doi: 10.3969/j.issn.0490-6756.2018.06.012

基于因果岭回归的多数据源科研主题识别方法

何增颖[1], 陈建锐[1], 钟足峰[2]
(1. 岭南师范学院网络与信息技术中心, 湛江 524048; 2. 岭南师范学院商学院, 湛江 524048)

摘 要: 为了有效解决多数据源科研主题的识别问题,基于因果岭回归建立了一种新的多数据源科研主题识别方法. 该方法首先给出了多数据源科研主题识别关键参数(如主题词的引用权重、状态密度)的评价指标;同时根据科研主题形态特征建立了特征函数,并基于因果岭回归给出了具体识别方法;最后,通过仿真实验深入研究了影响该识别方法的关键因素. 结果显示,与朴素贝叶斯、KNN 算法和 MGe-LDA 算法相比较,该方法在价值引用量、引用权重和前沿主题相似度等方面具有较大优势.

关键词: 多数据源; 科研主题; 识别方法; 形态特征; 因果岭回归
中图分类号: TP391 **文献标识码**: A **文章编号**: 0490-6756(2018)06-1204-07

The research topics identification with multiple data source based on causal regression

HE Zeng-Ying[1], *CHEN Jian-Rui*[1], *ZHONG Zu-Feng*[2]
(1. Network and Information Technology Center, Lingnan Normal University, Zhanjiang 524048, China;
2. Business School, Lingnan Normal University, Zhanjiang 524048, China)

Abstract: In order to effectively tackle the research topics identification with multiple data source, a new research topic identification method is presented based on causal regression. In this paper, the evaluation indicators are defined to identify the key parameters of research topics for multiple data source, such as the citation weight and status density of research topics, the feature function is established with morphological characteristics of research topics, and the research topics identification based on multiple data sources is modeled by causal regression. The experimental results show that the proposed method has great advantages in terms of value citation, citation weight and similarity with frontier topics, compared with Naive Bayes, KNN and Mge-LDA algorithm.

Keywords: Multiple data source; Research topics; Identification method; Morphological characteristics; Causal regression

1 引 言

随着网络信息技术的迅速发展,互联网已经成为人们获取和发布信息的最重要平台之一,如何在海量数据面前提供有效的识别方法,已经成为人们识别科研主题数据的主要手段. 目前,典型的科研主题数据识别方法有分层抽样法和焦点识别法等[1-5]. 文献[6]基于文本分割方法建立了一种主题分析模型,利用 LDA 为语料库采取背景词汇聚类方式将主题词扩充到待分析文本之外,以此挖掘隐藏于字词表面之下的文本内涵. 文献[7]结合三维马尔可夫模型介绍文本网数据的 4 个主要特点以

收稿日期: 2017-11-24
基金项目: 广东省科技厅公益研究与能力建设专项资金项目(2015A020219013)
作者简介: 何增颖(1981-), 女, 硕士, 实验师, 研究方向为人工智能. E-mail: hezy@lingnan.edu.cn

及对应的主题模型，并利用二型模糊系统阐述分布式单词计算方法和主题建模应用方法. 文献[8]针对微博主题情感分析过程中没有协同分析微博主题与微博情感的问题，基于多特征融合建立了一种微博主题情感挖掘方法，该方法将情感表情符号与用户性格情绪特征纳入图模型中，以此实现微博主题与情感的同步推导，使其具有更优的主题情感检测性能. 文献[9]基于概率话题模型提出了面向动态主题数的新闻焦点演化分析方法，该方法能够使得焦点数随着事件的发展而动态变化，但是不能有效刻画时间长度小于分割长度的焦点. 文献[10]针对 Web 社区识别方法忽略文本属性的问题，基于文本相似度的边容量分配建立了一种网页内容分析与链接分析的改进算法，并通过社区结点排序策略刻画节点和社区主题的相似度，以此来加强节点区分度. 文献[11]基于短语参数学习建立了主题模型，以此抽取在线评论中被评价实体的 aspect 和与之对应的 rating，并在实施过程中引入先验知识用来短语识别分析和等级预测精度分析. 文献[12]针对新闻焦点识别方法存在的演化跟踪偏斜和焦点识别不清等问题，基于焦点和时间联合建立了一种新闻焦点演化跟踪方法，该方法有效克服了跟踪算法依赖时间分割的局限性，能够识别出各焦点持续时间分布及报道力度.

