基于LDA模型的科技期刊主题演化研究

李湘东 张 娇 袁 满

(武汉大学信息管理学院 武汉 430072)

摘 要 提出一种基于LDA(Latent Dirichlet Allocation)潜在语义模型、全面研究科技期刊主题演化过程的方法。该方法根据科技期刊的特点引入时间因素,使用困惑度确定最优主题数目,通过LDA主题提取结果及JS散度,实现主题在强度和内容两方面的演化研究,并对不同时间窗口的主题稳定性做出相应分析。实验结果表明该方法可以较好地分析某一特定科技期刊的主题随时间的强度演化规律以及主题内容的演化趋势。

关键词 LDA模型 科技期刊 主题演化 主题稳定性

中图分类号 G350 **文献标识码** A **文章编号** 1002-1965(2014)07-0115-07

DOI 10.3969/j.issn.1002-1965.2014.07.021

On Topic Evolution of a Scientific Journal Based on LDA Model

Li Xiangdong Zhang Jiao Yuan Man

(School of Information Management, Wuhan University, Wuhan 430072)

Abstract This paper proposed a method based on LDA (Latent Dirichlet Allocation) Latent semantic model to comprehensively study scientific journal's topic evolution, which according to the characteristics of the scientific journals, added time factor into the model, then confused degree method was used to determine the optimal number of topics. Through LDA topic extraction and JS divergence, we tracked the journal topic trends in different time intervals and conducted topic evolution analysis on aspects of both contents and strength, and the stability of topics over time were also included. The results show that this method can analyze scientific journal's topic strength evolution over time better, in addition, the evolution trend of the journal topics contents can be described as well.

Key words LDA model scientific journal topic evolution topic stability

0 引 言

随着计算机与互联网的蓬勃发展,科技期刊由于其承载着科学研究在各个主题方面的最新成果而日益重要。因此,对科技期刊(以下简称期刊)主题进行研究具有重要的意义。期刊主题演化是指期刊主题在时间维度上的发展变化过程,主要是分析某一时期内期刊主题的变化程度以及不同主题随时间变化的规律,它有助于揭示学科领域在不同时期研究内容的变化程度及研究重点。近年来,使用不同方法对学科领域的研究主题进行分析,识别新兴领域、主题演化趋势成为科学结构及其演化研究的前沿热点问题。通过分析学科结构及其演化过程来把握不同领域研究主题的发生、发展、转移、汇聚等趋势,有助于科学家更快地捕获学科发展前沿及热点,有助于为科研决策者和管理者提供决策依据[1]。

1 相关工作

对期刊主题展开演化研究的方法众多,从定性分析到定量分析,从简单图表到信息可视化分析,从基于词频统计到基于模型的机器学习方法等[2]。这些传统的方法大都以文本中出现的词语作为主题的代表开展统计和分析,由于未涉及文本或者词语所包含的语义,使用这些传统方法开展主题演化研究的成果较为宏观和粗略,且受制于文本来源的期刊编辑部、数据库商或者研究者本人等人为设置的主题范围。LDA模型由于可以很好地模拟大规模语料的语义信息的特点开始引入到期刊主题演化研究领域。

1.1 LDA建模过程中主题数目的确定 在LDA建模时,现有研究通常采用文本聚类、贝叶斯模型或者

收稿日期:2014-03-29 修回日期:2014-05-04

作者简介:李湘东(1963-),男,博士,副教授,研究方向:信息检索、数据挖掘以及文本自动分类;张 娇(1989-),女,硕士研究生,研究方向:信息检索与数据挖掘;袁 满(1991-),男,硕士研究生,研究方向:信息计量。

困惑度计算等方法确定最优主题数目。学者贺亮[3]利用LDA模型抽取科技文献话题时,使用文本聚类方法确定主题提取数目,研究了ACL文集和NIPS文集的主题特性,文献[4]也用到此方法。由于聚类需要设定聚类的终止条件,操作需要人工干预,与传统方法研究期刊主题时一样,主题数目的确定掺杂较多的主观性和随意性,研究者们也在探索其他的方法。如学者Griffith T L,Steyvers M.[5]在验证所提出的马尔科夫蒙特卡洛方法(Markov Chain Monte Carlo:MCMC)可以近似推理LDA模型参数时,采用贝叶斯模型确定主题数目,提取PANS文集的隐含主题,并跟踪文档之间的相似性变化,随后文献[6]和[7]也参考该方法确定主题数目,对相关文本材料进行研究。贝叶斯模型有较好的数学原理作为其理论基础,但缺乏直观性。在使用LDA为期刊主题建模的研究中,更多学者[8-12]使用了更为直观的困惑度计算方法确定主题数目。作为语言模型中标准的评判准则,困惑度计算可以对分组数据建模,来评价各种模型的性能,困惑度越低,说明模型的泛化能力越强。该方法和贝叶斯模型一样,既具有较好的数学原理作为其理论基础,同时也不需要人为设定参数,即主题数目的确定排除了主观性和人为性,能够客观一致地体现出分析对象实际所包含的隐含主题。

