

基于多元关系融合的科技文本主题识别方法研究^{*}

许海云 武华维 罗 瑞 董 坤 李 婧

摘 要 当前文本主题获取方法大多依靠单一关联分析,不能全面分析可获取信息,难以准确获取科技发展主题。科技文献的主题词、作者和引文之间蕴含了以研究主题内容为纽带的语义关联关系,主题词共现关系、引文关系和合著关系分别从不同的角度展现了主题关联关系。因此,本文根据主题词之间语义关系距离的远近,将主题识别中主题词关联分为基础关系、强化关系和新增关系,在此基础上提出面向主题识别的多元关系抽取及关系融合方法;并以基因工程疫苗的研发与制备领域为例进行领域实证分析,利用 PathSelClus 算法实现基于多元关系融合的主题聚类,通过对比实验证明多元关系融合可以有效提高实证领域的文本主题聚类效果,而未来多关系融合主题识别则是需要重点关注的问题。图 4。表 6。参考文献 19。

关键词 文本主题识别 多元关系 数据融合 关系融合 主题聚类

分类号 G254

Topic Identification Based on Multi-Semantic Relation Fusion

XU Haiyun, WU Huawei, LUO Rui, DONG Kun & LI Jing

ABSTRACT

One of the typical characteristics of big data analysis is multivariate data relation processing. The multi-relationship analysis of topics refers to the analysis of the relationships established between topics and other measurable entities (MEs). There are many MEs in scientific or technological documents, and they relate directly or indirectly with knowledge units. However, the current topic acquisition methods for this document rely mostly on single association analysis, so it is difficult to obtain the topics of scientific or technological developments accurately. Therefore, finding the multi-relationships between the entities of a document is one of the key technologies for accurate topic identification in massive scientific or technological literatures.

This paper firstly reviewed the research status of multi-relations fusion in topic identification, summarized the various measurable relationships of topic terms in the scientific or technical literatures. The research

^{*} 本文系国家自然科学基金项目“基于科学—技术主题关联分析的创新演化路径识别方法研究(编号:71704170)”,中国科学院知识产权信息服务专项“面向干细胞领域知识发现的科研信息化应用”(编号:KFJ-EW-STS-032)研究成果之一。(This article is an outcome of the project “Study on the Recognition Method of Innovative Evolving Trajectory Based on Topic Correlation Analysis of Science and Technology” (No. 71704170) supported by National Natural Science Foundation of China and the project “E-Science Application for Knowledge Discovery in Stem Cells” (No. KFJ-EW-STS-032) supported by the Informationization Special Project of Chinese Academy of Sciences.)

通信作者:许海云,Email: xuhy@clas.ac.cn, ORCID:0000-0002-7453-3331 (Correspondence should be addressed to XU Haiyun, Email: xuhy@clas.ac.cn, ORCID: 0000-0002-7453-3331)

found that there are semantic relations between topic terms, authors and citations in the scientific or technical literatures based on the topic content, with their co-occurrence relations can reveal respectively the topic association from different perspectives. Based on the distance between the semantic distances of topic terms, we divided the topic terms associations in topic identification into basic relations, strengthened relations and additional relations. For the strengthened relations and additional relations, any type of MEs can be the intermediate node of the topic terms association. Choosing the appropriate intermediate MEs is especially important for fully establishing the semantic association between the topic terms. This paper chooses authors, references and citation literatures as the intermediate MEs of topic term strengthened relations and additional relations. Seven types of topic associations are formed by the topic terms and these MEs. The fusion relationship can make up for the lack of information of a single association relationship through obtaining more accurate topic association.

The acquisition of multiple topic associations is the basis of multi-relations fusion. Whether the multi-relations fusion algorithm can enhance the meaningful topic semantic association and weaken the noise correlation is also an important step to achieve multi-relationship topic clustering. This study gives a calculation method of both direct and indirect association weights of MEs with reference to Morris's definition of association weights of multi-relational MEs. Finally, a multi-relationship extraction and relationship fusion method for topic identification is proposed. Finally, this paper took genetic engineering vaccine as experimental field, through relational matrix acquisition algorithm proposed by self-programming, seven types of topic correlation matrices were extracted and the correlation association matrices were realized by PathSelClus algorithm. With a comparative analysis, it proves that multi-relations fusion can effectively improve the effect of topic clustering.

The PathSelClus relationship fusion used in this paper is merely one of various existing multi-relations fusion methods, and it is highly dependent on expert knowledge. The quality of the annotation results directly affects the clustering results, and there is no effective way to determine the number of clusters. We think that much work needs to be done to further the study in the future, such as, how is the effect of other fusion methods working in topic identification from text? What will the performance comparison be? At the same time, according the research objective, we will explore more fusion methods and integrate them to obtain fusion results with more information. 4 figs. 6 tabs. 19 refs.

KEY WORDS

Topic recognition based on text. Multiple relations. Data fusion. Relational fusion. Topic clustering.

0 引言

文本主题自动识别是多种情报分析,如文献分类、检索以及领域前沿、热点识别的基础。当前科技文献的数量呈超指数增长,并伴随文献类型的日趋多样化,因此,科技文本主题识别逐步发展为大数据环境下的文本主题识别。大

数据情报分析的典型特征之一是多元数据关系处理,一篇科技文献存在多个计量实体,不同实体或知识单元之间存在多种直接或间接关联。多元数据融合是指通过数据融合算法实现多种关联数据的整合,从而通过更丰富的信息抽取与甄别,获得更准确的实体或知识单元之间的语义关联关系。由此减弱单一关联关系因对实体关联特征表达的不充分而带来的分析偏差,

