基于 LDA 主题关联过滤的领域主题演化研究*

秦晓慧 [1,2]　乐小虬 [1]
[1](中国科学院文献情报中心　北京 100190)
[2](中国科学院大学　北京 100049)

摘要: 【**目的**】发现领域文献中主题的新生、消亡、继承、分裂和合并的演化轨迹。【**方法**】根据文献出版时间划分多个时间窗口, 通过 LDA 主题模型识别各个时间窗口中的主题; 利用主题关联(Topic Association)过滤规则确定相邻时间窗口主题间的演化关系; 形成连续时间段内主题新生、消亡、继承、分裂和合并的演化轨迹。【**结果**】在保证主题延续性的条件下, 更准确地识别主题的新生、消亡、继承、分裂和合并的演化类型。【**局限**】固定的时间窗口, 未考虑主题演化周期的多样性。【**结论**】该方法可以有效降低 LDA 主题模型中相似度较小主题的干扰, 提升主题演化关系识别的准确性。

关键词: 主题关联　主题演化　主题模型　LDA
分类号: TP393

1　引　言

领域主题演化指一个领域的主题内容与强度在研究过程中变化的现象[1], 能够帮助研究者深入了解主题产生、发展的过程。近年来, 有关领域主题演化的研究涌现出许多新思路和新方法, 其中改进LDA主题模型[2]是重要途径之一。常见的处理方式是利用 LDA 模型获取不同时间段出现的主题, 将相邻时间窗口的主题采用阈值法[3-7]或最大相似度法[8-9]等进行关联。这种方法虽然能描述主题内容随时间的演化过程, 但其准确性存在瑕疵, 经常会使无关主题引入到演化关系中。

为了解决上述问题, 本文提出通过制定主题关联过滤规则, 对相邻时间窗口间的主题进行关联分析, 以期减少非关联主题的干扰问题。将对处理流程、实现方法以及实验结果进行具体阐述。

2　相关研究

LDA 模型[2]是一个三层贝叶斯生成模型, 其基本思想为: 主题是一个在词表上的多项式分布, 而每篇文献对这些主题有一个特定的分布。由于它可以很好地模拟大规模语料的语义信息, 在主题演化领域有一定的优势, 学者们对其进行了一系列扩展工作, 如动态主题模型 DTM[10]、在线主题模型 OLDA[11]、连续时间模型 TOT[12]等; 其应用涉及电子邮件[2]、科研文献[13-14]、微博[15]、作者[16-17]等主题演化。

为全面分析主题内容的演化趋势, 常见的处理方式是根据文献的出版时间离散到相应的时间窗口[18], 利用LDA获取不同时间窗口出现的主题, 将相邻时间窗口间的主题关联, 进而获得主题的演化过程。其中相邻时间窗口的主题关联是主题演化分析的重要步骤, 主题能否关联决定着主题之间是否存在演化关系, 对

通讯作者: 秦晓慧, ORCID: 0000-0002-3084-2546, E-mail: qinxh@mail.las.ac.cn。
*本文系国家科技支撑计划子课题“基于文献知识网络的领域学术关系研究与示范”(项目编号: 2011BAH10B06-04)的研究成果之一。

主题演化结果有直接影响。本文归纳了4类常见的主　　题关联方法，如表 1 所示:

表 1　主题关联方法

类型	代表研究	说明	优点	缺点
直接关联法	DTM[10]、DMM[19]	每个时间窗口的主题数相同，主题一一对齐，认为不同时间窗口中序号相同的主题存在关联	主题关联可直接建立，无需进行更多计算	假设主题数目在整个时间段内不发生变化，是不合理的
相似度阈值法	楚克明等[3-4]	设定一个主题相似度阈值，用相似度衡量相邻时间窗口的主题相似度，当相邻时间窗口的主题相似度大于该阈值时，说明主题之间具有关联关系	可捕获主题的一对多、多对多演化现象	演化结果对阈值的大小比较敏感，阈值的确定需要较强的专业知识
距离阈值法	李保利等[5]、崔凯等[6]、胡吉明等[7]、唐晓波等[15]	设定一个主题距离阈值，用 K-L 距离衡量相邻时间窗口的主题差异性，当主题与时间相邻窗口的距离大于阈值时，认为主题发生改变，即主题不存在关联	引入参数较少，可探测到主题的新生、消亡现象	得到的主题演化都是一对一的，与科学研究中主题的融合、交叉、分裂等现象不完全吻合
最大相似度法	Lv 等 [8]、胡艳丽等[9]	对时间窗口 t 中的主题 T_i^t，找出时间窗口 t–1 和 t+1 中与 T_i^t 相似度最大的一个 T_r^{t-1}、T_k^{i+1}，分别定义为 T_i^t 的父主题与子主题。经向前与向后的推理，可获得多种演化现象	可捕获多种主题演化类型	可能在相似度小的主题间建立关联，而相似度更大的主题之间没有建立关联