1.2 基于LDA的期刊主题演化研究内容 主题演化是主题随时间逐渐发生改变的过程,其中包括新主题的产生、主题的消亡、主题的强度,以及主题内容向其他主题的迁移或渗透等,还包括主题的属性的变化,即主题本身的特征研究。

一方面,现有文献使用LDA分析期刊的主题属性时,研究内容主要有主题挖掘[11,13]、热点主题和冷门主题识别[5-6]、影响力主题[3]识别,还包括主题的交叉和融合[7,12]。具体看来,其中文献[3]定义了一种影响力话题的判定公式,基于文档的相似值判定,若某个主题被越多的主题关联,则它越具有影响力,并以期刊论文集进行实验证明方法的有效性;文献[6]对PNAS文集提取隐含主题,识别热点主题和冷门主题并分析其强度变化;文献[12]对期刊LIS(Library Information Science)1930-2009年间的文集主题建模,得到各时间段的主题-词汇概率表,并对高频率主题排名,基于相似度衡量比较不同阶段的主题交叉情况。

另一方面,主题的趋势演化研究主要包括主题的强度演化和内容演化,其中内容演化又可分为两类:一是局部主题内容演化,例如计算每个主题的内容偏移量及其随时间的演化趋势以及每个主题的内容发展脉络;二是全局的主题内容演化,即整个全局时间段内主题内容变迁的宏观发展趋势。学者李勇等[9]以传染病学领域代表性期刊LID(Lancet Infectious Diseases)为研究对象,采用基于KL(Kullback-Leibler)距离的主题相似性度量识别同一主题,通过监测不同时间段的主题-词汇分布,再用Z-Score方法计算同一主题随时间的偏移值,重点了分析HIV主题偏移量演化趋势。学者Andrew McCallum et. al[14]结合TOT(Topics over time)和LDA模型,对NIPS文献集的隐含主题也进行主题偏移的相关研究。其次,文献[5,8]是对文本材料的主题强度演化研究,文献[11]中学者王萍对三种国际教育技术期刊的进行主题挖掘,基于LDA文档建模和相似度测量,对具有相似主题内容的文档,使用多维坐标MDS(Multidimensional Scaling)方法直观地输出文档随时间的簇集分布情况。文献[10]中学者张才东等人获取干细胞领域的文本集,提取和识别具体主题,进而分析同一主题的内容偏移计算及其随时间的演变。也有一些学者[3,4,7]以多种期刊为实验对象研究,涉及的研究主题内容较全面。如文献[4]以管理学领域内十种主要期刊为研究对象,使用基于LDA模型的CTM模型(相关主题模型),对主题的属性特征以及主题的内容演化做了详细地实证分析;文献[3]对于ACL和NIPS两个文集进行LDA主题建模,通过文档-主题矩阵计算主题的文档支持率值作为其强度值,进而刻画随时间的强度变化;通过主题-词汇分布计算话题词在不同时间段的权重,得到话题的内容演化路径。文献[7]采用LDA对管理学领域三种核心期刊的4244篇文章的标题、摘要、关键词和发表时间的整合文本进行建模,通过主题间转移概率的计算,设置阈值构建主题的时序链,实现主题强度曲线和主题内容发展脉络的演化。

总之,现有的文献使用LDA模型研究期刊主题时,一是并没有很好地发挥LDA模型的数据效应,导致主题的特征研究不够全面和深入。二是对于某种特定期刊进行详细的实践研究时,仅仅涵盖了主题的少数方面特征的演化,对期刊主题演化研究在实践中的应用缺乏指导性和具体的可操作性。

本文在前人的研究基础上,将时间因素引入LDA建模中,采用困惑度计算方法确定最优主题提取数目,以Journal of Information Science期刊为实验材料,阐述了如何使用LDA建模研究某一期刊的主题特征演化,即通过LDA主题建模从大量的文献数据中获得抽象的主题信息,利用提取的主题词汇以及文献的发表时间来构建主题关联网络,用JS距离对期刊主题相似度进行衡量,并利用时间段的连续性分析主题的强度变化和主题内容的覆盖情况。最后通过实验,深入挖掘LDA模型数据,从活跃主题识别、主题强度计算、主题内容变化趋势分析以及主题稳定性评价方面比较系

