基于社会网络关注度的学科前沿热点挖掘

张　晖[1]，　杨小彦[2]，　赵旭剑[2]，　杨春明[2]，　李　波[2]

(1. 西南科技大学 理学院　四川 绵阳 621010；2. 西南科技大学 计算机科学与技术学院　四川 绵阳 621010)

摘要：从科研文献数据中挖掘出学科前沿热点是目前学术界和工业界亟待解决的问题. 社会网络可以及时反映信息传播的实际受欢迎程度,故提出一种基于社会网络关注度的学科前沿热点挖掘方法. 首先通过数据相关性分析、相关属性划分以及社会网络关注度因子挖掘,构建文献热度评价模型. 同时,采用文档主题生成模型(latent dirichlet allocation,LDA)从文献热度评价模型挖掘的科研文献中识别出该学科的前沿热点. 最后,在“artificial intelligence and image processing”学科的数据集上构建评价模型并进行多组对比实验,结果表明提出的方法有效提高了学科热点挖掘结果的前沿性,热点主题在时间维度上更具时效性.

关键词：文献热度评价模型；社会网络关注度；相关性分析；LDA 模型；因子挖掘

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1671-6841(2018)03-0046-07

DOI：10. 13705/j. issn. 1671-6841. 2017201

0　引言

随着信息资源数量和种类的急速增长,科学研究领域不断开拓,科研人员和学者在掌握学科前沿热点方面面临着越来越多的挑战. 如何快速、准确地从科研文献中提取和识别学科领域研究的前沿热点,对当前科研工作具有重要研究意义[1]. 传统研究方法主要是以电子期刊、学位论文等作为数据源,采用词频分析[2]、共词分析[3]、多维尺度分析[4]、社会网络分析[5]和其他分析模型[6-8]来识别前沿热点. 这类研究方法主要通过分析文献的学术传播热度来挖掘学科领域热点,仅考虑了领域知识在专业学术平台的影响力,忽视了科研文献在社会网络中的流行与传播,热点挖掘结果存在滞后、前瞻性较差等不足.

学科前沿热点挖掘可分为两类:第一类简单地考虑文献计量特征,包括词频分析、文献引用、关键词的共词或共现分析;第二类是使用广泛用于文本挖掘中的主题模型 LDA、HDP 等来识别主题热点. 由于引文和关键词能较好地描述科技文献的主题内容,因此国内外学者利用文献计量的引文分析法、可视化图谱[9]、关键词的词频分析和共词分析等方法进行学科前沿热点挖掘. 文献[3]运用文献计量中共词和文档共引,从高影响力文章、作者、期刊、机构和国家等角度出发,绘制知识图谱,分析抗癌研究领域的研究热点和整体发展趋势. 文献[10]基于 h 指数和引文分析法梳理了国内外碳市场研究领域的研究热点、发展趋势和主要区域分布等. 文献[11]利用 WoSCC 收录的 Treg 领域相关文献数据进行文献计量方法和 Citespace[12]绘制共引图谱,分析该领域的研究热点和发展演化趋势. 除此之外,基于主题模型的学科领域热点识别方法也得到了广泛运用,如文献[13]使用 LDA 模型从论坛语料中识别热点话题,并计算话题强度和特征关键词,以找到热点话题的发展及演化趋势,实验结果说明该方法是合理和有效的. 文献[14]提出基于主题模型的热点发现技术,实验表明该模型在文本挖掘方面的热点主题识别上具有明显的优势.

相较于专业学术平台,信息在社会网络环境下往往传播速度更快、范围更广,更能实时地体现传播对象的冷热程度及普遍性,充分保证学科热点的前沿性. 因此,科研文献在社会网络中的传播影响力分析对挖掘学科前沿热点具有重要研究意义和应用价值. 基于该思想,本文考虑了社会网络中文献传播的普及,提出了一种用于热点主题挖掘的方法. 将文献热度属性分为传统和社会属性,然后在社会网络环境中构建文献热度

收稿日期:2017-07-05

基金项目:四川省军民融合研究院开放基金项目(18sxb017);西南科技大学博士基金项目(12zx7116);四川省信息管理与服务研究中心科研基金项目(SCTQ2016YB13);四川省赛尔网络下一代互联网技术创新项目(NGII20170901).