但是，目前的识别往往精度不够高，数据源较混乱，主题相对不够集中. 因此，本文拟将基于形态特征与因果岭回归[13-17]，利用分层抽样和焦点识别来建立一种新的多数据源科研主题识别方法，同时通过仿真实验深入研究了影响该识别方法的关键因素.

2　性能刻画

当前，科研前沿进度隐含在一些科技文本数据、基金项目数据、科技论文中，揭示不同领域科研主题的发展，主要以学术文献的形式呈现，由于每种数据源对科学研究的贡献程度不同，针对多数据源科研主题材料，本文侧重对有效主题材料的识别. 本文研究多数据源科研主题获取来源是专门进行学术研究成果报道的学术性期刊，比如知网、IEEE、ACM、Elsevier、Springer 等多种数据源，这些基本反映当前世界各领域的科研水平. 对此，本文针对这些学术性期刊，基于主题特征值变化在动态时间序列下对多数据源科研主题进行有效识别：捕抓时域和频域下的形态特征，并保留充分的数据信息. 假设随机选取科研主题的两个特征 $Theme_1$ 和 $Theme_2$，计算主题余弦相似度参考值 TS，如下式.

$$TS = 1/\cos(Theme_1, Theme_2) \tag{1}$$

考虑到科学研究主要参考近期发表的文献集，即前沿成果推动科学研究进一步发展的情况，本文引用前沿主题相似度 α 来进行比较，如下式.

$$\alpha = \frac{Sum_t Ca(FT)}{Sum_t R(T)} \cdot (1/\cos(Theme_1, Theme_2)) \tag{2}$$

若引用的前沿主题相似度值偏高，则反映科研数据的语义聚类映射性好，该主题已经收到广泛关注，并伴随有大量研究发表，在相同时间系数和数据源下，本文根据某个目标科研主题的被引用量进行讨论：假设 $Sum_t Ca(FT)$ 表示前沿主题 FT 在时间 t 内被引用的数量之和，$Sum_t R(T)$ 代表相似科研主题 T 在 t 时间段的发文统计量.

首先对被引用的核心参考文献集进行研究，以便识别多数据源下的潜在价值，令 w_i 表示 $Topic_i$ 主题下的引用参数，分析识别研究主题的演化过程和研究热点、趋势，比较 i、j 主题的引用权重 χ，如下式.

$$\chi = \frac{\sum_{i=1}^{n} w_i(Topic_i) \times \sum_{j}^{n} w_j(Topic_j)}{\sqrt{(\sum_{i=1}^{n} w_i^2(Topic_i)) \times (\sum_{j=1}^{n} w_j^2(Topic_j))}} \tag{3}$$

科研主题隐含在不同的文本数据中，随着主题引用量的增大，主题热度不断提高，反复被引用的参考文献多次被使用之后，影响力和新颖度会下降，若重复引用率过高，甚至会使得该研究主题失去价值，因此，对于多次被引用的科研主题文献，识别该文献的可用性是十分重要的. 假设当前目标文献共有 $total\_num$ 可引用量，通过拟合计算年有效量，如下式.

$$Tpy = \sum_{i=1}^{n} year_{t_i}/total\_num \tag{4}$$

其中，$year_{t_i}$ 表示主题特征 t_i 的年出现量. 为了研究并区分不同年限时隙对科研主题的影响，对数据进行时间窗口划分，即不同年限的就近引用率不同，利用引用概率分布计算，得到两个不同年限主题状态的密度分别为 p 和 p'，如下式.

$$p = \int_i^n \frac{1}{total\_num} \sum_{i=1}^{n} f(t_i \to t_j)/Tpy(i \neq j) \tag{5}$$

$$p' = \int_i^E \frac{1}{total\_num}\sum_{i=1}^{n} f(t_i \to t_j \to t_E)/Tpy \quad (i \neq j \neq E) \tag{6}$$