统和全面地分析该期刊主题的演化特征,并根据实际背景对演化结果进行合理的讨论。

2　LDA 基本原理

为了更加符合现实世界中的文本特点,学者 Blei 等在 2003 年提出了潜在狄利克雷分配(Latent Dirichlet Allocation)模型[15],简称 LDA 模型。它是一个三层贝叶斯概率生成模型,主要思想是假设每个文档均是由多个主题混合而成,而每个主题则是多个词汇上的概率分布[16]。其后,T. L. Giffiths 在研究中对该模型做了改进,增加了主题 - 词汇分布的 Dirichlet 分布先验,改进后的模型如图 1 所示[17]。

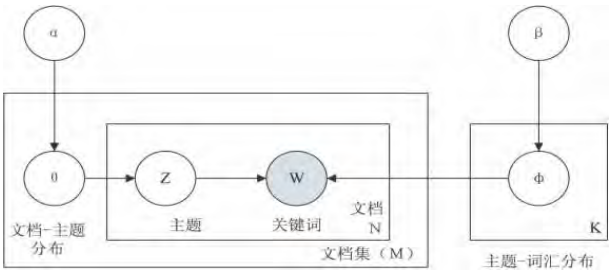

图 1　LDA 模型图

图中隐含变量包括超参数 α、超参数 β、文档 - 主题分布和主题 - 词汇分布,可观察变量仅有关键词 w,有向箭头代表条件概率,方框 M 代表的是文档集中每个文本迭代抽样的主题概率分布,方框 N 表示从主题分布中迭代抽取产生的文档的词汇。LDA

模型中生成文档的步骤如下所示:

a. 对于每个文档 $d \in D$,根据 $\theta_d \sim \mathrm{Dir}(\alpha)$,得到文档 d 上主题的多项式分布参数 θ_d;

b. 对于每个主题 $z \in K$,根据 $\varphi_Z \sim \mathrm{Dir}(\beta)$,得到主题 z 上词汇的多项式分布参数 φ_z;

c. 对于文档 d 中的词汇 $w_{d,j}$,根据多项分布 $z_{d,j} \sim \mathrm{Mult}(\varphi_z)$,得到主题 $z_{d,j}$,根据多项分布 $w_{d,j} \sim \mathrm{Mult}(\varphi_z)$,得到词汇 $w_{d,j}$。

在 LDA 概率生成模型中无法直接获得参数 θ 和 φ,因此,一般通过参数估计的方法近似推理参数值,常用的方法有变分推理(Variational Inference)、Laplace 近似、Gibbs 抽样算法和期望 - 扩散(Expectation Propagation)。其中 Gibbs 抽样算法实现简单,计算速度快,现广泛地应用 LDA 概率生成模型中。本文参考文献[12]的 Gibbs 抽样算法求解 LDA 模型的相关参数。

3　期刊主题演化研究框架

3.1　**演化模型**　对于期刊主题而言,演化主要由主题的三个方面组成,一是主题在时间轴上的强度变化,即主题的活跃度变化;二是主题在内容上的变化,即主题的内容迁移;三是主题的社会网络变化,即作者或科研机构的变化。对于本文而言,主要探讨前两种演化模式。以下是建模过程中的相关问题阐述。

3.1.1　**文本预处理**　本文对英文期刊进行主题演化分析实验,其预处理工作主要包括分词、去停用词、词根还原以及去标记信息等。英文文本由于其语言特点,可直接通过文本的空格、标点符号将单词划分。去停用词主要是将那些不能为文本分析提供有效信息的词语去除,通常如英文中的助词、代词、连词及副词等;其次,针对实验材料的特点,本文将停用词表进行扩充,加入一些本实验特有的、出现频率高且对实验结果没有意义的词语(如 journal; study; base 等),以避免对实验结果造成影响。词根还原则是将英文单词恢复成词根,这样处理之后可以较大程度地减少实验材料所包含特征项的数量,提高效率。本研究利用 Lucene 开源工具包完成对英文文本的去停用词和词根还原工作。

3.1.2　**困惑度确定模型最优主题数目**　主题是一个抽象的概念,根据不同的粒度划分语料集中的主题可以得到不同的主题数量,在 LDA 模型中主题的数量 K 需要预先给定,一般来讲语料集越大主题的数量越多,在不同的时间窗口主题的数量也是动态改变的。本文使用统计语言模型中常用的评价指标即困惑度[18]来确定最优主题数。困惑度为文档集中包含的各句子相似性几何均值的倒数,随句子相似性的增加而逐步递减。计算公式为

$$\mathrm{Perplexity(D)} = \exp\{-\frac{\sum_{m=1}^{M}\log_D(w_m)}{\sum_{m=1}^{M}N_m}\} \quad (1)$$

其中:N_m 表示第 m 个文档的长度,M 为文档集,$p(w_m)$ 表示 LDA 产生文档的概率:

$$p(w_m) = \sum_d \prod_{n=1}^{N}\sum_{j=1}^{T} p(w_j \mid z_j = j) \cdot p(z_j = j \mid w_m) \cdot p(d) \quad (2)$$