作者简介:张晖(1972—),男,四川绵阳人,教授,主要从事大数据与人工智能研究,E-mail:429219831@qq.com;通信作者:杨小彦(1991—),女,四川广安人,硕士研究生,主要从事大数据与自然语言处理研究,E-mail:18281607011@163.com.

评价模型,计算和分析文献关注度,挖掘具有社会传播影响力的学术论文. 其次,采用 LDA 算法对文献内容进行主题挖掘,生成学科前沿热点主题. 与已有的工作相比,本文的主要贡献在于:

1) 从数据相关性的角度分析文献传播的评价指标与文献热度的关联性,采用无监督学习方法进行各媒体指标的主成分分析,划分影响文献热度的热度评价指标主题类别,为测度文献的社会网络关注度指标奠定了基础.

2) 通过挖掘评价指标中的社会网络关注度因子,构建文献热度评价模型,计算文献社会传播热度(社会网络关注度),采用 LDA 主题模型对文献内容进行主题挖掘,生成学科前沿热点主题.

3) 利用 Altmetric[8] 获取"artificial intelligence and image processing"学科的 16 658 条论文记录数据集,对提出的学科前沿热点挖掘方法进行了对比实验. 实验结果表明,本文提出的方法有效提高了学科热点挖掘结果的时效性,热点主题在时间维度上与传统方法相比,挖掘结果更具有学科前沿性.

1　基于社会网络关注度的学科前沿热点挖掘

1.1　研究方法

本文工作主要为两部分:对指标数据进行相关性分析,采用无监督学习方法进行评价指标的主成分聚类,剖析出影响文献热度的媒体指标主题类别,挖掘社会关注度因子并构建文献热度评价模型;采用吉布斯抽样的 LDA 模型对科研文献内容进行学科前沿热点挖掘,生成学科前沿热点知识. 热点主题挖掘算法具体的方法流程如下所示.

输入: $K=\langle k_1, k_2, \cdots, k_n\rangle$, $A=\langle A_1, A_2, \cdots, A_j\rangle$, a_i 表示第 i 条数据 k_i的属性值, 文档标题 t_i, 文档集 D, 参数 P.

输出: 热点主题 topicList.

```
Begin
1:    if    a_i ← CorrelationAnalysis (k_i, A) & a_i ∈ A  then
2:          PrincipalComponentAnalysis(a_i, a_j);
3:          Document Popularity Model←ComponentScoreMatrix(a_i, a_j);
4:          D ← DocumentPopularity Model (t_i);
5:    else
            RemoveAttributes(a_i);
6:    end if
7:    topicList = GibbsSamplingLDA(D, P);
8:    return topicList;
End
```

1.2　文献热度评价模型构建

1.2.1　*相关性分析*　通过数据分析,本文采用皮尔逊(Pearson)相关性模型挖掘文献传播的媒体指标与文献热度的关联性,计算其相关系数并剔除弱相关或无相关的指标,最终提取出 6 个文献热度评价指标(Reddit、Bloggers、Twitter、Google + 、News、Facebook). 皮尔逊相关系数是用来反映两个变量线性相关程度的统计量. 皮尔逊相关系数用 $P_{X,Y}$表示,计算公式为

$$P_{X,Y}=\frac{X\text{ 和 }Y\text{ 的协方差}}{X\text{ 的标准差}\times Y\text{ 的标准差}}=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{(n-1)\sigma_X\times\sigma_Y}. \tag{1}$$

其中:n 为样本量;X_i 和 Y_i 分别为两个变量 X 和 Y 的观测值;σ_X 为变量 X 的标准差.

1.2.2　*热度评价模型指标主题挖掘*　6 个文献热度评价指标从不同维度表征了一篇论文在各媒体平台的传播影响力,然而通过研究发现每个评价指标及其代表的媒体平台都具有一定的主题性. 因此,本文考虑采用无监督学习方法进行热度评价指标的主成分分析,挖掘指标主题. 进行主成分聚类之前,需进行 KMO-

Bartlett 检验,计算出 KMO 值为 0.690,大于 0.5(KMO 值小于 0.5 不适合进行主成分分析)表明指标间存在共同因子,因此也说明可以进行主成分分析.