其中,$t_i \to t_j$ 表示主题特征 t_i 和其他主题特征 t_j 之间相同特征的共词数,$f(t_i \to t_j)$ 表示其共词数 n 出现的频率,同理比较两个以上主题词的密度 p',其中 $E=(1,2,\ldots,\mathrm{num})$. 伴随着被引用的年总比例 Tpy 增加,为进一步刻画科研数据可用价值,本文根据共词数计算该科研主题材料可实现的有价值引用量为 β,如下式.

$$\beta = \sqrt{\sum_{i=1}^{n} (t_i \to t_j)^2 \sin(\pi \mathrm{p})} \tag{7}$$

假设目的主题时间序列长度为 t_s,该主题文献平均已实现引用量的形态是 q,利用斜率近似值计算第 i 个时序段的形态特征拟合引用权重 $\bar{q}_i$,如下式.

$$\bar{q}_i = \frac{ts * \sum_{j=j_0}^{ts*i} j * q_j - (\sum_{j=j_0}^{ts*i} j)(\sum_{j=j_0}^{ts*i} q_j)}{ts * \sum_{j=j_0}^{ts*i} j^2 - (\sum_{j=j_0}^{ts*i} j)^2} \tag{8}$$

其中,j 表示引用基点值;$j_0 = t_s(i-1)+1, i=1,2,\cdots,t_s$.

对此,本文针对上述提出的主题相似度、引用权重、引用量等参量,利用具有因果关系的岭回归来刻画和识别多数据源科研主题,以此提高科研主题识别精度.

3 识别方法

岭回归是一种改良最小二乘估计法,大多应用于处理共线性数据分析,对具有因果关系的不定数据也能进行正则化处理,且能统一诊断并处理多重共线性问题,是一种针对多数据源信息实用性强大的方法. 由于多数据源数据是一种高维空间的数据集合,考虑到多数据源具有较强的可变性,本文在岭回归的基础上,结合因果关系对多数据源进行预测识别. 首先对多数据源进行降维:假设两个共线性数据 x、y,且 g 作为耦合向量函数,f 是非线性向量函数,定义数据误差关系是 $u=x-y$. 计算得到整函数 $\dot{u}$,如下式.

$$\dot{u} = (f(x) - f(y) - g(x) - g(y)) \tag{9}$$

$$Df(x) - Dg(x) = \frac{(f(x) - f(y) - g(x) - g(y))}{x - y} \tag{10}$$

其中,D 分别是非线性向量函数和耦合向量函数沿着轨迹分布的偏微分计算,根据 f、g 微分计算得到降维后的特征函数 $Df(x) - Dg(x)$,此时特征值可以用 $M(u_1,u_2)$ 表示,如下式.

$$u_1 = (1 - \frac{\dot{u}}{2}) + \frac{\sqrt{4\dot{u} - \dot{u}^2}}{2} \tag{11}$$

$$u_2 = (1 - \frac{\dot{u}}{2}) - \frac{\sqrt{4\dot{u} - \dot{u}^2}}{2} \tag{12}$$

当特征函数的特征值为复数,且 $\dot{u} \in (1,4)$,则 $\cos\theta = 1 - \frac{\dot{u}}{2}$,$\sin\theta = \frac{\sqrt{4\dot{u} - \dot{u}^2}}{2}$,此时特征值可以重写. 重写过程如下:首先确定动态方程,在某一段确定时间序列下,把各数据源下的主题看作一个因子,多维空间下分析引用过程中的调用为

$$x_f^{m+1} = (1+w)x_f^m - wx_f^{m-1} \tag{13}$$

利用改良最小二乘法对处理时序中所包含的数据,进行迭代计算,得到岭回归虚拟值 X,如下式.

$$X = x_f^m + wv_f^m\left[w^{n-1} + (1 - \varphi_1^{m+n} - \varphi_1^{m+n})\sum_{j=1}^{n-2} w^j\right] + \prod_{i=2}^{n}(1 - \varphi_1^{m+j} - \varphi_2^{m+j}) \tag{14}$$

其中,m 和 n 是信息量;w 是惯量因子;x_f^m 是位置向量;φ 是约束因子.