以保证后续主题识别得到更客观、更具真实意义的主题^[1-3]。然而,当前大多数文本主题自动抽取方法主要依靠一种关联关系,难以保证实体或知识单元关系的充分分析,造成很多自动主题抽取方法的实际应用效果不佳。因此,深入分析主题关联的实体关系,在此基础上实现多元关系融合,是准确识别海量科技文献的文本主题识别的关键技术^[4]。目前更多的研究都是只关注前两种方法的使用,并且多是基于单一关系展开分析,少数文献计量的研究人员通过多种文献计量关系研究文本主题,这些研究发现通过多元数据融合可以显著提高文本主题自动识别的效果^[5-7]。除了对基于论文分析的多元关系融合探索,也有研究人员尝试对专利技术主题的识别做多元关系融合分析。例如,Zhang^[8]将基于IPC的类别相似分析与基于文本内容分析的语义分析进行专利组合的混合相似度计算。

基于目前主题识别的多元关系融合研究现状,本研究系统归纳面向主题识别的多元关系抽取及融合的方法及应用,在此基础上,尝试构建科学且可操作的多元关系抽取及关系融合的主题识别方法。之后,以基因工程疫苗的研发与制备领域为例,利用PathSelClus算法实现本研究基于多元关系融合的主题聚类,验证所提出的多元关系抽取及关系融合的主题识别方法的优势。最后,总结不足,同时提出未来需要研究的主题。

1 主题识别的多元关系抽取及关系融合

1.1 科技文献中主题相关的关系类型

本研究中主题识别的多元关系多是指主题词(术语)基于其他计量实体建立起的各种关联关系。科技文本是由具有语义信息的主题词按照一定的逻辑结构构成的有序实体,主题词之间存在多种不同的物理位置关系,这些不同的位置关系隐含了不同的语义关联。因此,通过挖掘术语之间不同的关联关系,理论上可以发

现多种不同的语义关联的表征规律,进而通过融合多种语义关系提高文本主题识别的准确度。尤其是,科技文献的主题词、作者和引文之间蕴含了以研究主题内容为纽带的语义关联关系^[2]。

根据不同的计量分析目的,多元关系的融合方式多种多样。依据融合关系是过滤性还是兼容性,多元关系融合可分为串行融合与并行融合。其中,串行融合是指在一种关系约束下再进一步考虑其他关系,实质上通过不同关系的依次过滤处理,实现多种限制关系的叠加,并行融合是指同时考虑多种关系,最后获得融合的综合关系^[1]。串行融合可以通过多种信息的融合实现多源信息的交叉校正,并行融合可通过信息的相互补充增强实体或知识单元的关联关系,其中,串行融合是目前文本主题分析中常用的多关系融合。在已有的计量实践中,研究人员多是尝试将参考文献和主题词结合起来识别领域主题。主题词与参考文献的结合分析有两种不同的方式,一种是利用参考文献作为词间关系的限定,即串行融合方式^[9-12];另一种是通过引用构建参考文献—词之间的关系,即并行融合方式^[3,13-15]。本文尝试通过多元关系的并行融合实现文本主题识别,由此提高科技文本主题识别的可解理性。

我们认为主题词的共现关系、引文关系与合著关系分别从不同的角度展现了主题词之间的语义关联。因此,本文拟通过对主题词、作者和引文之间关系的并行融合方式,弥补单一主题关联关系的信息缺失或歪曲,获取更准确的主题关联关系。要实现这一过程,首先要获取多元关系类型,之后通过融合算法实现多元关系的融合。本研究中,我们首先通过融合共现、引用和合著关系形成多元关系融合网络(见图1)。

1.2 主题识别的多元关系类型

主题词的共现关系、引文关系与合著关系对于主题间语义关联的表征能力不同,本研究根据多元关系中主题词之间的语义远近不同,将

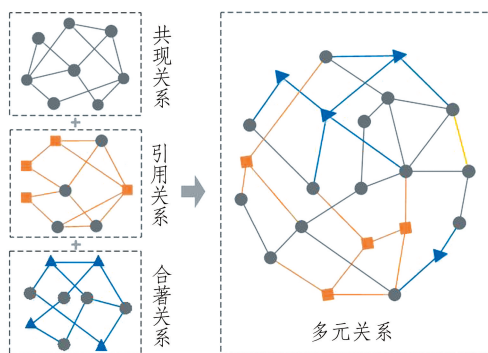


图1 文献实体关系融合示意图

面向主题识别的三种关系分为基础关系、强化关系和新增关系,具体含义如下:

(1)基础关系。我们认为,出现在同一篇文献中的主题词语义关系最为紧密,并且关联程度与共现次数成正比。主题词节点基于同一文献的直接共现关系是主题词关联分析的基础,由此反映的主题词语义距离最为相近,因此我们将其定义为基础关系。

(2)强化关系。主题词之间并非只有在同一篇文献中共现才发生关联,通常主题词并不出现在同一篇文献中,但仍可以通过其他文本实体建立关联。例如,出现在不同文献中但却拥有同一作者的主题词,因研究人员(作者)知识的累积性和延续性而存在语义关联;同理,出现在不同文献中但却拥有同一参考文献/同一引证文献的主题词,因文献引用的借鉴与继承而存在语义关联。这种关联弱于基础关系,但也反映了主题词之间的语义相关性,同时强化了主题词原有的关联,因此研究将其定义为强化关系。

(3)新增关系。除了上述两种关联关系,主题词之间还存在间接关联关系,即主题词之间不能通过某一计量实体直接产生关联,而需要以特定计量实体为中间桥接点建立间接关联。例如,科研合作是科学研究的主要形式,作者合作过程中伴随着研究人员知识的积累与扩散,尤其是隐性知识的传播多通过科研合作的形式发生。因此,出现在不同文献中且作者不同的

主题词,如果各自的作者曾有过合著活动,那么,可以认为这些主题词之间存在一定的语义关联。同理,文献引用体现科学知识纵向传播的连续性和继承性,同时也包含知识的横向交叉与渗透,是知识积淀与延伸的重要载体。因此,如果出现在不同文献中且拥有不同参考文献/引证文献的主题词,但各自的参考文献/引证文献曾有引文耦合/共被引关系,那么,可以认为这些主题词之间存在语义关联。这种关联关系弱于基础关系和强化关系,因此在一般的主题关联分析中,往往不会考虑这种间接关联,但是它可以作为一种新增的关系进一步增强主题词之间的语义关联,为直接关联关系提供补充,因此,研究将其定义为新增关系。