虽然这些热度指标能提取出共同因子,这些共同因子能聚类到一起构成几个主成分,还需通过解释总方差进行分析,各指标解释总方差如表1所示. 根据主成分的提取原则(主成分对应的特征值应大于1,主成分积累的总方差尽可能大(50%以上)),只有成分1和成分2的特征值合计大于1,而且这两个成分积累的总方差比较大,达到61.149%,没有影响原始数据的共同度,表明可以将6个指标提取出两个主成分. 同时,如表2所示,对各指标进行了主成分载荷矩阵,Facebook、Google +、Twitter、Reddit、Bloggers 5类指标对成分1贡献较大,而成分2则主要依赖于News. 因此,文献热度评价指标可以划分为两个主题:由 Facebook、Google +、Twitter、Reddit、Bloggers 等社交平台构成的社会网络媒体即社会属性;News 为代表的传统网络媒体即传统属性.

表1 各指标解释总方差

Tab.1 Total variance explained

成分	初始特征值			提取平方和载入		
	合计	方差/%	累积/%	合计	方差/%	累积/%
1	2.513	41.875	41.875	2.513	41.875	41.875
2	1.156	19.274	61.149	1.156	19.274	61.149
3	0.913	15.209	76.358			
4	0.59	8.817	85.175			
5	0.507	8.452	93.627			
6	0.382	6.373	100.000			

表2 主成分载荷矩阵

Tab.2 Principal component load matrix

指标	成分1	成分2
Reddit	0.553	-0.102
Bloggers	0.597	0.572
Twitter	0.716	-0.340
Google +	0.726	-0.312
News	0.464	0.749
Facebook	0.772	-0.211

1.2.3 文献热度评价模型构建 由1.2.2节可知文献热度由具有不同主题属性的6个评价指标综合决定,各指标特征满足 AHP 分析法中的单层次模型,因此采用 AHP 模型[15]思想构建文献热度评价模型:

$$p_{soc} = \lambda_1 \cdot R_i + \lambda_2 \cdot B_i + \lambda_3 \cdot T_i + \lambda_4 \cdot G_i + \lambda_5 \cdot N_i + \lambda_6 \cdot F_i, \tag{2}$$

$$p_{tra} = \gamma_1 \cdot R_i + \gamma_2 \cdot B_i + \gamma_3 \cdot T_i + \gamma_4 \cdot G_i + \gamma_5 \cdot N_i + \gamma_6 \cdot F_i. \tag{3}$$

其中:p_{soc}表示第 i 篇文献在社会网络媒体上的关注度,即文献热度;p_{tra}表示第 i 篇文献在传统网络媒体上的关注度;λ_1 代表 Reddit 指标的权重;R_i 代表 Reddit 对第 i 篇文献的引用数;λ_i 是指第 i 个评价指标在整体评价中的相对重要程度,权重越大则该指标的重要性越高,对文献热度的影响就越大.

由热度评价指标主题类别可知,本文可采用主成分分析提取两个主成分,借鉴文献[16]使用主成分分析各指标数据之间的潜在关系,利用回归法计算出成分得分系数,各指标成分得分实际上是一个相对值,即该样本偏离所有样本均值的程度,正值说明超过平均水平,负值说明低于平均水平,正负值正好将各指标划分到对应的主成分中,能很好地区分各主成分的主题类别,结果如表3.

表3 成分得分系数矩阵

Tab.3 Component score coefficient matrix

指标	成分1	成分2
Bloggers	0.237	0.495
Twitter	0.285	-0.294
Google +	0.289	-0.270
News	0.185	0.648
Facebook	0.307	-0.182

从表3可知,Twitter、Facebook、Google +等指标与主成分1密切相关,系数均在0.2以上,由此可以发现主成分1中贡献比较大的指标主要用于测度学术论文在社会网络媒体中的传播影响,通过在线社交过程中的交互行为传播所产生的影响力,是最具社会网络关注度的因子,也是本文研究的重点. News 对主成分2相关系数较高,故主成分2可以代表用于测度学术论文在新闻等传统网络媒体中传播所产生的影响力. 从成分得分系数矩阵确定指标权重得到文献热度评价模型:

$$p_{soc} = 0.22 \cdot R_i + 0.237 \cdot B_i + 0.285 \cdot T_i + 0.289 \cdot G_i + 0.185 \cdot N_i + 0.307 \cdot F_i, \tag{4}$$

$$p_{tra} = -0.088 \cdot R_i + 0.495 \cdot B_i - 0.294 \cdot T_i - 0.27 \cdot G_i + 0.648 \cdot N_i - 0.182 \cdot F_i. \tag{5}$$