对于相邻时间窗口间主题的关联，直接关联法无需计算主题相似度，但必须固定时间窗口的主题数，不能有效地反映新主题的产生、旧主题的消亡等现象；其他三种方法可根据经验值、困惑度[2]等将时间窗口中的主题数设为可变量，通过计算主题相似度或主题距离决定主题是否存在关联，相对来说应用更灵活。但也各有不足之处：在相似度阈值法中，主题演化关系的确定依赖于一个固定的阈值，且对阈值的大小比较敏感，例如，相似度阈值太大会导致相邻时间窗口间存在演化关系的主题过少，而相似度阈值法太小会使无关主题引入到主题演化中，因此阈值的设定需要较强的专业知识；最大相似度法可能会丢失相似度比较大的主题关联，而误将相似度小的主题关联在一起。基于以上原因，本文采用主题关联过滤方法判别主题间是否存在较强的关联，进而分析主题的演化。

3　基于 LDA 主题关联过滤的领域主题演化

3.1　基本思路

(1) 将时间序列划分为若干个长度固定的不重叠时间窗口，根据文献的出版时间把文献划入到相应的时间窗口，利用LDA 主题模型识别不同时间窗口出现的主题；

(2) 计算相邻时间窗口中主题的相似度，建立主题关联；

(3) 经主题关联过滤，筛选出有效的主题关联；

(4) 针对相关联的主题，经过沿时间序列向前或向后的推理，判别主题的演化关系类型。

3.2　处理流程

(1) 领域主题识别。针对不同时间窗口的文献集，经 LDA 模型得到该时间窗口中的主题，此处的主题表示为一个分布在一组主题词上向量；

(2) 主题关联建立。对步骤(1)中得出的主题向量，计算相邻时间窗口间主题的相似度，然后根据相似度建立主题关联；

(3) 主题关联过滤。制定三条主题关联过滤规则，对步骤(2)中形成的主题关联进行过滤，若经过滤后的主题关联仍有效，则认为主题之间存在演化关系；

(4) 主题演化关系判别。分析已建立的主题演化关系，依据时间先后顺序向前或向后推理，确定主题的演化关系类型。

整体分析框架如图 1 所示:

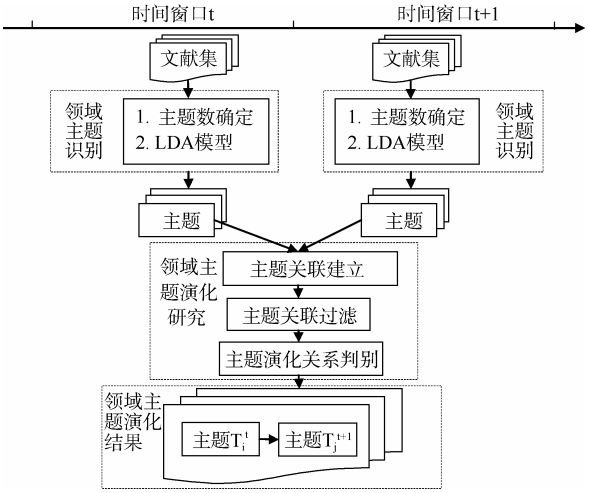

图 1　领域主题演化分析框架

3.3 领域主题识别

将时间序列划分为若干个长度为 L 的时间窗口，依据文献的出版时间划入到相应的时间窗口。假设时间窗口 t 的文献集为 C_t，每个时间窗口内的文献数不同，相应的主题数也应随着时间动态变化。因此作为 LDA 模型的输入参数之一，主题数目 K 必须事先确定，本文借鉴文献[20]的方法确定各时间窗口的主题数。

本文定义主题是分布在一组主题词上的向量，即：

$$T=\{(v_1,p_1),(v_2,p_2),\cdots,(v_i,p_i),\cdots,(v_n,p_n)\}$$

其中，v_i是与主题 T 相关的词，p_i是主题 T 在该词上的分布概率。在 LDA 中，时间窗口 t 中的第 i 个主题可以表示为：

$$T_i^t=\{(v_1,p(v_1|T_i^t)),(v_2,p(v_2|T_i^t)),\cdots,(v_j,p(v_j|T_i^t)),\cdots,(v_n,p(v_n|T_i^t))\}$$

其中，$v_j\in V$，V 为C_t的词表，$p(v_j|T_i^t)$为主题T_i^t中词v_j的概率。文献则定义为在这些主题上的多项分布，因此文献 d 可表示为：

$$d=\{(T_1^t,p(T_i^t|d)),(T_2^t,p(T_2^t|d)),\cdots,(T_i^t,p(T_i^t|d)),\cdots,(T_K^t,p(T_K^t|d))\}$$

其中，$p(T_i^t|d)$为主题T_i^t在文献 d 中的概率。

3.4 主题关联建立

经 LDA 识别出的不同时间窗口间的主题是相互独立的，为分析主题的演化需对相邻时间窗口的主题建立关联。本文认为主题的演化过程中必然存在内容的延续与改变，即相邻时间窗口间存在关联的两个主题具有内容相似性。通过计算相邻时间窗口间主题的相似度来衡量主题间内容的延续性，并建立主题关联。