对于强化关系与新增关系而言,任意类型的计量实体都可以成为主题词关联的中间节点,但是不同计量实体对主题词的语义连接力不同,因此选择恰当的中间计量实体对于充分建立主题词之间的语义关联显得尤为重要。因此本文选择以作者、参考文献和引证文献作为主题词强化关系与新增关系的中间计量实体,由主题词和这些计量实体形成的多元关系的具体类型有7种,详见表1。

1.3 多元关系融合

1.3.1 已有关系权重计算方法

基础关系、强化关系和新增关系的中间关联实体不同,如何基于其他计量实体获取多元关系的权重计量方式是实现多元关系融合的主题识别中的关键步骤。

Morris 通过 $f_2 \cdot f_1$ 复合函数的形式计量不同关联关系的权重,其中, f_1 函数是获得两类计量实体 i 和 j 基于任意个体计量实体 k 的关联关系, f_2 函数的功能是获得 i 和 j 基于计量实体集中所有个体的关联关系^[16-17]。

对于 f_1 函数, Morris 给出了三种函数,包括矩阵相乘函数、功能叠加函数和反向闵可夫斯基函数。其中,矩阵相乘函数简单便捷,却在很大程度上放大了关系权值,对非二值矩阵不具

表1 面向主题识别的多元关系类型

多元关系类型		关系示例	计量实体类型
基础关系	主题词 共现关系 (TPT)	主题词 T_i 与主题词 T_j 在同一篇文献 P_m 中的直接共现关系: $T_i-P_m-T_j$	主题词
强化关系	基于作者的 强化关系 (TAT)	主题词 T_i 与主题词 T_j 出现于不同文献中,但拥有共同作者 A_i , 则主题词 T_i 与主题词 T_j 的关系通过作者节点 A_i 得到了强化: $T_i-A_i-T_j$	主题词 作者
	基于参考文献的 强化关系 (TRT)	主题词 T_i 与主题词 T_j 出现在不同文献中,但引用了相同文献 R_i ,则主题词 T_i 与主题词 T_j 的关系通过参考文献节点 R_i 得到了 强化: $T_i-R_i-T_j$	主题词 参考文献
	基于引证文献的 强化关系 (TCT)	主题词 T_i 与主题词 T_j 出现在不同文献中,但同时被文献 C_i 引 用,则主题词 T_i 与主题词 T_j 的关系通过引证文献节点 C_i 得到了 强化: $T_i-C_i-T_j$	主题词 引证文献
新增关系	基于作者的 新增关系 (TAPAT)	主题词 T_i 与主题词 T_j 出现于不同文献中,且没有共同作者,但 主题词 T_i 对应的作者 A_i 与主题词 T_j 对应的作者 A_j 拥有合著 文献 P_m ,则主题词 T_i 与主题词 T_j 通过作者 A_i 、文献 P_m 、作者 A_j 形成了新的关联关系: $T_i-A_i-P_m-A_j-T_j$	主题词 作者
	基于参考文献的 新增关系 (TRPRT)	主题词 T_i 与主题词 T_j 出现于不同文献中,且没有共同参考文 献,但主题词 T_i 对应的参考文献 R_i 与主题词 T_j 对应的参考文 献 R_j 共同被文献 P_m 引用,则主题词 T_i 与主题词 T_j 通过参考 文献 R_i 、文献 P_m 、参考文献 R_j 形成了新的关联关系: $T_i-R_i-P_m-R_j-T_j$	主题词 参考文献
	基于引证文献的 新增关系 (TCPCT)	主题词 T_i 与主题词 T_j 出现于不同文献中,且没有共同引证文 献,但主题词 T_i 对应的引证文献 C_i 与主题词 T_j 对应的引证文 献 C_j 共同引用了文献 P_m ,则主题词 T_i 与主题词 T_j 通过引证 文献 C_i 、文献 P_m 、引证文献 C_j 形成了新的关联关系: $T_i-C_i-P_m-C_j-T_j$	主题词 引证文献

有适用性;功能叠加函数可有效降低关系权值,但弱化了某些分量对结果的积极影响,不利于从整体上评判计量实体的关联强度;反向闵可夫斯基函数借鉴了解决电流、电导问题的测度方法,与科技文献中知识通过多元关系发生连通和流动的机理具有一致性,因此在原理上最为接近,但是该方法对各个分量作等同处理,没有消除不同分量量纲的影响。

对于 f_2 函数,当 i 和 j 通过多条路径产生关联时,每一条路径都增强了两者之间的关系,取最大值或最小值的方法都可能遗漏重要信息,难以从整体上体现实体 i 和实体 j 的关联,因此,求和函数作为 f_2 在计量实体关系分析应用场景中更具合理性。

1.3.2 关系矩阵获取与关系权重赋值

本文在参照 Morris 对多元关系计量实体的关联权重定义方式上,给出计量实体直接关联与间接关联权重的计算方式,并借助数学公式表示这些关联关系。根据文献及其主题词、作者、引文 3 种节点类型之间的关系,构建表 1 中的 TPT、TAT、TRT、TCT、TAPAT、TRPRT、TCPCT 七种关联矩阵,表征主题词之间的基础关系、强化关系和新增关系。

(1) 基础关系矩阵(共词矩阵)

假设分析文献集合为 Q ,主题词 T_i 与主题词 T_j 出现在同一篇文献 P 中,满足此条件的 T_i 与 T_j 所在的文献集合为 Q_0 , $Q_0 \subseteq Q$, T_i 在文献 P 中的频次为 n_{pi} , T_j 在文献 P 中的频次为 n_{pj} ,文献 P 的所有主题词频次为 N_p (见图 2)。假设在