2　实验

2.1　文献热度评价实验

本文利用 Altmetric 跟踪“artificial intelligence and image processing”学科的指标数据进行数据分析与处理,构建文献热度评价模型. 因为社会网络媒体具有时间优势. 故以主成分 1 构建的具有社会网络关注度的文献热度评价模型进行学科前沿热点数据挖掘实验. 通过该模型挖掘出社会网络媒体上比较受用户关注以及具有热度的前沿热点文献数据,列举了社会网络关注度排名前 10 的文献数据,结果如表 4 所示.

表 4　社会网络关注度排名前 10 文献数据

Tab. 4　Top 10 papers ranked by the social-network-based evaluating model

p_{soc}值	文献标题	Reddit	Bloggers	News	Twitter	Google +	Facebook
780. 243	Peer-review practices of psychological…	0	12	3	2 586	39	93
218. 444	Judgments About Fact and Fiction…	8	8	13	476	19	232
123. 100	The prevalence of statistical reporting…	0	2	0	425	2	3
74. 505	Shut up and pet me! Domestic dogs…	0	5	6	191	2	56
63. 618	Embodied Cognition is Not What …	2	10	0	185	12	15
60. 808	Differences in negativity bias …	2	3	2	207	1	0
51. 040	Scent of the familiar: An fMRI study …	4	0	0	176	0	0
44. 610	Algorithms for Solving Rubik's Cubes	1	1	9	148	0	1
43. 590	The weirdest people in the world?	2	6	5	137	5	1
41. 284	Disentangling canid howls across…	1	9	21	91	6	24

从表 4 可知,在排名靠前的这 10 篇文献中,Twitter、Facebook 和 Google + 等社会网络媒体上的文献引用数较大,其社会网络关注度的 p_{soc}值较大,说明通过主成分 1 构建的文献热度评价模型能很好地体现文献的社会网络特性以及热度.

由基于社交网络的模型评估的文献数据被表示为 dataset 1,而 dataset 2 表示由基于传统媒体的模型评估的文献数据. 为了评估这两种模型挖掘出的文献在人工智能和图像领域的影响,本文引入 *NCII* 指数来测量文献的影响力. 论文的引用次数与其出版时间有很大关系,一般来说,论文出版时间越早,引用的可能性就越大. 这导致在不同时间出版的论文很难比较它们的影响力. 因此,考虑到出版时间对参考文献数量的影响,因此提出了一种新的 *NCII* 指数[17],其计算公式为

$$NCII = \frac{\text{论文引用次数}}{\text{当前时间} - \text{论文发表时间}}. \tag{6}$$

以最近五年内的文献作为前沿信息,通过式(6)计算传统媒体和社交网络媒体的文献影响力,验证两种媒体识别出的科研文献的时效性及影响力,结果如图 1 所示. 从图中可知,社交媒体挖掘的文献的影响力值均大于传统媒体,说明社会网络媒体挖掘的文献数据更具时效性和影响力.

2.2　主题模型实验

利用吉布斯抽样的 LDA 主题模型,以 dataset 1 为实验数据挖掘出 50 个潜在主题及其代表关键词. 主题是由一系列关键词组成,而每个词对主题的贡献率各不相同,因此,选择每个主题贡献率最大的 8 个单词表征该热点主题. 根据不同主题的关键词表示,本文对各个热点主题进行话题语义抽象. 由于篇幅有限,表 5 给出了 10 个主题的挖掘结果.

同时,本文分析了主题模型计算出的潜在话题分布情况如图 2 所示,从图 2 可知,自然语言处理、算法优化、情感分析、深度学习等热点主题在“artificial intelligence and image processing”领域较其他主题占的比重较大,更为热门;而图像识别、大数据应用、可视化等热点主题在该领域发展较为均衡.