(1) 主题相似度

LDA中每个主题在词汇上的概率分布可看作一个$1\times|V|$的向量，可采用余弦相似度[21]计算主题间的相似程度。主题T_i^t与主题T_j^{t+1}的相似度为：

$$Sim(T_i^t,T_j^{t+1})=\frac{|T_i^t|\times|T_j^{t+1}|}{\sqrt{(T_i^t)^2}\times\sqrt{(T_j^{t+1})^2}} \quad (1)$$

考虑到在计算过程中，文献集合不断引入新词汇，两个时间窗口中的词表不完全相同，根据LDA思想与Gibbs[20]抽样的估计方式，对于时间窗口 t 内未出现的词汇 w，主题T_i^t在 w 上的概率不为 0，而是符合一个统一的概率分布，笔者定义该分布公式为：

$$p(w|T_i^t)=\frac{\beta}{n_{T^t}^{(\cdot)}+V\times\beta} \quad (w\in V^{t+1},w\notin V^t) \quad (2)$$

其中，V 是V^t与V^{t+1}的并集，β 是 LDA 模型参数，$n_{T^t}^{(\cdot)}$是主题T_i^t包含的总词数。

(2) 建立主题关联

针对时间窗口 t 中的主题T_i^t，设时间窗口 t+1 内的主题按照与 T_i^t 的相似度从大到小排序，其中，与 T_i^t 相似度最大的是 T_j^{t+1}，记 T_j^{t+1} 为 T_i^t 的后向主题：

$$T_j^{t+1}=post(T_i^t)$$

针对时间窗口 t+1 中的主题 T_j^{t+1}，时间窗口 t 内的主题按照与 T_j^{t+1} 的相似度从大到小排序，其中，与 T_j^{t+1} 相似度最大的是 T_i^t，记 T_i^t 为 T_j^{t+1} 的前向主题：

$$T_i^t=prior(T_j^{t+1})$$

以上两种情况均认为T_i^t与T_j^{t+1}之间存在主题关联。

3.5 主题关联过滤

主题关联建立后，能够找出可能存在演化关系的主题对，文献[8-9]直接将上述主题关联认定为演化关系，但其结果会导致存在演化关系的两个主题相似度较低，主题延续性不明显。为提高主题演化关系的准确度，本文采用主题关联过滤规则去除无效的主题关联。

笔者定义以下三条过滤规则：

(1) 设定相似度阈值ε，时间窗口 t+1 内的主题与 T_i^t 相似度的最大值为 $MAX_{T_i^t}$，若 $MAX_{T_i^t}<\varepsilon$，则 T_i^t 与 T_j^{t+1} 之间主题关联无效；

(2) 如果 T_j^{t+1} 是 T_i^t 的后向主题，即时间窗口 t+1 内的主题按照与 T_i^t 的相似度从大到小排序，T_j^{t+1} 的排序位置为第 1；设时间窗口 t 内的主题按照与 T_j^{t+1} 的相似度从大到小排序，T_i^t 所在的排序位置为第 s (s>2)，若存在主题 T_k^t 所处的排序位置是$\rho\in(1,s)$，且$post(T_k^t)\neq T_j^{t+1}$，即 T_k^t 与 T_j^{t+1} 不能建立关联，则认为 T_j^{t+1} 与 T_i^t 的关联无效；

(3) 时间窗口 t 内的主题与T_j^{t+1}的相似度最大值为 $MAX_{T_j^{t+1}}$，按照与T_j^{t+1}相似度从大到小排序，T_i^t所处的排序位置为第 s (s>2)。设定阈值$\mu\in(0,1)$，当 $Sim(T_i^t,T_j^{t+1})<\mu\times MAX_{T_j^{t+1}}$ 时，T_i^t 与 T_j^{t+1} 的主题关联无效。

对于建立的主题关联，采用以上三条规则过滤，将过滤后仍存在关联的主题判定为具有较强的关联性，即主题具有演化关系。

3.6 主题演化关系的判别

为解决主题演化中演化关系的判别问题，本文对

经关联过滤后的关联主题进行前向、后向的推理分析，并把它们的关系分为新生、消亡、继承、分裂和合并 5 类，如图 2 所示：

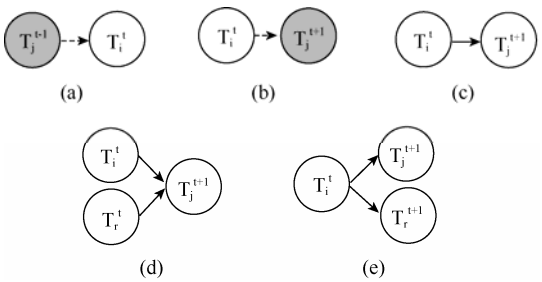

图 2　领域主题演化形式

其中：a 为主题新生；b 为主题消亡；c 为主题继承；d 为主题合并；e 为主题分裂。其中，虚线箭头表示主题间演化关系不存在。

(1) 新生：若主题 T_i^t 没有前向主题，时间窗口 t–1 中也不存在后向主题是 T_i^t 的主题，则认为 T_i^t 是 t 内新生的主题；

(2) 消亡：若主题 T_i^t 没有后向主题，时间窗口 t+1 中也不存在前向主题是 T_i^t 的主题，则认为 T_i^t 是 t 内消亡的主题；

(3) 继承：设主题 T_i^t 的后向主题是 T_j^{t+1}，且 T_j^{t+1} 的前向主题是 T_i^t，则认为 T_j^{t+1}、T_i^t 是同一主题在不同时间段的表示，即 T_j^{t+1} 继承于 T_i^t；

