概念在词典中的映射关系,但是脱离文本内容,并且一些未记录的词会导致识别的新词缺失,无法全面揭示主题。基于潜在语义索引的LDA方法能够扩大所识别主题的语义覆盖率,但是容易掺杂噪声,并且需要人工指定聚类系数,这种经验值难以获取。

基于网络中心度的方法较基于词频和词间关系的方法有很大改进,它综合考虑了文本中的多种句法、语义关系,代表性的指标有点度中心度、中介中心度和接近中心度。很多学者提出针对不同网络的中心度算法,但目前的研究大多是基于规模较小的无向图,对于规模较大的复杂网络算法效率很低,有的甚至不能实现,

而且并未同时考虑术语间的多种关系。

基于子图挖掘的文本主题识别方法已显优势，利用图中的结点和边不仅可清晰揭示文本中的复杂关系，而且凝聚子图也能表征文本中的凸显信息。况且属性图中模式的挖掘方法可将图的拓扑结构和结点属性关联起来，将具有相同属性的结点关联成一个紧密子团，能较好揭示多重文本图中的核心主题，是一种可待深入探索的新的研究思路，但是对主题的数量和规模大小如何确定还不明确。虽有研究者提出通过设定参数来实现，但该参数的取值也是人为主观确定的，而且该算法实现也很复杂。目前基于属性图的模式挖掘中，对结点属性个数和性质考虑得较少，大多是对同质的属性进行分析。文本关系图包含多重关系，而且术语结点及关系存在多种不同性质的属性，在聚类的过程中，不仅要考虑结点共享属性的个数，同时还要分析在属性个数相同的情况下，不同属性组合对聚类结果的影响。目前，对聚类结果的选取并没有统一的标准，在计量指标的基础上，还要结合定性的评价方法对聚类效果进行分析，选取最优的聚类方案。

5.2　未来研究思路

文本关系图除具有一般网络关系图的特征之外，还具有语言网络自身的特性。文本是由具有语义信息的术语按照一定的逻辑结构构成，这些术语之间除了在物理位置上关联外，还存在句法上的支配从属关系和隐含的语义关联。术语结点存在多种不同性质的属性，如表示术语自身性质的属性，词干、词性等，表示术语位置的属性，所属文本编号、句子编号、标题或摘要等，以及表征术语所属关系的属性。由于术语之间存在多种关系，如共现关系、句法关系和语义关系，相应的文本关系图中各边也存在多种属性。与其他关系属性图相比，文本关系图在属性数量和属性性质上都更为复杂，因此在核心主题识别过程中不仅要考虑图或网络自身的拓扑结构，还要分析结点属性和边属性在不同数量、不同性质，以及不同组合情况下所识别出的文本子主题之间存在的差异，以及产生差异的原因。

基于上述文本主题识别领域的研究背景和现有方法存在的不足，未来研究的思路为，抽取某主题相关论文集中的术语和术语间的共现、句法和语义关系，构建多重文本关系叠加模型，三种关系相互叠加补充，可避免文本信息的丢失。通过对叠加后的多重文本关系图的特点进行深入分析，抽取其中紧密关联的 clique 子团，根据 clique 子团间的相似性距离和结点之间共享属性的情况，将同主题的 clique 子团进行聚类，识别文本集中所包含的重要子主题。通过对多重文本关系图中 clique 子团的识别和聚类，揭示文本中的核心主题及其知识结构，挖掘和发现文本内潜在的知识和组织模式。基于 clique 子团聚类的文本主题识别方法，不仅可凸显文本集中结构和内容紧密关联的重要内容，而且基于这些凸显信息进行聚类，还可在不减弱文本集信息量的情况下降低计算复杂度。

参考文献

[1] Ryan G W, Bernard H R. Techniques to identify themes[J]. Field Methods, 2003, 15(1): 85-109.

[2] Ye C L. The research of theme identification in scientific documents[J]. Computer Science and Automation Engineering (CSAE), 2012(3): 715-718.

[3] Martin B, Eklund P W. From concepts to concept lattice: a border algorithm for making covers explicit[M]//Formal concept analysis. Berlin: Springer. 2008: 78-89.

[4] Chau R, Tsoi A C, Hagenbuchner M. A concept link graph for text structure mining[C]//Proceedings of the Thirty-Second Australasian Conference on Computer Science, 2009: 141-150.

[5] Popping R. Knowledge graphs and network text analysis[J]. Social Science Information,2003,42(1):91-108.

[6] Aggarwal C,Zhao P. Towards graphical models for text processing[J]. Knowledge and Information Systems,2013,36(1):1-21.