困惑度随着主题数目的增加而逐渐减小,困惑度越低模型的泛化能力越强,性能越优。

3.2　**期刊主题演化的度量方法**

3.2.1　**主题相似性度量**　本文采用 Jensen - Shannon 散度方法来衡量主题相似度。对于测试文本集合 C,其词汇表为 V,经过 LDA 建模后,可以得到词汇 V 的分布,也就是 LDA 中的主题 T,为了将 LDA 模型的主题扩展到测试集中,本文采用平滑技术将不同主题的词分布进行扩展,汇总到一个整体的词汇表上。假设经过 LDA 建模之后有 T_1、T_2 这两个主题,其中 T_1 是词汇表 V_1 上的分布,T_2 是词汇表 V_2 上的分布,利用平滑技术将 T_1,T_2 扩展为 T_1',T_2',此时 T_1',T_2' 是在词表 $V = V_1 \cup V_2$ 上的分布,假定 P 是主题 T_1' 在词表 V 上的分布,q 是主题 T_2' 在词表 V 上的分布。则 T_1' 和

T_2' 的主题相似度使用 JS 距离衡量为：

$$\text{Similarity}(T_1', T_2') = \text{Jensen-Shannon}(p,q) \quad (3)$$

其中

$$\text{Jensen-Shannon}(p,q) = 1/2\ [D(p \mid\mid m) + D(q \mid\mid m)]\ m = 1/2\ [p+q]$$

JS 距离的区间为 [0,1]，两个主题的 JS 距离越小，相似度就越大，关联的可能性也越大。因此本文通过设定阈值来决定期刊各年间刊载的论文主题之间是否具有演化关系。即当两个主题的相似度小于该阈值时，就判定这两个主题在内容上是具有继承关系的。

3.2.2 主题强度度量 主题强度主要描述了主题热门程度，在某一时刻关于某个主题的文章数量越多，说明该主题的强度越高，可以被认为是热门主题。本文借鉴了文献 [19] 的主题强度度量方法，假定 θ_k^d 为文档 d 中主题 k 所占的比例，则 t 时间窗口 k 主题的强度 θ_k^t 表示为

$$\theta_k^t = \frac{\sum_{d-1}^{M} \theta_k^d}{M} \quad (4)$$

用 θ_k^t 表示当前时间窗口的主题 k 的强度，可以得出期刊主题文档中一系列随时间变化的主题强度的不同值，依据此可绘制出主题强度变化曲线图，从宏观上分析该主题的趋势。同时，主题强度配合代表主题的各关键词和文章列表联合分析可以得出期刊主题文档中一系列随时间变化的主题强度的不同值，依据此可绘制出主题强度变化曲线图，从宏观上分析该主题的趋势。同时，主题强度配合代表主题的各关键词和文章列表联合分析可以得出该主题具体的含义，区分背景词汇和通用词汇所代表的主题，识别出有价值的主题。

3.2.3 主题稳定性 期刊主题稳定性分析是观察期刊所收录论文的主题在一定时间周期里的发展及波动状况，并对这种波动进行分析，期望为描述期刊不同方面的发展趋势做出依据。基于此本文在期刊主题强度演化的基础上，借鉴统计学中变异系数的思想提出主题稳定性度量方法，用来衡量期刊主题的稳定程度，稳定性系数具体定义为：对主题的一系列强度值 θ_k^t，计算其在所有时间窗口中的变异系数 V_k，然后计算所有主题的变异系数的平均值，得到整个时间段的主题稳定性系数 V。具体计算公式如下：

$$V_k = \frac{\sigma_{\theta_k^t}}{\bar{X}_{\theta_k^t}} \ (\sigma \text{ 为标准差，} \bar{X} \text{ 为平均值}) \quad (5)$$

$$V = \text{mean}(V_k) \quad (6)$$

相对于影响因子来说，期刊的稳定性是一个相对的评判标准。如果主题的发展缓慢或波动较小，那么稳定性相对较好，相反如果主题发展剧烈或者波动较大则稳定性差。对期刊的稳定性研究有助于发现学科的发展趋势以及学科演变、发展的规律。

4 实验结果与讨论

本实验选取信息科学领域中较为权威的期刊《Journal of Information Science》(以下简称 JIS)作为研究对象，从 ProQuest 数据库中抽取了其 1981－2011 之间刊载的学术性文章的内容(共 2 608 篇论文条目)，选取内容特征中的标题、主题词、摘要组成短小文本作为实验材料，并将期刊数据按年代划分成 31 个时间窗口并存档。由于论文发表存在较显著的时滞性，2011 年 JIS 文献数据并不全面，不能完全客观反映该年度的科研成果，因此，本文实验结果主要呈现 30 个时间窗口的主题分析。