(4) 合并：主题 T_i^t 的后向主题是 T_j^{t+1}，而 T_j^{t+1} 的前向主题是 T_r^t，且 $r \neq i$，则认为 T_j^{t+1} 由 T_r^t 与 T_i^t 合并而得。若时间窗口 t 内有两个或两个以上主题的后向主题都是 T_j^{t+1}，同样认为这些主题在时间窗口 t+1 中合并为 T_j^{t+1}；

(5) 分裂：主题 T_j^{t+1} 的前向主题是 T_i^t，而 T_i^t 的后向主题是 T_r^{t+1}，且 $r \neq j$，则认为 T_j^{t+1}、T_r^{t+1} 由 T_i^t 分裂而得。若时间窗口 t+1 内有两个或两个以上主题的前向主题都是 T_i^t，则认为这些主题由 T_i^t 分裂而成。

4　实验及结果分析

4.1　数据预处理

为验证基于 LDA 主题关联过滤的领域主题演化方法的有效性，本文从 SCI 下载 2001 年–2012 年肿瘤领域 20 种期刊，类型是 Article 与 Proceeding Paper 的文献共 66 164 篇。

利用 NCI 叙词表 (National Cancer Institute thesaurus, NCIt)[22]，分别从标题、关键词、摘要中抽取肿瘤领域的词汇，并把同义词归并为同一形式。此外还利用文献自身的关键词对词表进行扩充。

将时间序列划分为 12 个长度 L 为 1 年的时间窗口，根据文献的出版时间把文献划入到相应的时间窗口。借鉴文献[20]的方法确定各个时间窗口的最优主题数。各时间窗口文献数、主题词数以及对应的最优主题数如表 2 所示：

表 2　数据集各时间窗口所含文献数、主题词数及最优主题数

文献集	文献数	主题词数	最优主题数
2001 年文献	5 157	10 339	103
2002 年文献	4 979	10 303	86
2003 年文献	5 698	11 090	110
2004 年文献	5 825	11 166	93
2005 年文献	6 396	11 255	105
2006 年文献	6 020	11 347	95
2007 年文献	5 870	11 151	100
2008 年文献	5 570	10 798	105
2009 年文献	5 566	12 106	121
2010 年文献	5 366	12 161	118
2011 年文献	4 882	11 638	102
2012 年文献	4 835	12 580	96

4.2　主题识别结果

利用 LDA 模型获取每个时间窗口的主题，选取各个主题中分布概率 Top8 的词汇作为特征词表示主题内容。表 3 是 2012 年的部分主题：

表 3　2012 年部分主题

主题	主题内容
T_{61}^{2012}	Breast Carcinoma, Estrogen, Estrogen Receptor, Progesterone Receptor, Progesterone, MCF7, Negative Lymph Node, Triple-Negative Breast Cancer Finding
T_{83}^{2012}	Immunity, CD8B1, Peptide, Antigen, Immunotherapy, CD4 Gene, Vaccine, T-Lymphocyte
T_{55}^{2012}	Radiation Therapy, Electromagnetic Radiation, Oncology, Rash and Dermatitis Adverse Event Associated with Radiation Excluding Chemo-radiation, Thoracic, Stereotaxic Radiosurgery, Ionizing Radiation, Radiation Therapy Oncology Group

(1) 模型识别的主题具有明确的研究方向。观察 T_{61}^{2012} “乳腺癌(Breast Carcinoma)”、T_{83}^{2012} “免疫(Immunity)”、T_{55}^{2012} “放射疗法(Radiation Therapy)”三个主题, 发现主题间内容差别较大, 区分明显。

(2) 各主题中分布概率较高的主题词能够涵盖该主题的内容。以主题 T_{61}^{2012}“乳腺癌(Breast Carcinoma)”为例, 美国国家癌症研究所发布[22]: 乳腺癌(Breast Carcinoma)是近些年常见癌症之一。另外根据实际文献观察, 近年来新辅助化疗越来越多地应用于可切除的乳腺癌治疗中, 但化疗对乳腺癌雌激素受体、孕激素受体的影响有很大争议, 不少学者为此展开研究。模型将雌激素(Estrogen)、雌激素受体蛋白(Estrogen Receptor)、助孕素(Progesterone)这些词汇归入同一主题中, 说明了该主题的涵盖内容。