对单位矩阵 I,列满秩时,因子 y 表达为 $X\delta = y$,根据梯度下降法求解

$$\delta = (X^TX)^{-1}X^Ty \tag{15}$$

当列矩阵不是满秩时,给损失函数增加一个正则化项.

$$s = \| X\delta - y \|^2 + \| \Gamma\delta \|^2, \Gamma = \alpha I \tag{16}$$

$$\delta(\alpha) = (X^TX + \alpha I)^{-1}X^Ts \tag{17}$$

其中,$\delta(\alpha)$ 随 X 的改变而变化.

基于上述最小二乘回归方法,此处利用具有因果关系的岭回归来建立多数据源科研主题识别方法,使其具有更高的数值稳定性和计算精度. 本文给出如下算法步骤.

step1 初始化各参数,确定数据来源,从常见期刊源下载并确定科研主题,依据较近年限作为参考值,设置统一的环境变量.

step2 降维定义 $\dot{u}$,得到特征函数 $Df(x) - Dg(x)$,通过特征函数计算出两个显著 $Theme_1$ 和 $Theme_2$,并刻画前沿主题相似度 α,结合 α 确定特征值 M.

$$M = \begin{cases} (1 - \frac{\dot{u}}{2}) + \alpha\frac{\sqrt{4\dot{u} - \dot{u}^2}}{2} \\ (1 - \frac{\dot{u}}{2}) - \alpha\frac{\sqrt{4\dot{u} - \dot{u}^2}}{2} \end{cases} \tag{18}$$

step3 开始调用计算,对引用权重 $\bar{q}_i$,比较价

值引用权重 $\bar{q}_i'$，如下式.

$$\bar{q}_i' = \beta \cdot \frac{ts * \sum_{j=j_0}^{ts*i} j * q_j - (\sum_{j=j_0}^{ts*i} j)(\sum_{j=j_0}^{ts*i} q_j)}{ts * \sum_{j=j_0}^{ts*i} j^2 - (\sum_{j=j_0}^{ts*i} j)^2} \tag{19}$$

step4　根据总引用量 *total_num* 计算被引用的年总比例 Tpy，由于就近引用原则影响，考虑不同年限概率密度分布 p、p'，用有价值引用量 β 优化回归方程.

$$X' = \sqrt{\pi\beta(x_f^m + \omega v_f^m[\omega^{n-1} + (1-\varphi_1^{m+n} - \varphi_1^{m+n})\sum_{j=1}^{n-2}\omega^j] + \prod_{i=2}^{n}(1-\varphi_1^{m+j} - \varphi_2^{m+j}))} \tag{20}$$

step5　根据回归方式计算输出值 $\delta'(\alpha)$ 得到岭迹，如下式.

$$\delta'(\alpha) = (X^T X' + \alpha I)^{-1} X^T s' \tag{21}$$

step6　输出各指标值.

step7　仿真结束

4　数学仿真

为了验证基于形态特征与因果岭回归识别算法的有效性，识别被引用数据，本文从相关科研学报的官网上每种学报随机下载 100 份文献样本，设置确定的引用值，在 MATLAB 平台进行仿真实验. 设置不同科研主题引用量，不同发表时间的年限，并考虑数据源影响，分析不同数据源学报，引用说服力强的学报. 同时将本文算法和常见数据挖掘算法朴素贝叶斯、KNN 算法、MGe-LDA(Mutually Generative Latent Dirichlet Allocation)算法[18] 在被引用量参数上进行比较，实验数据统计结果如表 1 所示，三种算法关于被引用量参数都存在一定程度的误判或漏判，但是本文提出的算法总体情况优于其他两种算法.