4.1 期刊隐含主题的挖掘结果 对 JIS 文献主题建模时，根据困惑度确定最优主题数目，实验表明当主题数设为 21 时困惑度值最小，因此最优主题数目为 21。

本实验设定每个主题下关键词的提取个数为 40 个，各词汇按概率依次由大到小输出，由于篇幅原因，此处列举了 6 个最活跃主题，并选取每个主题包含的概率排在 top10 的关键词作为主题的含义代表，并结合其他输出词汇提供的信息，得到热点主题的识别结果，如表 1 所示。

表 1 热点主题识别结果

Topic 0	Topic 1	Topic3	Topic4	Topic5	Topic9
information	information	data	electronic	electronic	studies
science	knowledge	systems	technology	sites	library
analysis	systems	online	uk	analysis	academic
retrieval	libraries	statistical	system	presented	citation
methods	research	analysis	issues	science	universities
algorithms	management	searches	region	activities	colleges
networks	development	project	usage	web	publishing
comparative	theory	models	search	journal	technological
index	dissemination	metadata	national	internet	communication
web	social	mining	area	network	copyright

由此可知，JIS 文献的主题集中在 Knowledge management，Information science theory and policy，Application and practice，Information technology and processing，Education and government。通过主题－词汇的分

布可以具体分析主题的内容含义，例如主题 0 主要是信息检索领域的内容，包括 retrieval，algorithms，networks，index，methods 侧重于方法的创新；主题 1 则主要是信息管理领域的内容，包括 knowledge management，information dissemination，social ，research，theory 侧重于理论的研究和探索；主题 3 主要是数据挖掘领域，包括 metadata，statistical，models，analysis，project 侧重于对数据统计分析方法的研究；主题 4 主要是信息领域各国的发展状况，包括 journals，uk，region，national，issues 探讨世界范围内的信息科学发展对比研究；主题 5 是计算机与网络相关的，包括 electronic，internet，web，sites，activities 侧重于信息领域对新兴技术的应用；主题 9 主要是图书情报与教育领域，包括 library，colleges，universities，academic，citation 探讨图书情报在高校图书馆的发展及科研机构的科研能力评价。

4.2　期刊主题的强度演化结果　为了能够从宏观上更好的展现主题强度，本文将实验得到的主题强度值分段进行呈现，利用线性分段方法，对不同时期的主题强度演变进行分析（本文选取 10 个主题为例呈现结果），依次分为 1980s、1990s、2000s 三个时间段，并对每一个时间段的文档进行主题强度计算，得出每个主题的强度变化规律。

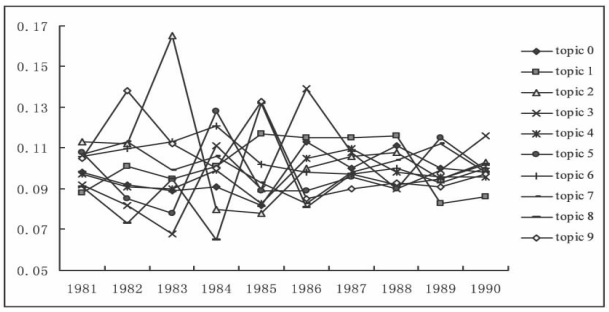

图 2　1980s 主题强度演化趋势

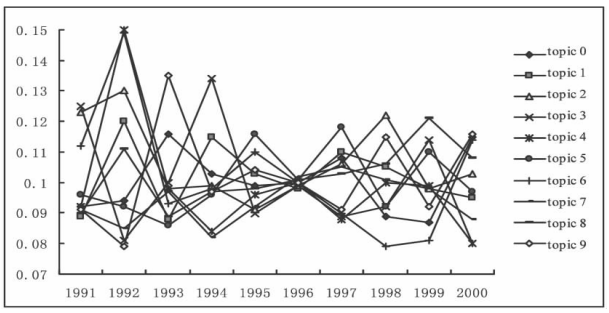

图 3　1990s 主题强度演化趋势

图 2、图 3 和图 4 分别展示了这三个时间段所选取 10 个主题的强度变化趋势图，其中横轴代表时间，纵轴主题的强度值。通过对比和分析三个时间段内 10 个主题的强度变化情况，可得出以下结论：1980s 的前期同一时间主题的强度值跨度较大，主题之间的强度值差异也较大，而到了后期主题强度差异减小，并趋于平衡和稳定，说明这一时期期刊主题波动不大，主题较分散；1990s 的主题强度特点是前期有些主题强度特别大，持续性的集中报道较多，在 1996 年出现汇聚，各个主题的强度几乎相等，说明该年度的主题分布比较平均，同时也说明一些高热度的主题也开始降温；2000s 的趋势与 1980s 情况正好相反，前期的主题强度值跨度较小，最小的仅限于 0.03，说明报道的主题的较分散，且这期间主题的波动不大，但是在后期主题强度值的波动性非常大，某些主题的强度特别高，说明该阶段出现了新的研究热点。