4.3 主题关联过滤结果

按照 3.4 节中叙述的方法, 计算相邻时间窗口主题间的相似度, 并建立主题关联。根据 3.5 节的方法进行主题关联过滤, 其中过滤参数设置为: ε=0.20, μ=0.5。并与文献[9]中未经主题关联过滤的演化结果进行比较。

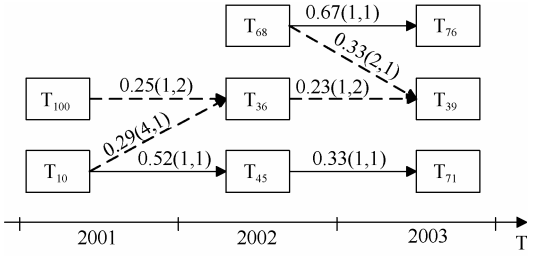

图 3 2001 年–2003 年部分主题的演化关系

图 3 是从结果中选取的 2001 年–2003 年若干个主题的演化图。箭头方向表示主题的演化方向, 其中实线箭头表示经本文方法过滤后的主题关联。箭头上数字表示箭头两端的主题之间的相似度大小及排序, 如 T_{100}^{2001} 与 T_{36}^{2002} 之间的数字显示: 两主题的相似度为 0.25, 2002 年的主题按照与 T_{100}^{2001} 的相似度从大到小排序, T_{36}^{2002} 所处排序位置为第 1; 2001 年的主题按照与 T_{36}^{2002} 的相似度从大到小排序, T_{100}^{2001} 所处排序位置为第 2。图 3 中各个主题内容如表 4 所示:

表 4 图 3 中各主题内容

主题	主题内容
T_{10}^{2001}	Immunity, Peptide, Antigen, Epitopes, Vaccine, Cytotoxic T-Lymphocyte, CD8B1, Immunotherapy
T_{100}^{2001}	PTGS2 Enzyme, Ultraviolet B Radiation, Sulindac, Carcinogenesis, Nonsteroidal Antiinflammatory Drug, Chemical, Epidermis
T_{36}^{2002}	Peptide, Enzyme, Synthesis, Molecule, Dose-dependent, Affinity, Modulation Action, Carcinogen
T_{45}^{2002}	Antigen, Epitopes, Immunotherapy, Vaccine, Immunity, Cytotoxic T-Lymphocyte, CD8B1, Interleukin-12
T_{68}^{2002}	Antibody, EGFR, Receptor, Growth Factor, Epidermal Growth Factor, Gefitinib, Monoclonal Antibody, Kinase
T_{39}^{2003}	Monoclonal Antibody, Molecule, Therapeutic, IgG, Affinity, Radiolabeled, Synthesis, Radioimmuno-therapy
T_{71}^{2003}	Immunotherapy, Vaccination, CD8B1, Vaccine, CD4 Gene, Immunity, Epitopes, Antigen
T_{76}^{2003}	EGFR, Receptor, Growth Factor, Epidermal Growth Factor, Gefitinib, Kinase, PTK2, Tyrosine Kinase Inhibitor

由图 3 可见, 未经主题关联过滤时: 2001 年–2002 年的主题演化中, 主题 T_{45}^{2002} 继承于 T_{10}^{2001}, 同时 T_{10}^{2001} 又与 T_{100}^{2001} 合并成 T_{36}^{2002}; 2002 年–2003 年的主题演化中, T_{71}^{2003} 继承于 T_{45}^{2002}, T_{76}^{2003} 继承于 T_{68}^{2002}, 同时 T_{68}^{2002} 又与 T_{36}^{2002} 合并生成 T_{39}^{2003}。而经本文方法的处理, T_{36}^{2002} 与 T_{10}^{2001}、T_{36}^{2002} 与 T_{100}^{2001}、T_{39}^{2003} 与 T_{68}^{2002}、T_{39}^{2003} 与 T_{36}^{2002} 的关联都被过滤。

在时间窗口 2003 内, T_{68}^{2002} 的相似主题中, T_{39}^{2003} 排序在第 2 位, 两者的相似度是 0.33, 不足第一位 T_{76}^{2003}(0.67)的 1/2。根据 3.5 节中过滤规则(3), T_{39}^{2003} 与 T_{68}^{2002} 的关联无效。另外, T_{39}^{2003} 与 T_{68}^{2002}、T_{71}^{2003} 与 T_{45}^{2002} 的相似度同为 0.33, 但两个演化关系的判定结果不同, 观察两对主题的主题词, 可发现 T_{71}^{2003} 与 T_{45}^{2002} 中概率较高的相同词更多, 延续性更强。反映出本文的过滤方法可以针对不同数据、不同主题特征灵活取值, 无需通过较强的专业知识设定固定阈值。

主题过滤方法还能够自动识别相似度较低的关联, 并将其过滤。如 2002 年中的主题按照与 T_{10}^{2001} 的相似度的大小排序, T_{36}^{2002} 仅占第 4 位, 而第 2 位和第 3 位的主题未与 T_{10}^{2001} 建立关联, 根据 3.5 节中过滤规则(2), T_{36}^{2002} 与 T_{10}^{2001} 的关联无效。另外在表 4 中可以看到, T_{10}^{2001} 与 T_{45}^{2002} 都是关于免疫与抗体的内容, 关