集合 Q_0 中有 k 篇文献,即 T_i 与 T_j 在 k 篇文献中共现,在集合 Q_0 中, T_i 与 T_j 的关系权重为:

$$d_{ij} = \sum_{p=1}^{p=k} \left(\frac{n_{pi}}{N_p} + \frac{n_{pj}}{N_p} \right), \text{其中, } \frac{n_{pi}}{N_p} \text{ 表示 } T_i \text{ 与文献 } P \text{ 的关联度,数值越高,说明 } T_i \text{ 与文献 } P \text{ 的关联度越强;同理, } \frac{n_{pj}}{N_p} \text{ 表示 } T_j \text{ 与文献 } P \text{ 的关联度,数值越高,说明 } T_j \text{ 与文献 } P \text{ 的关联度越强。} d_{ij} \text{ 数值越高,说明 } T_i \text{ 与 } T_j \text{ 在集合 } Q_0 \text{ 中的总体关联强度越大。根据此公式,获得集合 } Q \text{ 中任意两个主题词之间的关系权重,将关系权重作为主题词相关系数,可获得主题词关联矩阵,此矩阵即为基础关系矩阵。}$$

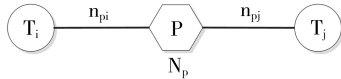


图2 基础关系权重示意图

(2) 强化关系矩阵(以基于作者的强化关系矩阵为例)

假设分析文献集合为 Q , 主题词 T_i 与主题词 T_j 未出现在同一篇文献中,但两者拥有共同作者 A , 满足以上条件的 T_i 与 T_j 所在的文献集合为 $Q_1, Q_1 \subseteq Q$, 在 Q_1 中有 u 个这样的作者 A 。 T_i 与作者 A 的关联频次(Q_1 中,作者 A 撰写的所有 T_i 的频次)为 n_{ai} , T_j 与作者 A 的关联频次(Q_1 中,作者 A 撰写的所有 T_j 的频次)为 n_{aj} , 在集合 Q_1 中,作者 A 撰写的所有主题词的频次之和为 N_a (见图3), T_i 与 T_j 基于作者 A 的强化关系权重为: $d_{ij} = \sum_{a=1}^{a=u} \left(\frac{n_{ai}}{N_a} + \frac{n_{aj}}{N_a} \right)$ 。其中, $\frac{n_{ai}}{N_a}$ 表示 T_i 与作者 A 的关联度,数值越高,说明 T_i 与作者 A 的关联强度越大;同理, $\frac{n_{aj}}{N_a}$ 表示 T_j 与作者 A 的关联度,数值越高,说明 T_j 与作者 A 的连接力越强。 d_{ij} 反映了 T_i 与 T_j 通过作者建立的关联强度,数值越高,说明 T_i 与 T_j 在文献集 Q_1 中基于作者的关联强度越大。根据此公式,可获得总文献集 Q 中任意两个主题词之间基于作者的强化关系权重,将此关系权重作为主题词相关系数,可获得主题词关联矩阵,即为基于

作者的强化关系矩阵。



图3 强化关系权重示意图

(3) 新增关系矩阵(以基于作者的新增关系矩阵为例)

假设分析文献集合为 Q , 主题词 T_i 与主题词 T_j 未出现在同一篇文献中,且不拥有共同作者,但 T_i 的作者 A_k 与 T_j 的作者 A_r 有合著关系,满足以上条件的 T_i 与 T_j 所在的文献集合为 $Q_2, Q_2 \subseteq Q$ 。 T_i 与作者 A_k 的关联频次(Q_2 中,作者 A_k 撰写的所有 T_i 的频次)为 $n_{a_k i}$, T_j 与作者 A_r 的关联频次(Q_2 中,作者 A_r 撰写的所有 T_j 的频次)为 $n_{a_r j}$, 作者 A_k 与 A_r 的所有合著主题词频次为 n , 在集合 Q_2 中,作者 A_k 撰写的所有主题词的频次之和为 N_{a_k} , 作者 A_r 撰写的所有主题词的频次之和为 N_{a_r} (见图4)。假如在文献集 Q_2 中,满足 A_k 与 A_r 条件的组合 c 有 t 种, $C(A_k, A_r) \in [1, t]$, 即 T_i 与 T_j 可以通过 t 种 A_k 与 A_r 组合中的任意一种或多种产生关联,在文献集 Q_2 中, T_i 与 T_j 基于作者 A 的新增关系权重为: $d_{ij} = \sum_{c=1}^{c=t} \left(\frac{n_{a_k i}}{N_{a_k}} \times \frac{n}{N_{a_k}} + \frac{n_{a_r j}}{N_{a_r}} \times \frac{n}{N_{a_r}} \right)$ 。

其中, $\frac{n_{a_k i}}{N_{a_k}}$ 表示 T_i 与作者 A_k 的关联度,数值越高,说明 T_i 与作者 A_k 的关联越强;同理, $\frac{n_{a_r j}}{N_{a_r}}$ 表示 T_j 与作者 A_r 的关联度,数值越高,说明 T_j 与作者 A_r 的关联越强。 d_{ij} 为 T_i 与 T_j 在文献集 Q_2 中基于作者的关联强度,数值越高,关联强度越大。根据此公式,可获得总体文献集 Q 中任意两个主题词之间基于作者的新增关系权重,将此关系权重作为主题词相关系数,可获得主题词关联矩阵,此矩阵即为基于作者的新增关系矩阵。