表 1　实验结果比较

Tab. 1　Comparison of experimental results

数据源	样本数量(份)	被引用量(次)	本文算法(次)	朴素贝叶斯(次)	KNN 算法(次)	MGe-LDA 算法(次)
计算机学报	100	97	96	90	79	75
软件学报	100	98	97	89	78	75
IEEE	100	96	95	87	75	74
电子信息学报	100	99	98	89	77	76
Physica A	100	97	96	88	79	78
自动化学报	100	96	94	85	80	80
Nature	100	95	94	84	76	79
Safety Science	100	98	97	83	78	78
Physical Review E	100	96	95	85	73	79
Transportation Research Part B	100	95	94	88	79	75
Applied Mathematical Modelling	100	98	98	86	76	73
Building and Environment	100	95	94	89	73	74
计算机研究与发展	100	99	98	87	74	76
控制与决策	100	98	97	89	78	78
系统仿真学报	100	97	97	88	79	75
控制理论与应用	100	98	98	89	74	77
模式识别与人工智能	100	96	95	87	73	75
机器人	100	98	97	84	77	75
科学通报	100	97	96	87	72	78
平均检测率(%)			99.07	89.74	78.73	74.44

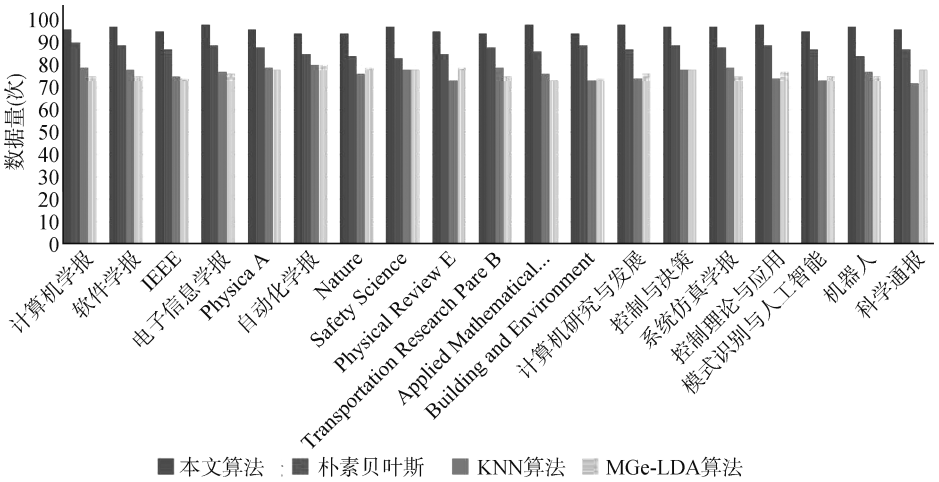

图 1 被引用数据量比较结果

Fig. 1 Comparison of the cited data

这里定义检测率=被引用量正确检测数量/样本数量. 如图 1 所示,KNN 算法、MGe-LDA 算法识别效果不佳,检测率未达到 80%,且对于不同数据源计算能力差别不大;朴素贝叶斯在数据处理方面有较好的反应;本文比较检测率在 80%~90%左右. 本文算法与朴素贝叶斯、KNN 算法、MGe-LDA 算法相比,本文算法的计算结果最接近实际值,对于不同数据源的检测率均在 95%以上.

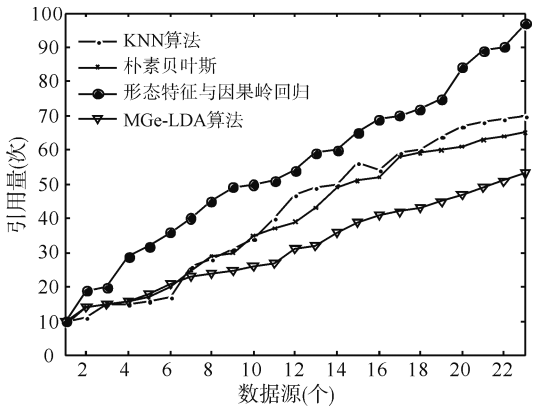

图 2 多数据源下引用量比较

Fig. 2 Comparison of citation in multiple data sources

为了对比数据源数量对算法识别能力的影响,本文模拟 23 个数据源环境下,使用四种识别算法,计算出科研主题文献引用量,比较结果如图 2 所示. 当只有较少数据源时,四种算法得到的引用量差别不大,随着数据源增多,KNN 算法逐渐超过朴素贝叶斯,但优化性能处理力度不大;KNN 算法和朴素贝叶斯算法随着数据源增多逐渐达到极值;MGe-LDA 算法识别能力低于三种算法;本文算法性能随着数据源的增多,能够较好达到识别的效果,并且随着数据源的增多,识别性能依然较好.