系更为密切, 而 T_{36}^{2002} 主要内容为缩氨酸, 与免疫的相关性不强。

4.4　主题演化结果分析

以主题“免疫(Immunity)”为例分析其自 2001 年–2012 年的演化情况, 如图 4 所示。可以看出 T_{61}^{2005} 是 2005 年的新生主题, T_{44}^{2008} 在 2008 年后消亡; 主题在整个时间段上的演化过程中, 2001 年–2006 年逐年继承, 在 2007 年、2011 年分别发生一次合并, 在 2008 年、2010 年分别发生一次分裂。图 4 中各主题的内容如表 5 所示。

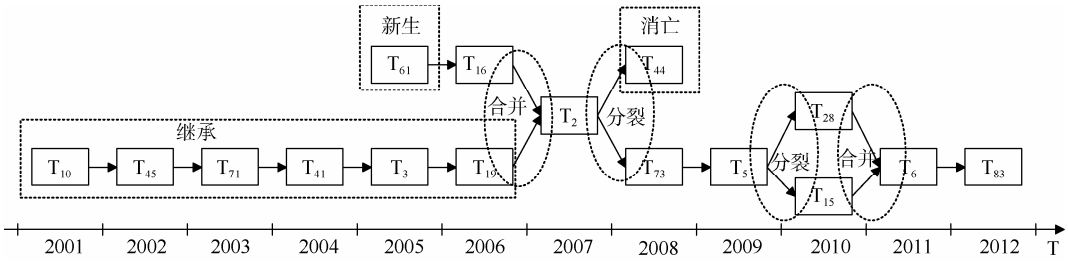

图 4　主题“免疫(Immunity)”2001 年–2012 年演化图

表 5　主题“免疫(Immunity)”的主题内容

主题	主题内容
T_{10}^{2001}	Immunity, Peptide, Antigen, Epitopes, Vaccine, Cytotoxic T-Lymphocyte, CD8B1, Immunotherapy
T_{45}^{2002}	Antigen, Epitopes, Immunotherapy, Vaccine, Immunity, Cytotoxic T-Lymphocyte, CD8B1, Interleukin-12
T_{71}^{2003}	Immunotherapy, Vaccination, CD8B1, Vaccine, CD4 Gene, Immunity, Epitopes, Antigen
T_{41}^{2004}	Vaccine, Vaccination, CD8B1, Immunotherapy, Immunity, Antitumor, Immune Response, Cytokine
T_3^{2005}	CD4 Gene, Cytokine, CD8B1, Therapeutic Interleukin-2, Interleukin, T-Lymphocyte, Macrophage, Immunity
T_{61}^{2005}	Antibody, Peptide, Monoclonal Antibody, Molecule, Therapeutic, IgG, Affinity, Radiolabeled
T_{19}^{2006}	CD8B1, CD4 Gene, Cytokine, Antigen, Cytotoxic T-Lymphocyte, Epitopes, Immunity, T-Lymphocyte
T_{16}^{2006}	Peptide, Endometrial, Immunotherapy, Vaccination, Vaccine, MHC Class-II Protein, Endometrial Carcinoma, STS Gene
T_2^{2007}	Peptide, CD4 Gene, CD8B1, Antigen, Immunotherapy, Immunity, Epitopes, Vaccine
T_{73}^{2008}	CD8B1, CD4 Gene, Cytokine, Immune Response, Macrophage, Lymphocyte, T-Lymphocyte, Immunity
T_{44}^{2008}	Antigen, Peptide, Immunotherapy, Vaccine, Vaccination, Epitopes, Cytotoxic T-Lymphocyte, Immunity
T_5^{2009}	Peptide, CD8B1, Vaccine, CD4 Gene, Immunity, Antigen, Vaccination, T-Lymphocyte
T_{15}^{2010}	Antitumor, CD8B1, T-Lymphocyte, CD4 Gene, Antigen, Immunotherapy, Macrophage, Therapeutic Interleukin-2
T_{28}^{2010}	Vaccination, Disease Response, Vaccine, Peptide, Immune Response, Efficacy, Therapeutic Interferon, Immunity
T_6^{2011}	CD8B1, Antitumor, Antigen, Immunotherapy, Vaccine, Immunity, T-Lymphocyte, Vaccination
T_{83}^{2012}	CD8B1, Peptide, Antigen, Immunotherapy, CD4 Gene, Vaccine, T-Lymphocyte, Immunity

(1) 根据图 4 与表 5 所示, 主题演化过程中主要内容明显, 而每年内容有细微变化。

由表 5 可以看出, 主题在整个时间段上的演化中免疫(Immunity)、抗体(Antigen)、CD8B1 等位基因(CD8B1)、免疫疗法(Immunotherapy)等主题词出现概率较大, 说明该主题主要内容是免疫的物质、方法等。同时, 随着时间变化, 主题中词汇概率每年都有上下波动, 如 CD8B1 等位基因(CD8B1)、CD4 基因(CD4 Gene)在 2001 年–2006 年的概率排序逐年升高, 反映该主题的研究中有很多微观基因的讨论。