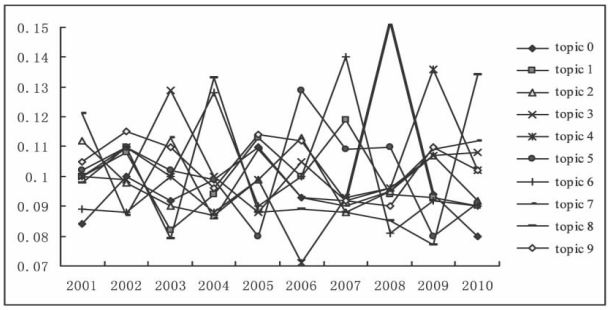

图 4　2000s 主题强度演化趋势

其次，通过本文提出的主题稳定性指标公式得到 1980s 的主题稳定指数是 8.19，1990s 的主题稳定指数是 14.71，2000s 的主题稳定指数是 6.83。由于稳定指数越低说明稳定性越强，说明期刊刊载的文章内容具有较强的连续性，并注意主题的均衡，因此从 JIS 主题稳定指数的动态变化来看，1980s 稳定指数较高，这个跟 1980s 刊载的文章篇数有一定的关系，因为这个时间段，该期刊刊载的文献数量不稳定，同时也与期刊的办刊能力和办刊风格有很大的关系。1990s 该期刊的主题稳定指数很高，稳定性较差，而该时期该期刊的文献数量逐渐稳定，这与当时的期刊主题热点变动有很大关系，据估计与信息科学领域的学科结构或新旧主题交替的原因有关，因此本文认为需要对具体的主题强度进行研究，以分析出原因。

通过追踪 30 年间的主题热度的变化趋势，提取出在 1990s 非常活跃的主题进行判断，强度演化如图 5 所示，图中横轴表示时间，纵轴代表活跃主题的强度。从图 5 可以看出，1990s 出现了较多新的主题，比如 Electronic information，Online database，Electronic journal

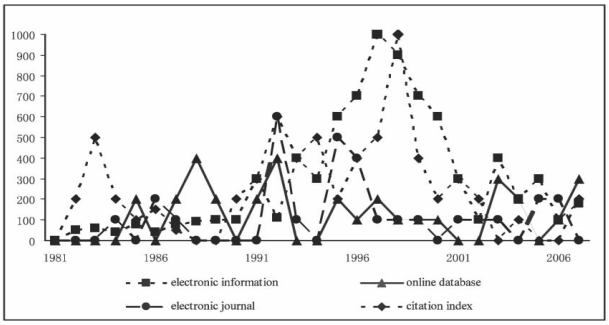

图 5　活跃主题的强度变化

等主题，这是由于1993美国推出“信息高速公路”计划，互联网、通信等技术飞速发展，对信息学科也产生了很大的影响，电子信息也逐渐成为学者关注的焦点。因此，这些主题自出现后持续受到学术界较高的关注，这些变化对1990s主题的分布产生了影响，因而引起主题强度趋势的变化。其次，可以发现该期刊非常注重信息理论的研究，在1990s期间引文相关的研究相当活跃，其中Citation index，Citation analysis热度很高，这期间学术界对于引文分析进行了改进研究，特别是提出了许多对影响因子的改进或者补充研究。1998年，美国科技信息研究所所长尤金·加菲尔德(Eugene Garfield)博士在《科学家》(The Scientists)杂志中叙述了影响因子的产生过程，引起了学术界的热烈讨论。同时，随着全球化趋势的到来，在1990s图书情报领域出现了许多世界范围内的学术研究能力、学科研究发展状况等对比研究，采用引文分析方法对世界各地区的高校和科研机构的学术能力以及不同学科结构之间的对比研究。