[7] Aggarwal C,Zhao P. Graphical models for text:a new paradigm for text representation and processing[C]// Crestani F,Marchand-Maillet S. The 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:ACM,2010:899-900.

[8] Dmitry P. Identifying the pathways for meaning circulation using text network analysis[EB/OL]. [2015-10-16]. http://noduslabs.com/publications/Pathways-Meaning-Text-Network-Analysis.pdf.

[9] Malioutov I, Barzilay R. Minimum cut model for spoken lecture segmentation[C]//Carpuat M, Duh K. Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2006: 25-32.

[10] Diesner J,Carley K M. Revealing social structure from texts[J]. Causal Mapping for Research in Information Technology,2004,81(3):65-72.

[11] Coursey K,Mihalcea R. Topic identification using Wikipedia graph centrality[C]//Chelba C,Kantor P,Roark B. Proceedings of Human Language Technologies:the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics,Companion Volume:Short Papers. Stroudsburg,PA:Association for Computational Linguistics,2009:117-120.

[12] 吴思竹. 基于语言网络的文本主题中心度计算方法研究[D]. 中国科学院文献情报中心,2011. (Wu Sizhu. A study on textual topic identification by centrality algorithms based on language network[D]. Beijing:National Science Library,Chinese Academy of Sciences,2011.

[13] Zhao Z Y. Topic oriented community detection through social objects and link analysis in social networks[J]. Knowledge-Based Systems,2012,26(1):164-173.

[14] Kas M,Carley K M,Carley L R. Trends in science networks:understanding structures and statistics of scientific networks[J]. Social Network Analysis and Mining,2012,2(2):169-187.

[15] Gjoka M,Smith E,Butts C. Estimating clique composition and size distributions from sampled network data[C]// Computer Communications Workshops(INFOCOM WKSHPS),IEEE Conference. Toronto,ON:IEEE,2014:837-842.

[16] Zheng G,Tesfay A,Huang X,et al. A clique-based approach to the identification of common gene association sub-networks[J]. Applied Mathematics,2013,4(6):893.

[17] Ngamassi L,Maitland C,Tapia A H. Humanitarian interorganizational information exchange network:how do clique structures impact network effectiveness?[J]. International Journal of Voluntary and Nonprofit Organizations,2014,25(6):1483-1508.

[18] Mougel P N,Rigotti C,Gandrillon O. Finding collections of k-clique percolated components in attributed graphs[M]// Fayyad U M,Piatetsky-Shapiro G,Smyth P, et al. Advances in knowledge discovery and data mining. Berlin:Springer,2012:181-192.

[19] Verma A,Butenko S. Network clustering via clique relaxations:a community based approach[J]. Graph Partitioning and Graph Clustering,2012,58(8):129-143.

[20] Quiniou S,Cellier P,Charnois T,et al. Graph mining under linguistic constraints to explore large texts[J]. Computación y Sistemas,2013,17(2):239-250.

[21] Derényi I,Palla G,Vicsek T. Clique percolation in random networks[J]. Physical Review Letters,2005,94(16):160-202.

[22] Boginski V, Butenko S, Shirokikh O, et al. A network-based data mining approach to portfolio selection via weighted clique relaxations[J]. Annals of Operations Research, 2014, 216(1): 23-34.

[23] Palla G. The critical point of k-clique percolation in the E-R graph[J]. Journal of Statistical Physics, 2007, 128(2): 56-72.

[24] Fan J, Chen X. Generalclique percolation in network evolution[EB/OL]. [2015-10-16]. http://arxiv.org/abs/1309.4535.

[25] Gregori E, Lenzini L, Mainardi S. Parallel k-clique community detection on large-scale networks[J]. IEEE Transactions on Parallel and Distributed Systems, 2013, 24(8): 1651-1660.

[26] Polanco X, Juan E S. Text data network analysis using graph approach[C]// Guerrero-Bote V P. I International Conference on Multidisciplinary Information Sciences and Technologies, vol. 2. Mérida, Spain: Open Institute of Knowledge, 2006: 586-592.

[27] Zhang H, Fiszman M, Shin D, et al. Clustering cliques for graph-based summarization of the biomedical research literature[J]. BMC Bioinformatics, 2013, 14(1): 182-191.

[28] Zubcsek P, Chowdhury I, Katona Z. Information communities: the network structure of communication[J]. Social Networks, 2014, 38(6): 50-62.

[29] Ah-Pine J, Jacquet G. Clique-based clustering for improving named entity recognition systems[C]//Association for Computational Linguistics. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2009: 51-59.