(2) 主题在整个时间段的演化过程中, 存在主题的新生、消亡、继承、分裂和合并 5 种类型的演化关系。

主题 T_{61}^{2005} 的主要内容为抗体(Antibody)、缩氨酸(Peptide)等, 由于在 2004 年没有与其关联的主题, 被判定为新生主题; 同样, 由于在 2009 年没有与 T_{44}^{2008} 相关联的主题, 判定 T_{44}^{2008} 为消亡主题。

主题T_{19}^{2006}概率较高的主题词为: CD8B1等位基因(CD8B1)、CD4基因(CD4 Gene)、细胞活素(Cytokine), T_{16}^{2006}概率较高的主题词为: 缩氨酸(Peptide)、子宫内膜(Endometrial)、免疫疗法(Immunotherapy), 可了解到前者更侧重于基因等微观物质的研究, 后者则更侧重于免疫方法的讨论。这两个主题的主题词与T_2^{2007}都有较大的重合度, 相似度分别为 0.38 和 0.44, 可以推测T_2^{2007}是由T_{19}^{2006}与T_{16}^{2006}合并而成的。

2010 年, 主题 T_5^{2009} 分裂成两个主题 T_{15}^{2010} 和T_{28}^{2010}。根据其主题词可看出T_{15}^{2010}主要内容为抗癌抗菌素(Antitumor)、CD8B1 等位基因(CD8B1)等免疫物质, 而T_{28}^{2010}主要内容是针对种痘(Vaccination)、疾病反应(Disease Response)等现象的讨论。但两个主题均未脱离关于主题“免疫(Immunity)”的分析, 可看作“免疫(Immunity)”的两个子主题, 这些结果与对该领域的理解是一致的。

5 结 语

本文选用LDA主题模型, 采用主题关联过滤方法对领域主题的演化进行探索, 并以 SCI 中肿瘤领域的相关文献为例对该方法的有效性进行验证, 可以得出如下结论:

(1) 本文的方法有效地区分了主题的演化关系类型, 探测领域文献中主题的新生与消亡, 揭示主题继承时内容的变化, 反映主题的合并与分裂原因、合并内容以及分裂方向。

(2) 主题关联过滤有效地降低了LDA主题模型中相似度较小的主题的干扰, 从而提升主题演化关系(新生、消亡、继承、分裂和合并)的识别准确性。

(3) 过滤方法针对不同数据、不同主题特征灵活设定主题关联阈值, 解决了演化关系确定时对固定阈值大小过于敏感的问题。

但是, 本文把时间窗口大小设置为不可变, 未考虑主题演化周期的多样性。因此下一步的研究中将考虑根据不同主题的特点, 灵活设置时间窗口的大小, 以提高主题演化分析的合理性与准确性。

参考文献:

[1] 李勇, 安新颖. 基于 LDA 的主题演化研究[J]. 医学信息学杂志, 2013, 34(2): 57-61. (Li Yong, An Xinying. Research on Topic Evolution Based on LDA [J]. Journal of Medical Informatics, 2013, 34(2): 57-61.)

[2] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. The Journal of Machine Learning Research, 2003, 3: 993-1022.

[3] 楚克明, 李芳. 基于 LDA 模型的新闻话题的演化[J]. 计算机应用与软件, 2011, 28(4): 4-7, 26.(Chu Keming, Li Fang. LDA Model-based News Topic Evolution [J]. Computer Applications and Software, 2011, 28(4): 4-7, 26.)

[4] 楚克明. 基于LDA的新闻话题演化研究[D]. 上海: 上海交通大学, 2010.(Chu Keming. The Reaearch on Topic Evolution for News Based on LDA Model [D]. Shanghai: Shanghai Jiaotong University, 2010.)

[5] 李保利, 杨星. 基于 LDA 模型和话题过滤的研究主题演化分析[J]. 小型微型计算机系统, 2012, 33(12): 2738-2743. (Li Baoli, Yang Xing. Analyzing Research Topic Evolution with LDA and Topic Filtering [J]. Journal of Chinese Computer Systems, 2012, 33(12): 2738-2743.)

[6] 崔凯, 周斌, 贾焰, 等.一种基于 LDA 的在线主题演化挖掘模型[J]. 计算机科学, 2010, 37(11): 156-159, 193. (Cui Kai, Zhou Bin, Jia Yan, et al. LDA-based Model for Online Topic Evolution Mining [J]. Computer Science, 2010, 37(11): 156-159, 193.)

[7] 胡吉明, 陈果. 基于动态 LDA 主题模型的内容主题挖掘与演化[J]. 图书情报工作, 2014, 58(2): 138-142. (Hu Jiming, Chen Guo. Mining and Eolution of Content Topics Based on Dynamic LDA [J]. Library and Information Service, 2014, 58(2): 138-142.)