4.3 期刊主题的内容演化结果 通过LDA建模工具，已获得不同主题的主题－词汇概率分布以及主题在不同文档上的概率分布。因此，主题之间的关联则转变成两个概率密度的关联，根据JS距离公式，得到主题相似性度量结果。其次，通过分析不同时间窗口主题词的特点，获得主题含义，并将主题内容映射到不同的时间窗口内，最后分析主题内容演化。具体步骤为：

a. 根据LDA主题提取获得相邻时间窗口的文档－主题分布及主题－词汇分布，并利用JS相似度计算，获得主题相似度的分布矩阵。

b. 根据矩阵分布，设定主题相关度的阈值进行主题间关系的判定，如果取值低于阈值则说明主题相关性较高，具有延续性，如果取值高于阈值则说明主题相关性较低，反映了主题的消亡现象。

c. 根据连续多年的主题间相似度情况，构建内容演化网络图。

为了能够较好的展现多年主题变迁的宏观发展趋势，把握JIS期刊主题内容的全局演化，本文将以上获得的实验数据，导入内容分析和文本挖掘工具WordStat中，研究主题内容的发展脉络。WordStat包含大量探测性的数据分析和文字工具，可以用来检测文档信息内容之间的关系，使用其等级聚类和多维标度分析可以断定词或类之间的关系，相关性分析(Correspondence Analysis)则可以断定关键词和不同组别个体之间的关系[20]。通过对JIS期刊的主题随时间的相关性分析，得到主题的内容演化如图6所示。

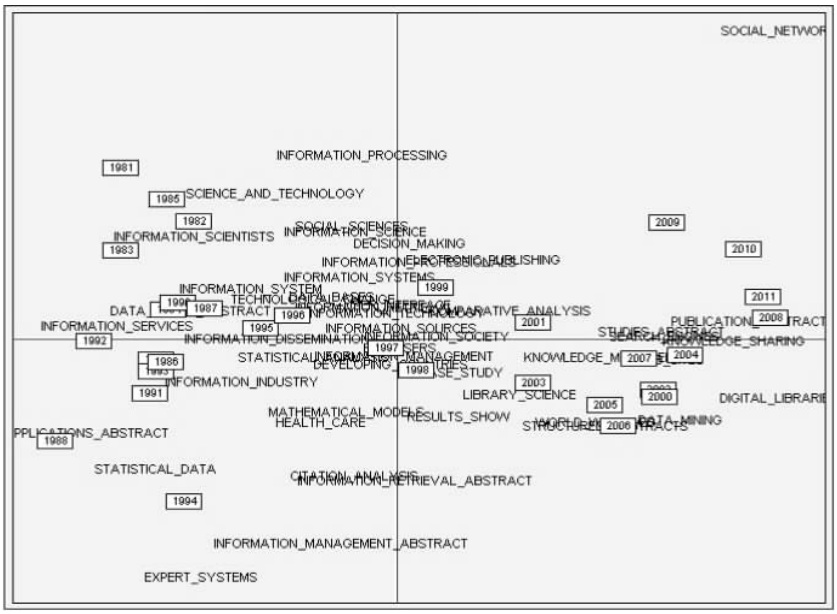

图6 JIS期刊主题内容演化结果

图中每个年代值的位置代表了该年度主题的集中趋势，整体图形展现了随着时间的变化，研究主题的逐渐更新和变迁。1980s的大部分年代和1995，1996分布在一个象限，侧重于信息服务方面，如Information Service，Information Process，Information System等主题领域，1990s的大部分年代集中在信息科学领域的相关理论和应用，如Expert System，Citation Analysis，Statistical Data，Health Care等，2000s以后逐渐趋向于计算机网络技术和数据挖掘，电子图书馆等，如Digital Library，Data Mining，Retrieval Abstract等。从这些离散的年代分布图，可以看出主题随时间变化的大小，如1983与1984的距离较远，说明主题的变动较大，而1995和1996的距离较近，说明主题的变动缓慢，稳定性较强。而这一趋势与学者用自组织映射的方法[21]得出的研究结果是相吻合的。同时，观察1990s的相对分布位置可以发现，相比较1980s和2000s，1990s中的年代分布地较为分散，如1994，1992其他相近年份分隔较远，说明这段时间主题的差异较大，这也很好地论证了本文在主题强度演化1990s期间的主题稳定性相对较差的结论。

5 结论与展望

本文在近年来国内外研究工作的基础上，提出了基于LDA模型和JS散度相结合的、使用困惑度自动确定文本建模的最优主题数，以研究某一特定期刊的

主题演化趋势方法,从强度、内容和稳定性等方面,全面研究了期刊主题的演化特点。结论表明该方法可以较好地分析某一特定期刊的主题随时间的强度演化规律以及主题内容的演化趋势。

然而,为了增强本研究的实践意义,可以扩展期刊主题的研究模式,加入如主题生命周期、主题寿命等其他更加复杂的影响主题的因素,从而更加深入地研究期刊主题的特征演化规律。其次,对期刊主题的稳定性测度,还需进一步系统的研究,力图为期刊的评价以及更深层次的期刊主题趋势预测提供更多的依据,这是我们下一步的研究方向。

参 考 文 献

[1] Batagelj V, Mrvar A. Pajek Analysis and Visualization of Large Networks [M]. Graph Drawing Software, 2004:77 - 103.

[2] 俞宇楠. 研究主题的动态演化及知识流动特性分析 [D]. 哈尔滨:哈尔滨工业大学,2011.