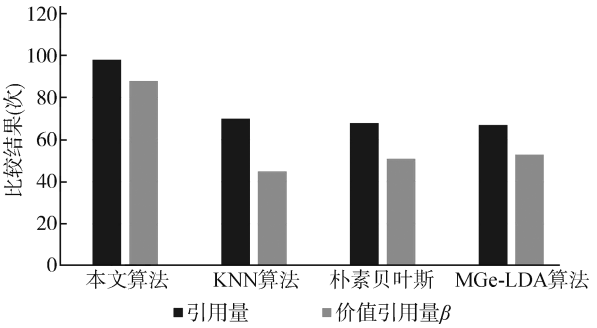

图 3 价值引用量比较

Fig. 3 Comparison of value quotation

同时这里基于引用量变化,比较三种算法的价值引用量 β, 如图 3 所示. KNN 算法和朴素贝叶斯算法的引用量与 β 计算结果均不大,且两者之间的跨度大,反应出数据丢失情况严重;MGe-LDA 算法跨度对比相对较小,数据完整性较好,但该算法对数据引用处理性能不佳. 实际效果下,本文算法引用量为 98%左右,基于该引用量下得到的 β 值达 88%,可见本文算法对数据的适应性能佳.

考虑多种数据源环境下,类似主题识别影响,利用引用权重刻画算法性能,如图 4 所示. 当数据源在 1~3 范围内,本文算法计算得到的引用权重值低于 KNN 算法、MGe-LDA 算法、朴素贝叶斯,当数据源大于等于 4 之后,本文算法的优势体现出来,随着数据源的增多,本文算法计算得到的引用权值也逐渐增大,接近正比例变化;当数据源大于 16 时,KNN 算法、MGe-LDA 算法、朴素贝叶斯的增加速度已经趋于平稳,可见,对于多数据源的数据,本文算法的自使用性较好.

科研主题需要严谨,考虑到反面特征定位影响科研主题识别,使用前沿主题相似性来比较不同算

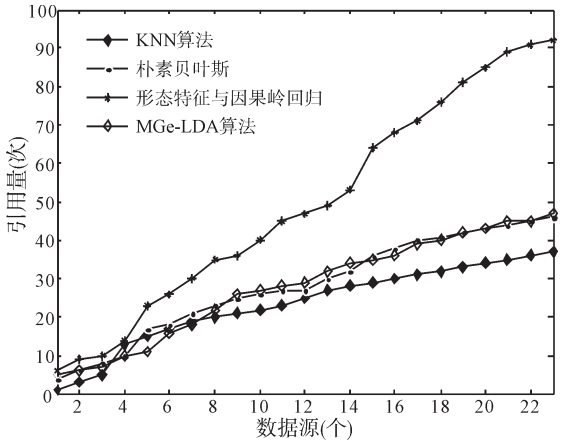

图 4　引用权重比较
Fig. 4　Comparison of reference weight

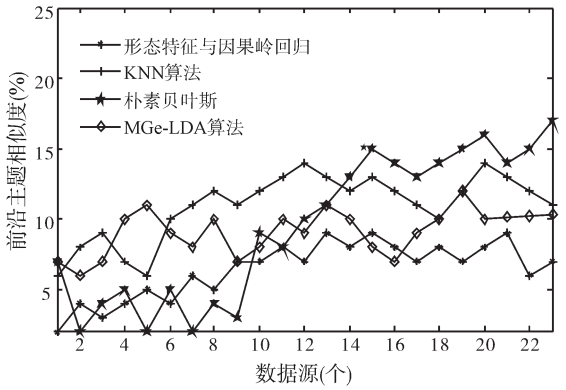

图 5　前沿主题相似度比较
Fig. 5　Comparison of similarity in frontier topics

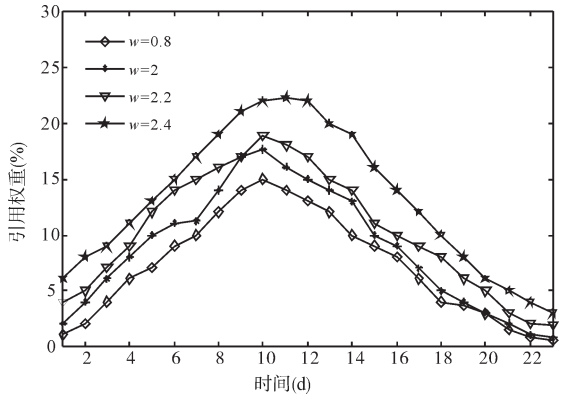

图 6　不同引用参数 w 下的引用权重比较示意图
Fig. 6　Comparison of reference weight in different index