[8] Lv N, Luo J, Liu Y, et al. Analysis of Topic Evolution Based on Subtopic Similarity [C]. In: Proceedings of the 2009 International Conference on Computational Intelligence and Natural Computing, 2009, 2: 506-509.

[9] 胡艳丽, 白亮, 张维明. 一种话题演化建模与分析方法[J]. 自动化学报, 2012, 38(10): 1690-1697. (Hu Yanli, Bai Liang, Zhang Weiming. Modeling and Analyzing Topic Evolution [J]. Acta Automatic Sinica, 2012, 38(10): 1690-1697.)

[10] Blei D M, Lafferty J D. Dynamic Topic Models [C]. In: Proceedings of the 23rd International Conference on Machine Learning. 2006: 113-120.

[11] Alsumait L, Barbara D, Domeniconi C. On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking [C]. In: Proceeding of the 8th IEEE International Conference on Data Mining. IEEE, 2008: 3-12.

[12] Wang X, McCallum A. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends [C]. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006: 424-433.

[13] 贺亮, 李芳.科技文献话题演化研究[J]. 现代图书情报技术, 2012(4): 61-67. (He Liang, Li Fang. Topic Evolution in Scientific Literature [J]. New Technology of Library and

Information Service, 2012(4): 61-67.)

[14] 范云满, 马建霞. 利用 LDA 的领域新兴主题探测技术综述[J]. 现代图书情报技术, 2012(12): 58-65. (Fan Yunman, Ma Jianxia. Review on the LDA-based Techniques Detection for the Field Emerging Topic [J]. New Technology of Library and Information Service, 2012(12): 58-65.)

[15] 唐晓波, 王洪艳. 基于潜在狄利克雷分配模型的微博主题演化分析[J]. 情报学报, 2013, 32(3): 281-287. (Tang Xiaobo, Wang Hongyan. Analysis of Microblog Topic Evolution Based on Latent Dirichlet Allocation Model [J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(3): 281-287.)

[16] 史庆伟, 乔晓东, 徐硕, 等.作者主题演化模型及其在研究兴趣演化分析中的应用[J]. 情报学报, 2013, 32(9): 912-919. (Shi Qingwei, Qiao Xiaodong, Xu Shuo, et al. Author-topic Evolution Model and Its Application in Analysis of Research Interests Evolution [J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(9): 912-919.)

[17] Xu S, Shi Q, Qiao X, et al. Author-topic over Time (AToT): A Dynamic Users' Interest Model [A].// Mobile, Ubiquitous, and Intelligent Computing [M]. Springer Berlin Heidelberg, 2014: 239-245.

[18] 单斌, 李芳. 基于 LDA 话题演化研究方法综述[J]. 中文信息学报, 2010, 24(6): 43-49, 68. (Shan Bin, Li Fang. A Survey of Topic Evolution Based on LDA [J]. Journal of Chinese Information Processing, 2010, 24(6): 43-49, 68.)

[19] Wei X, Sun J, Wang X. Dynamic Mixture Models for Multiple Timeseries [C]. In: Proceedings of the 20th International Joint Conference on Artificial Intelligent, Hyderabad, India. 2007: 2909-2914.

[20] Griffiths T L, Steyvers M. Finding Scientific Topics [C]. In: Proceedings of the National Academy of Sciences of the United States of America. 2004: 5228-5235.

[21] Manning C D, Schütze H, Raghavan P. 信息检索导论[M]. 王斌译. 北京: 人民邮电出版社, 2011. (Manning C D, Schütze H, Raghavan P. Introduction to Information Retrieval [M]. Translated by Wang Bin. Beijing: Post & Telecom Press, 2011.)

[22] National Cancer Institute. NCI Thesaurus Hierarchy [EB/OL]. [2014-02-14]. http://ncim.nci.nih.gov/ncimbrowser/pages/source_hierarchy.jsf?&sab=NCI.

作者贡献声明:

秦晓慧: 文献调研, 细化研究方向及技术方法路线, 设计实验方案, 数据采集、清洗与结构化, 编程及实验结果分析, 论文撰写与最终版本修订;

乐小虬: 提出论文研究方向和思路, 设计研究方案及技术路线, 修改文章部分章节。

收稿日期: 2014-10-08
收修改稿日期: 2014-11-13

Topic Evolution Research on a Certain Field Based on LDA Topic Association Filter

Qin Xiaohui[1, 2] Le Xiaoqiu[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] To detect the birth, extinction, development, merge and split of topic evolution of the literatures in a certain field. [**Methods**] This paper divides time windows according to the publication data of the literatures, and LDA model is applied to extract topics from each time window automatically. The topic association filter rules are used to determine evolution relationships between topics in adjacent time windows. Form a topic evolution path in a continuous time period. [**Results**] Considering the continuity of the topics, different types of topic evolution could be detected with high accuracy. [**Limitations**] This method fixes the size of time windows without considering the diversity of topic evolution cycles. [**Conclusions**] This method can effectively reduce the interference of topics with smaller similarity in LDA, and enhance accuracy of evolution relation recognition.

Keywords: Topic association Topic evolution Topic model LDA