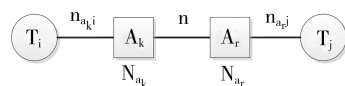


图4 新增关系权重示意图

1.3.3 关系融合

多种主题关联关系的获取是多关系融合的基础,而多关系融合算法能否增强有意义的主题语义关联而减弱噪音关联,也是实现多元关系主题聚类的重要问题。

目前,基于元路径提取进行异构信息网络融合的研究不断涌现。本文使用 PathSelClus 算法实现 7 种关联矩阵的融合,计算综合主题词的综合相似度。PathSelClus 算法使用概率图模型的方法来定义问题,主要包括三个子模型:元路径生成模型;用户标注的种子对象模型;元路径的权重模型^[18]。

该模型在多元关系(元路径)的层次上对异构信息网络中某一类目标节点进行聚类。PathSelClus 算法将用户的意图通过标引种子对象的方式引入到模型,因此,聚类是在用户标注少量数据结果的引导下进行的,更能够符合用户的意图。同时,不同的关联矩阵结合其权重,可以更好地分析聚类结果的语义关系来源,因此可解释性更强。

2 实证分析

2.1 数据获取与分析

本文选择基因工程疫苗(Gene Engineered Vaccine)的研发与制备领域作为实证分析对象。基因工程疫苗是新型疫苗类型的主要组成部分,也是生物医药领域的重点发展分支之一。本文实证数据来自中国知网论文数据库,分别通过基因工程疫苗研发与制备的相关概念和组成要素的组合,制定四组检索式,汇总检索结果并经人工排杂后共获取 4 315 篇研究论文,检索日期截至 2017 年 6 月。

第一步:数据预处理。数据预处理阶段的主要目标是从文献数据集中获取主题词、作者、参考文献、引证文献等字段作为分析对象。其中,作者、参考文献、引证文献等字段直接从原始数据中抽取,作者出现频次不少于 3 次、引证文献出现次数不少于 2 次、参考文献出现次数

不少于 2 次的记录。主题词字段需要通过文本挖掘方法从标题与摘要中抽取,本文借助 Derwent Data Analyzer(DDA)进行主题词预处理的基本流程^[19],DDA 在传统主题词清洗方法(如停词表、模糊语义处理等)的基础上引入当前文献计量学领域较为成熟的文本分析方法进行主题词清洗。为保证数据集的全面性,选择覆盖率高的分析字段:combined keywords + phrase,之后利用词表清洗工具和基于词形的相似度合并命令进行初步的合并,通过该步骤可以除去数字、停用词、大量跟分析主题无关的词,最终选择词频不小于 3 的主题词作为待分析主题词集。

第二步,抽取关联矩阵。通过自编程序按照 1.2 关系矩阵获取算法,抽取 TPT、TAT、TRT、TCT、TAPAT、TRPRT、TCPCT 共 7 种主题关联矩阵。

第三步,多元关系融合。使用 Matlab 实现 7 种矩阵的融合。

(1) 输入数据的准备

Scell:该变量存储了 M 条元路径的集合 P₁, P₂, ..., P_M 的邻接矩阵,Scell 的每一个元素为元路径关系矩阵 W_m。即 Scell = [W₁, ..., W_m],为 1×m 的 cell 数据类型。

SeedsMat:种子对象矩阵,矩阵的大小为 n×k, n 表示属于目标类型 T 的聚类对象的个数, k 为聚类的个数。

(2) 算法的输入

①异构信息网络 G。②聚类的目标类型 T。③聚类的个数 K。④每一个聚类的种子对象 L₁, ..., L_k,其中种子对象 L_k可以为空,即不需要对每一个聚类都指定样例。⑤以目标类型 T 为起点的 M 条元路径的集合 P₁, P₂, ..., P_M。

对每一个元路径,计算其邻接矩阵(adjacency matrix W_m),也叫关系矩阵(relation matrix)。计算方法是将原路径中每一对顶点表示的邻接矩阵依次相乘。以引文网络为例:元路径 APV 的关系矩阵 W = W_{AP} × W_{PV}。

(3) 算法的输出

①每一条元路径 P_m 的权重为 α_m (>= 0)。

②属于目标类型的对象的聚类结果,对任一 t_i , 当 $t_i \in T$, 有 k 维向量 $\theta_i = (\theta_{i1}, \dots, \theta_{ik})$, 其中 $\sum \theta_i = 1$ 。Matlab 代码实现的使用: function [thetaMat, betaCell, piVec, accuracy, nmi_value] = PathSelect_v3(Scell, SeedsMat, lambda, ground-

Truth)。

第四步,关系聚类与主题命名。对融合矩阵进行关系聚类,在领域专家参与下给予主题命名。表2为关系融合后的主题词关系矩阵,表3为专家参与下的聚类主题。

表2 主题词关系矩阵(部分)

	25aa- ELISA	2 型猪 链球菌	Ag85A	Ag85B	BALB/c 小鼠恒 定链	CCL17	CCL5	CCR4	CCR5	CD4+ CD25+ 调节性 T 细胞	CD80
23kD 膜蛋白	0	0	0	0	0	0	0	0	0	0	0
25aa- ELISA	0	0	0	0	0	0	0	0	0	0	0
2 型猪 链球菌	0	0	0	0	0	0	0	0	0	0	0
AMA1 基因	0	0	0	0	0	0	0	0	0	0	0
Ag85A	0	0	0	0.031 4	0	0	0	0	0	0	0
Ag85B	0	0	0.031 4	0	0	0	0	0	0	0	0
B2L 基因	0	0	0	0	0	0	0	0	0	0	0
BAC	0	0	0	0	0	0	0	0	0	0	0
BALB/c 小鼠恒 定链	0	0	0	0	0.104 0	0	0	0	0	0	0
B 细胞 表位	0	0	0	0	0	0	0	0	0	0	0
CCL17	0	0	0	0	0	0	0	0	0	0.031 1	0.049 2
CCL5	0	0	0	0	0	0	0	0	0	0.031 1	0.049 2
CCR4	0	0	0	0	0	0	0	0	0	0.031 1	0.049 2
CCR5	0	0	0	0	0	0	0	0	0	0.031 1	0.049 2
CD4+ CD25+ 调节性 T 细胞	0	0	0	0	0	0.031 1	0.031 1	0.031 1	0.031 1	0	0.027 6
CD80	0	0	0	0	0	0.049 2	0.049 2	0.049 2	0.049 22	0.027 6	0