[30] Wang J, Jia X, Zhang L. Identifying and evaluating the internet opinion leader community through k-clique clustering[J]. Journal of Computers, 2013, 8(9): 2284-2289.

[31] Ji X, Mitchell J E. Branch-and-price-and-cut on the clique partitioning problem with minimum clique size requirement[J]. Discrete Optimization, 2007, 4(1): 87-102.

[32] Jaehn F, Pesch E. New bounds and constraint propagation techniques for the clique partitioning problem[J]. Discrete Applied Mathematics, 2013, 161(13): 2025-2037.

[33] Mougel P N, Rigotti C, Plantevit M, et al. Finding maximal homogeneous clique sets[J]. Knowledge and Information Systems, 2014, 39(3): 579-608.

[34] Krems J A, Dunbar R I M. Cliquesize and network characteristics in hyperlink cinema[J]. Human Nature, 2013, 24(4): 414-429.

[35] Moser F, Colak R, Rafiey A, et al. Mining cohesive patterns from graphs with feature vectors[EB/OL]. [2015-10-16]. http://epubs.siam.org/doi/pdf/10.1137/1.9781611972795.51.

[36] Silva A, Meira Jr W, Zaki M J. Structural correlation pattern mining for large graphs[C]//Proceedings of the Eighth Workshop on Mining and Learning with Graphs, 2010: 119-126.

[37] Ge R, Ester M, Gao B J, et al. Joint cluster analysis of attribute data and relationship data: the connected k-center problem, algorithms and applications[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2008, 2(2): 7-20.

[38] Miyoshi Y, Ozaki T, Ohkawa T. Frequent pattern discovery from a single graph with quantitative itemsets[C]// IEEE International Conference on Data Mining Workshops. Miami, Florida, USA, 2009: 527-532.

[39] Berlingerio M, Bonchi F, Bringmann B, et al. Mining graph evolution rules[M]//Daelemans W, Morik K. Machine learning and knowledge discovery in databases. Berlin: Springer, 2009: 115-130.

[40] Crémilleux B, Soulet A, Kléma J, et al. Discovering knowledge from local patterns in sage data[EB/OL]. [2015-10-16]. https://cremilleux.users.greyc.fr/papers/igiDMMKM09SageFinal.pdf.

[41] Nowicki K, Snijders T A. Estimation and prediction for stochastic block structures[J]. Journal of the American

Statistical Association,2001,96(455):1077-1087.

[42] Wang X R,Mohanty N,McCallum A. Group and topic discovery from relations and text[C]//LinkKDD '05:Proceedings of the 3rd International Workshop on Link Discovery. New York:ACM,2005:28-35.

[43] McCallum A,Corrada-Emmanuel A,Wang X. The author-recipient-topic model for topic and role discovery in social networks, with application to Enron and academic email [EB/OL]. [2015-10-15]. https://people. cs. umass. edu/~mccallum/papers/art-siam05s. pdf.

[44] Sese J,Seki M,Fukuzaki M. Mining networks with shared items[C]//Proceedings of the 19th ACM International Conference on Information and Knowledge Management,CIKM 2010. Toronto,Ontario,Canada,2010:1681-1684.

[45] Cheng H,Zhou Y,Huang X,et al. Clustering large attributed information networks:an efficient incremental computing approach[J]. Data Mining and Knowledge Discovery,2012,25(3):450-477.

[46] Dorndorf U,Pesch E. Fast clustering algorithms[J]. ORSA Journal on Computing,2004,6(2):141-153.

[47] Brusco M J,Köhn H F. Clustering qualitative data based on binary equivalence relations:neighborhood search heuristics for the clique partitioning problem[J]. Psychometrika,2009,74(4):685-703.

[48] Satuluri V,Parthasarathy S. Scalable graph clustering using stochastic flows:applications to community discovery [C/OL]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris,France,2009:737-746[2015-10-15]. http://www1. se. cuhk. edu. hk/~seem5010/slides/kdd09-satuluri. pdf.

[49] Xu Z,Ke Y,Wang Y,et al. A model-based approach to attributed graph clustering [C]//Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. New York,NY,USA:ACM,2012:505-516.

[50] Achtert E,Goldhofer S,Kriegel H P,et al. Evaluation of clusterings-metrics and visual support[C]//IEEE 28th International Conference on Data Engineering (ICDE). Arlington,Virginia,USA,2012:1285-1288.

郭红梅 中国科学院文献情报中心馆员。北京 100190。

张智雄 中国科学院文献情报中心研究馆员,博士生导师。北京 100190。

(收稿日期:2015-07-31)