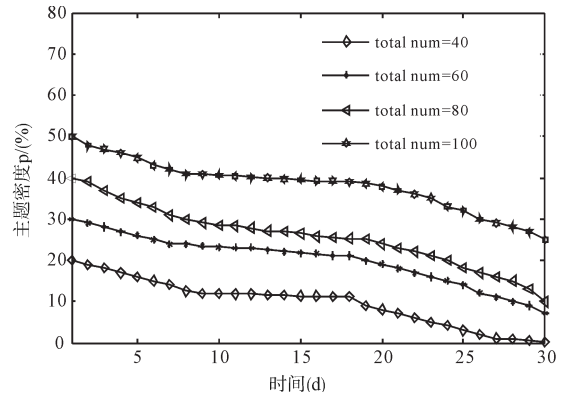

图 7　不同可引用量 total_num 下主题密度比较
Fig. 7　Comparison of subject density in different number of cited total_num

法，计算过程受数据源数量影响，如图 5 所示. MGe-LDA 算法、KNN 算法在数据源少的情况下，前沿主题相似度的刻画较好，而朴素贝叶斯和本文算法在少数据源情况下，对于主题相似度区分效果不佳. 但在多数据源情况下，本文算法和朴素贝叶斯呈现出较好性能，能够有效提高主题相似度. 并且在数据源数量达到 14 之后，其性能优于朴素贝叶斯. 说明本文算法能够有效处理多数据源主题相似度.

根据上述引用权重的计算公式，本文在 23d 内针对某一特定 *Topic* 主题，基于不同引用参数的科研数据计算引用权重百分率，计算结果如图 6，引用权重随着时间的增加先增后减，在第 10d 至 12d 左右，引用权重达最大值. 整个过程描述为：当新颖科研项目出现，引用权重伴随研究价值上升. 若该主题成为热议，被多次引用，新颖度随着引用次数降低，此时引用权重会逐渐下降.

主题密度受可引用总量值的影响，主题密度与价值引用量相关. 本文比较可引用量分别为 40、60、80、100 下的主题密度值 p，图 7 所示，根据曲线之间的跨度可以看出 p 与 num_total 呈现正比例变化，total_num 在 60～80 之间时，变化稍减，价值引用量较大或较小时，变化明显. 根据每条曲线的变化趋势，可以看出随着时间的增加，p 越来越小，在 7～14 范围内，加速度有所减小，可见此过程科研主题数据被研究率高.

最后，本文比较价值引用量 β^2 和主题密度值 p 的之间关系，结果如图 8 所示. 对价值引用量进行处理得到 $\frac{\arcsin\beta^2}{\pi}$，与 p 是正弦关系. 由此可见，密度和被引用量是振荡关系，而不是受数据直接的单调递增或者递减影响.

5　结　论

针对多数据源科研主题的识别问题，本文利用形态特征和因果岭回归建立了一种新的多数据源科研主题识别方法. 该方法首先给出了多数据源科研主题识别关键参数（如主题词的引用权重、状态