表3 基因工程疫苗聚类主题(关系融合)

1979—2000 年	2001—2005 年	2006—2010 年	2011—2015 年
#0 Th1/Th2 免疫应答 #1 猪感染相关病毒的免疫机制及基因工程疫苗 #2 病毒基因疫苗的研发及免疫学 #3 乙肝及肿瘤相关基因疫苗的构建 #4 人乳头瘤病毒基因工程疫苗 #5 禽畜基因工程疫苗的研究及安全性评价 #6 疫苗候选抗原的筛选 #7 病毒及其疫苗的生产制备技术 #8 防龋疫苗的研究 #9 DNA 疫苗免疫动物的实验 #10 合成肽疫苗 #11 禽畜类基因重组疫苗	#0 抗肿瘤核酸疫苗的构建及其表达 #1 自杀性 DNA 疫苗及免疫效应 #2 双启动子 DNA 疫苗载体的构建及免疫应答 #3 禽感染相关核酸疫苗(针对弓形虫、新城疫、口蹄疫、鸡艾美耳球虫) #4 抗独特型单克隆抗体疫苗的制备 #5 肿瘤基因疫苗 #6 结核分枝杆菌 DNA 疫苗的构建及免疫效果 #7 DNA 疫苗的转染载体及免疫途径 #8 DNA 疫苗免疫原性研究及候选抗原基因 #9 禽流感 DNA 疫苗的构建及免疫保护效果 #10 DNA 疫苗的佐剂选择(协同信号分子、黏附因子等) #11 多表位疫苗的制备 #12 合成肽疫苗 #13 疫苗共表达载体的构建及其诱导的免疫应答 #14 新孢子虫病和鸡球虫病相关基因疫苗 #15 糖尿病和流感病毒相关 DNA 疫苗 #16 亚单位疫苗和 DNA 疫苗 #17 噬菌体展示颗粒疫苗及转基因植物疫苗 #18 免疫避孕疫苗及基因重组疫苗	#0 肽疫苗和活载体疫苗的设计及免疫学 #1 新型疫苗的研究(免疫调节型 DNA 疫苗、标记疫苗、重组弱毒疫苗) #2 癌症相关疫苗及其佐剂 #3 疫病相关基因工程疫苗 #4 感染性疾病疫苗及佐剂 #5 多表位融合蛋白疫苗及免疫应答 #6 免疫避孕疫苗、肺炎相关蛋白疫苗 #7 活载体疫苗和表位 DNA 疫苗 #8 禽类基因疫苗 #9 肿瘤 DNA 疫苗及免疫治疗 #10 鸡新城疫—传染性支气管炎—减蛋综合征基因工程疫苗和抗龋疫苗 #11 多表位重组疫苗 #12 鸡球虫病核酸疫苗的构建及其免疫保护效果 #13 血吸虫 DNA 鸡尾酒疫苗、水产相关疫苗 #14 HPV 和帕金森基因工程疫苗、猪瘟兔化弱毒疫苗	#0 新型猪瘟疫苗(标记疫苗、活载体疫苗、核酸疫苗) #1 动物 DNA 疫苗及抗独特型抗体疫苗 #2 DNA 疫苗佐剂研究(纳米粒制备、刺激因子、C3d) #3 融合基因疫苗 #4 蛋白疫苗、甲病毒复制子载体疫苗及分子佐剂研究 #5 禽类相关基因工程疫苗 #6 重组活载体疫苗 #7 基因工程亚单位疫苗和合成肽疫苗 #8 免疫策略及免疫效果 #9 变态原疫苗及多表位疫苗 #10 新型疫苗载体

2.2 对比分析

2.2.1 对比试验结果

为验证多元关系抽取及关系融合方法在主题识别中的优势,本文运用单一共词分析形成的聚类主题作为对比分析。表4为专家参与下得到的与关系融合方法相同时间分段的聚类主题。

2.2.2 对比分析

总体上,对比试验中,单一共词聚类的抽取词过多,每个主题包含的抽取词相比关系融合方法多出至少一倍,主题意义泛化,导致多个主题之间缺乏差异度,主题命名难度增大,有效聚类不足。相比之下,关系融合方法效果更好,主题之间的差异度较明显,得到了更多有效聚类。

表 4 基因工程疫苗聚类主题(共词分析)