[3] 贺　亮. 基于话题模型的科技文献话题发现与趋势分析 [D]. 上海:上海交通大学,2012.

[4] 马秀敏. 中国典型管理期刊文献主题发现与演化分析 [D]. 大连:大连理工大学,2011.

[5] Griffiths T L, Steyvers M. Finding Scientific Topics [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(1):5228 - 5235.

[6] 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.

[7] 张红卫. 基于科技文献的时序主题链构建方法研究 [D]. 大连:大连理工大学,2013.

[8] 杨　星,李保利,金明举. 基于 LDA 模型的研究领域热点及趋势分析 [J]. 计算机技术与发展,2012,22(10):67 - 69.

[9] 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.

[10] 张才东,清　强. 基于 LDA 的主题演化文本挖掘模型 [C] //全国知识组织与知识链接学术交流会议,2011:208 - 211.

[11] 王　萍. 面向教育技术学文献数据的主题挖掘 [J]. 现代教育技术,2009,19(5): 46 - 50.

[12] Sugimoto C R, Li D, Russell T G, et al. The Shifting Sands of Disciplinary Development: Analyzing North American Library and Information Science dissertations using latent Dirichlet allocation [J]. Journal of the American Society for Information Science and Technology,2011,62(1): 85 - 204.

[13] 黄泽明. 基于主题模型的学术论文推荐系统研究 [D]. 大连:大连海事大学,2013.

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

[15] David M Blei, Andrew Y Ng, Michael I Jordan. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research,2003(3): 993 - 1022.

[16] Steyvers M, Griffiths T. Probabilistic Topic Models [J]. Handbook of Latent Semantic Analysis, 2007,427(7):424 - 440.

[17] Griffiths T L, Steyvers M, Tenenbaum J B. Topics in Semantic Representation [J]. Psychological review, 2007,114(2):211 - 244.

[18] Cao J, Xia T, Li J, et al. A Density - based Method for Adaptive LDA Model Selection [J]. Neurocomputing, 2009,72(7): 1775 - 1781.

[19] 崔　凯. 基于 LDA 的主题演化研究与实现 [D]. 长沙:国防科学技术大学,2010.

[20] 张　蕊,邱均平,周黎明. 计算机辅助内容分析软件进展研究 [J]. 图书情报工作,2005,49(6):29 - 32.

[21] 安　璐. 基于自组织映射的期刊主题研究 [D]. 武汉:武汉大学,2009.

(责编:贺小利)

(上接第 164 页)

Text - Based Methods Towards Automatic Web Thesaurus Construction [A] //Proceedings of the American Society for Information Science and Technology [C]. American: Wiley online library,2004,41(1):463 - 471.

[5] 周全明. 全文检索系统后控制技术研究 [D]. 北京: 空间政治学院,1995.

[6] Pierre Senellart, Vincent D. Blondel. Automatic Discovery of Similar Words, Chapter in: Survey of Text Mining [J]. Springer - Verlag,2003.

[7] Lesk M. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an ice Cream Cone [C]. New York: Proceedings of the 5th Annual Conference on Systems Documentation,1986:24 - 26.

[8] 李　赟,黄开妍,任福继,等. 维基百科的中文语义相关词获取及相关度分析计算 [J]. 北京邮电大学学报, 2009,32(3): 109 - 112.

[9] Masaki Murata, Toshiyuki Kanamaru, Hitoshi Isahara. Automatic Synonym Acquisition Based on Matching of Definition Sentences in Multiple Dictionaries [J]. CICLing 2005, LNCS 3406, 2005:293 - 304.

[10] Wu Hua, Zhou Ming. Optimizing Synonym Extraction Using Monolingual and Bilingual Resource [A]. Stroudsburg: Proceedings of the Second International Workshop on Paraphrasing, 2003:72 - 79.

[11] 崔　航, 文继荣, 李敏强. 基于用户日志的查询扩展统计模型 [J]. 软件学报, 2003,14 (9) :1593 - 1599.

[12] 张选平, 马　琮, 蒋　宇, 等. 一种基于概念抽取的相关词推荐模型 [J]. 微电子学与计算机,2003,23(5):163 - 165,169.

[13] 贺德方, 乔晓东, 朱礼军, 等. 汉语科技词系统(新能源汽车卷) [M]. 北京:科学技术文献出版社,2012.

[14] Jarneving B. A Comparison of Two Bibliometric Methods for Mapping of the Research Front [J]. Scientometrics,2005,65(2): 245 - 263.

[15] Sternitzke C, Bergmann I. Similarity Measures for Document Mapping: A Comparative Study on the Level of an Individual Scientist [J]. Scientometrics,2009,78(1):113 - 130.

(责编:刘影梅)