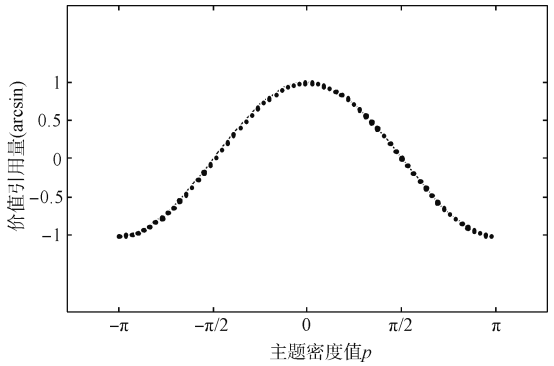

图 8 价值引用量 β^2 与主题密度值 p 的变化关系

Fig. 8 The relationship between of value quotation β^2 与 and subject density p

密度)的评价指标,同时根据科研主题形态特征建立了特征函数,并基于因果岭回归给出了具体识别方法.最后,通过仿真实验深入研究了影响该识别方法的关键因素.结果显示,与朴素贝叶斯、KNN 算法和 MGe-LDA 算法相比较,该方法在价值引用量、引用权重和前沿主题相似度等方面具有较大优势.在后续研究中,可以考虑用户偏好、主题类型等参数来完善科研主题识别方法.

参考文献:

[1] Tang H, Tan S, Cheng X. A survey on sentiment detection of reviews[J]. Expert Syst Appl, 2009, 36: 10760.

[2] Rao Y, Li Q, Mao X, *et al*. Sentiment topic models for social emotion mining[J]. Inform Sciences, 2014, 266: 90.

[3] 胡斐, 罗立民, 刘佳, 等. 基于时空兴趣点和主题模型的动作识别[J]. 东南大学学报: 自然科学版, 2011, 41: 962.

[4] 周亚东, 刘晓明, 杜有田, 等. 一种网络话题的内容焦点迁移识别方法[J]. 计算机学报, 2015, 38: 261.

[5] 朱靖波, 姚天顺. 文本内容主题的识别方法[J]. 东北大学学报:自然科学版, 2002, 23: 425.

[6] 石晶, 范猛, 李万龙. 基于 LDA 模型的主题分析[J]. 自动化学报, 2009, 35: 1586.

[7] 曾嘉, 严建峰, 龚声蓉. 复杂文本网数据的主题建模进展[J]. 计算机学报, 2012, 35: 2431.

[8] 黄发良, 冯时, 王大玲, 等. 基于多特征融合的微博主题情感挖掘[J]. 计算机学报, 2017, 40: 872.

[9] Lu R, Xiang L, Liu M R, *et al*. Discovering news topics from microblogs based on hidden topics analysis and text clustering[J]. PR & AI, 2012, 25: 382.

[10] 张宪超, 徐雯, 高亮, 等. 一种结合文本和链接分析的局部 Web 社区识别技术[J]. 计算机研究与发展, 2012, 49: 2352.

[11] 吕品, 汪鑫, 罗宜元, 等. 基于主题模型的(Aspect, Rating)摘要生成方法研究[J]. 电子学报, 2016, 44: 3036.

[12] 刘玉文, 吴宣够, 郭强. 网络热点新闻焦点识别与演化跟踪[J]. 小型微型计算机系统, 2017, 38: 738.

[13] 邓赵红, 张江滨, 蒋亦樟, 等. 基于模糊子空间聚类的 0 阶岭回归 TSK 模糊系统[J]. 控制与决策, 2016, 31: 882.

[14] 黄宴委. 基于核岭回归的非线性内模控制[J]. 控制与决策, 2009, 24: 1100.

[15] 李海林, 郭崇慧. 基于多维形态特征表示的时间序列相似性度量[J]. 系统工程理论与实践, 2013, 33: 1024.

[16] 李海林, 梁叶. 基于数值符号和形态特征的时间序列相似性度量方法[J]. 控制与决策, 2017, 32: 451.

[17] 赵超, 唐亚勇. 分位点门限自回归时间序列模型的贝叶斯方法[J]. 四川大学学报: 自然科学版, 2016, 53: 748.

[18] Xing C, Wang Y, Liu J, *et al*. Hash tag-based subevent discovery using mutually generative LDA in Twitter[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, Arizona: AAAI Press, 2016.

引用本文格式:

中 文: 何增颖, 陈建锐, 钟足峰. 基于因果岭回归的多数据源科研主题识别方法[J]. 四川大学学报: 自然科学版, 2018, 55: 1204.

英 文: He Z Y, Chen J R, Zhong Z F. The research topics identification with multiple data sources based on causal regression [J]. J Sichuan Univ: Nat Sci Ed, 2018, 55: 1204.