1979—2000 年	2001—2005 年	2006—2010 年	2011—2015 年
#0 基因工程疫苗 (亚单位疫苗和活载体疫苗) #1 病毒及肿瘤相关 基因疫苗及免疫学 研究 #2 DNA 疫苗的制备 #3 基因疫苗载体 构建 #4 合成肽疫苗、基 因疫苗(乙肝、血吸 虫、肿瘤相关) #5 基因工程疫苗的 设计及免疫效果 评价 #6 基因工程疫苗的 研究(细菌、病毒) 及抗原的筛选 #7 禽畜类基因重组 疫苗的研究 #8 肽疫苗及肿瘤基 因工程疫苗(DNA 疫苗、细胞疫苗) #9 疫苗设计与免疫 应答	#0 基因重组疫苗和 DNA 疫苗 的构建及其免疫应答机制和 效果 #1 疫苗候选分子、基因重组 疫苗载体设计(真核表达载 体、叶绿体载体等) #2 病毒感染相关基因疫苗的 研究 #3 基因工程活载体疫苗、亚 单位疫苗及免疫学研究 #4 禽类相关基因疫苗的研究 (DNA 疫苗、基因重组疫苗、 亚单位疫苗等)及免疫评价 #5 基因工程疫苗的制备方法、 佐剂的选择及免疫学研究 #6 DNA 疫苗融合蛋白疫苗的 制备及免疫学研究 #7 疫苗候选抗原、表位融合 蛋白疫苗、免疫机制及免疫效 果评价 #8 主题词范围过多,难以 命名 #9 基因工程疫苗和核酸疫苗 的制备 #10 主题词范围过多,难以 命名 #11 主题词范围过多,难以 命名 #12 表位疫苗及新型疫苗载 体的设计 #13 基因工程疫苗,抗独特型 单克隆抗体疫苗 #14 核酸疫苗和基因工程疫 苗的设计及免疫策略选择	#0 禽类相关基因疫苗的设计 (基因重组疫苗、表位疫 苗、DNA 疫苗、标记疫苗、 肽疫苗等) #1 病毒及肿瘤相关基因 疫苗 #2 基因工程疫苗的设计 (载体的构建、共表达、标记 等) #3 疫苗的抗原筛选、佐剂 选择及基因工程疫苗的 构建 #4 疫苗佐剂研究、基因疫 苗、DNA 疫苗 #5 禽类和肿瘤相关疫苗的 研究(基因工程疫苗、DNA 疫苗、蛋白疫苗等) #6 主题词范围过多,难以 命名 #7 主题词范围过多,难以 命名 #8 主题词范围过多,难以 命名 #9 主题词范围过多,难以 命名 #10 主题词范围过多,难以 命名 #11 主题词范围过多,难以 命名 #12 主题词范围过多,难以 命名	#0 病毒相关疫苗(基 因工程活载体疫苗、表 位疫苗、DNA 疫苗) #1 基因工程疫苗载体 构建(活载体、甲病毒 复制子等)及表达优化 #2 疫苗的制备方法 研究 #3 表达载体的构建 (反向遗传操作、共表 达等)、基因工程疫苗 的构建及免疫效力 检测 #4 肿瘤相关基因疫苗、 疫苗载体及表达系统 的研究、疫苗分子佐剂 的研究 #5 疫苗的构建及优化、 通用疫苗、疫苗生物信 息学分析 #6 新型疫苗载体(不 同启动子、活载体、甲 病毒复制子、DNA、细 菌等)、自杀性 DNA 疫 苗、疫苗免疫途径及免 疫效果评价

在不同时间窗,主题聚类存在以下具体差异。

(1) 1979—2000 年

相同点:关系融合方法与单一的共词聚类发现的研究热点大体一致,从疾病类型来看,包括了肝炎、肿瘤、禽畜(猪、鸡等)、其他病毒类

(HPV、轮状、流感等),从疫苗种类来看,包括了亚单位疫苗、减毒活疫苗、基因重组疫苗、DNA 疫苗、肽疫苗等,从制备技术上对疫苗的载体构建、表达、免疫动物实验、效果评价等都进行了相关研究。

区别;关系融合方法识别了新的技术主题,如:防龋齿疫苗的研究,变形链球菌是龋病的主要致病菌,该阶段主要利用基因重组技术、植物表达载体等开展可食用防龋疫苗研究。

(2) 2001—2005 年

关系融合方法与单一的共同聚类发现的研究热点大体一致,但关系融合方法识别出新技术主题免疫避孕疫苗的研究。单一共同分析方法有 3 个主题由于主题词范围过多,难以实现主题命名(主题 8、主题 9 和主题 11)。

(3) 2006—2010 年和 2011—2015 年

关系融合方法与单一的共同聚类发现的研究热点大体一致。关系融合方法未发现新的主题,这也可能与基因工程疫苗领域在该阶段的快速发展、主题分化有关。这一阶段的关系融合方法依旧优于单一的共同聚类方法,后者由于抽取得到的主题词数量过多,导致主题泛化,

不易进行主题识别的命名。2006—2010 年,单一共同分析方法有 7 个主题,由于主题词范围过多,难以实现主题命名(主题 8—12)。2011—2015 年的单一共同分析方法仅得到 6 个聚类主题,相对于关系融合方法的聚类主题粒度较粗。

2.2.3 原因分析

首先,单一共同分析的每一个主题中为什么出现那么多的关键词?单一共同分析的对比实验中很多关键词在每一个主题上的分布概率相同,即该算法根据所获的信息无法判断该关键词属于哪个主题。如表 5 中,term1 为用户标注的种子词,属于主题 0。而 term2 和 term3 在主题上的概率相同,同为“最大值”,则被“误认”隶属于所有的 8 个主题。该情况在单一共同分析中非常普遍,因此出现主题中的关键词非常多且无法进行专家解读的情况。

表 5 单一共同分析隶属聚类主题示例

	主题 0	主题 1	主题 2	主题 3	主题 4	主题 5	主题 6	主题 7
Term1	1	0	0	0	0	0	0	0
Term2	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
Term3	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125

多元关系融合算法输出的主题结果为每一个关键词隶属的概率最大的主题,如表 6 为主题词 term 1 在 8 个主题上的分布概率分布,term1 就会划分为概率值最大的主题 7。因此,

本研究通过 PathSelClus 算法实现基础关系、强化关系和新增关系的关系融合,有效增强了主题语义关联而减弱噪音关联,提高了聚类主题识别的效果。

表 6 关系融合隶属聚类主题示例

主题 0	主题 1	主题 2	主题 3	主题 4	主题 5	主题 6	主题 7
0.094 696	0.094 696	0.094 696	0.094 696	0.094 696	0.113 06	0.094 696	0.318 76

3 结语

当前文本主题获取方法大多依靠单一关联分析,不能全面分析可获取信息,难以准确获取科技发展主题。考虑到科技文献的主题词、作

者和引文之间蕴含了以研究主题内容为纽带的语义关联关系,且不同关联关系对于主题间语义关联的表征能力不同,本文将面向主题识别的共现关系、引文关系与合著关系分为基础关系、强化关系和新增关系,作为本文中多元关系融合的主题识别的关系类型。在此基础上,构

建了完整的基于多元关系抽取及多关系融合的主题识别的方法。并以基因工程疫苗的研发与制备领域为例,利用 PathSelClus 算法实现多元关系融合的主题识别,对比分析证明,多元关系融合可以有效提高主题聚类的效果。我们认为,未来基于多关系融合的主题识别需要重点关注以下两个问题。

首先,多元关系表征与获取是未来重点研究方向之一。任意两个计量实体间的关系都具有特定的语义信息,不同关系对研究目标的表征程度不尽相同,理清不同关系的信息表征特点,并筛选出具有融合价值的多元关系类型,是当前数据关系融合的难题。未来考虑全面衡量主题间的共现、语法和语义关系。同时,在关系复杂、关联实体较多的情况下,可利用超网络、元路径等表示多元关系,在间接关联关系获取时需建立完整的计算关联权重的方法体系,根

据路径长短分层次获取这些关系,提升各种关系对目标的表征精度。

其次,对其他领域数据关系融合方法在文本主题识别中的适用性进行验证。多关系融合算法能否增强有意义的主题语义关联而减弱噪音关联,是影响多元关系主题聚类的关键要素。本文中的 PathSelClus 关系融合仅是众多已有多元关系融合方法中的一种,PathSelClus 算法高度依赖专家的知识 and 判断,标注结果的好坏直接影响聚类的结果,并且缺少有效的方式确定聚类的个数 K 的值。

那么,已有的成熟的其他融合方法在文本主题识别中的效果如何,性能对比如何?同时,如何结合研究目标探索更多的数据关系融合方法,将多种融合方法综合集成,以获取蕴含更大信息量的融合结果。这些都是将来需要进一步研究的主要内容。

参考文献

- [1] 许海云,董坤,隗玲,等.科学计量中多源数据融合方法研究述评[J].情报学报,2018,37(3):318-328.
(Xu Haiyun, Dong Kun, Wei Ling, et al. Research on multi-source data fusion method in Scientometrics[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(3): 318-328.)
- [2] Xu H Y, Yue Z H, Wang C, et al. Multi-source data fusion study in scientometrics[J]. Scientometrics, 2017, 111(2), 773-792.
- [3] Janssens F, Zhang L, De Moor B. Hybrid clustering for validation and improvement of subject-classification schemes[J]. Information Processing & Management, 2009, 45(6): 683-702.
- [4] 许海云,董坤,刘春江,等.文本主题识别关键技术研究综述[J].情报科学,2017,35(1):153-160.(Xu Haiyun, Dong Kun, Liu Chunjiang, et al. A review on topic identification of scientific text files [J]. Information Science, 2017, 35(1): 153-160.)
- [5] Van Den Besselaar P, Heimeriks G. Mapping research topics using word-reference co-occurrences: a method and an exploratory case study[J]. Scientometrics, 2006, 68(3): 377-393.
- [6] Wen B, Horlings E, Mariëlle V D Z. Mapping science through bibliometric triangulation: an experimental approach applied to water research[J]. Journal of the Association for Information Science & Technology, 2016.
- [7] Dong K, Xu H, Luo R. An integrated method for interdisciplinary topic identification and prediction: a case study on information science and library science[J]. Scientometrics, 2018, 115(2): 849-868.
- [8] Zhang Y, Shang L, Huang L, et al. A hybrid similarity measure method for patent portfolio analysis[J]. Journal of Informetrics, 2016, 10(4): 1108-1130.
- [9] Calero-Medina C, Noyons E C. Combining mapping and citation network analysis for a better understanding of the

- scientific development: the case of the absorptive capacity field[J]. Journal of Informetrics, 2008, 2(4): 272-279.
- [10] He X, Ding C H, Zha H. Automatic topic identification using webpage clustering[C]//Data Mining, 2001 ICDM 2001, Proceedings IEEE International Conference on, 2001: 195-202.
- [11] He X, Zha H, Ding C H. Web document clustering using hyperlink structures[J]. Computational Statistics & Data Analysis, 2002, 41(1): 19-45.
- [12] Wang Y, Kitsuregawa M. Evaluating contents-link coupled web page clustering for web search results[C]//Proceedings of the eleventh international conference on Information and knowledge management, 2002: 499-506.
- [13] Janssens F, Glänzel W, De Moor B. Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis[C]// Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007: 360-369.
- [14] Janssens F A. Clustering of scientific fields by integrating text mining and Bibliometrics[D]. Belgium: Katholieke Universiteit Leuven, 2007.
- [15] 郭红梅, 孔贝贝, 张智雄. 基于多重文本关系图中 clique 子团聚类主题识别方法研究[J]. 情报学报, 2017, 36(5): 433-442. (Guo Hongmei, Kong Beibei, Zhang Zhixiong. Study on textual topic identification by clustering clique structure in multi-relationship text graph[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(5): 433-442.)
- [16] Morris S A, Yen G G. Construction of bipartite and unipartite weighted networks from collections of journal papers[J]. Physics, 2005.
- [17] Morris S A. Unified mathematical treatment of complex cascaded bipartite networks: the case of collections of journal papers[D]. Stillwater: Oklahoma State University, 2005.
- [18] Sun Y, Norick B, Han J, et al. PathSelClus: integrating meta-path selection with user-guided object clustering in heterogeneous information networks[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2013, 7(3): 11.
- [19] Derwent data analyzer[EB/OL]. [2018-10-07]. <https://www.thevantagepoint.com/tda-home.html>.

许海云 中国科学院成都文献情报中心, 中国科学技术信息研究所。四川 成都 610041。

武华维 中国科学院成都文献情报中心, 中国科学院大学。四川 成都 610041。

罗 瑞 中国科学院成都文献情报中心, 中国科学院大学。四川 成都 610041。

董 坤 山东理工大学科技信息研究所。山东 淄博 255200。

李 婧 中国科学院成都文献情报中心。四川 成都 610041。

(收稿日期: 2018-05-12; 修回日期: 2018-10-01)