2.3　传统与社会网络媒体挖掘对比实验

本文以 dataset 1 和 dataset 2 为实验数据进行学科热点主题挖掘,选取对热点主题贡献最大的文献的发表时间作为该主题的热点时间对比分析,两种媒体类型挖掘的热点主题对比结果如表 6 所示.

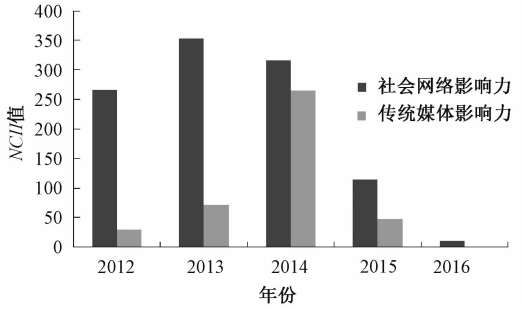

图1 社会网络与传统媒体 *NCII* 影响力

Fig. 1 The *NCII* of social-network and tradition-media influence

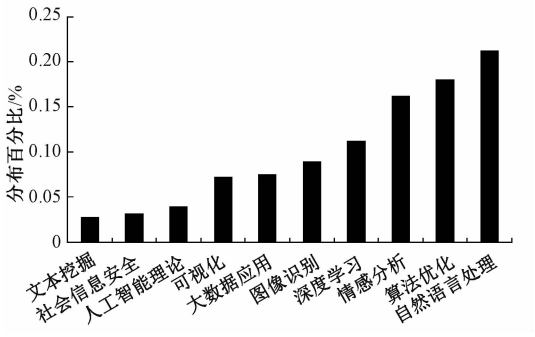

图2 热点主题潜在话题分布

Fig. 2 Potential topics distribution of hotspots

表5 前沿热点主题

Tab. 5 Frontier hot topics

热点主题(话题语义)	关键词
自然语言处理	language; processing; linguistic; learning; word; structure; information; syntactic; sentence; chunks
情感分析	emotional; brain; psychological; fear; specific; categories; affective; facial; discrete; instances
社会信息安全	social; search; security; web; access; information; sentiment; privacy; government; tools
文本挖掘	data; web; twitter; text; media; wikipedia; information; research; sentiment; computer
语音识别	speech; sound; strategy; signals; recover; spectral; implementation; video; audio; impaling
推荐系统	user; recommendation; system; evaluation; experience; visualization; quality; algorithm; collaborative; framework;
大数据应用	data; learning; big; distributed; mining; mapreduce; large; hadoop; framework; research
深度学习	learning; machine; deep; processing; mining; computing; data; techniques; analysis; clinical
算法优化	algorithms; time; set; number; analysis; model; results; space; small; structure
图像识别	image; recognition; vision; computer; features; object; active; search; number; training

从表6可知,两者有2个热点话题相似,其中相似话题“情感分析”和“自然语言处理”的热点时间不同,社会网络媒体挖掘的这两个热点主题时间比较新,原因是随着深度学习的深入研究,直接推动情感分析和自然语言处理等相关技术的发展,使两者也成了较新的研究热点.除了共同热点外,社会网络媒体挖掘的热点主题还包含一些特有的前沿热点信息,如“社会信息安全”、“大数据应用”、“算法优化”和“可视化”等,这些热点概念提出时间较晚,概念较新,近年来在学术著作中有大量的研究,其热门趋势也越来越明显,故也成了该领域的研究热点.

2.4 主题模型与共词分析法对比实验

同时,为了进一步验证本文方法的挖掘结果更具学科准确性,以 dataset 1 为数据分别采用 LDA 主题模型和基于关键词的共词分析法[14]进行学科前沿热点挖掘对比实验,表7给出了这两种方法的热点挖掘结果.

从表7可得,两种方法挖掘的热点有30%是相似的.本文挖掘“人工智能与图像处理”领域的前沿热点,主题模型挖掘出关于图像处理方面的热点有“图像识别”,其在2015年过后被广泛关注,是该领域的研究热点;然而传统的研究方法并没有挖掘出关于图像方面的热点,故该方法存在一定的不足.除上述热点主题均只通过主题模型挖掘出,传统研究方法也并未挖掘出“深度学习”、“文本挖掘”、“可视化”、“社会信息安全”等新技术或新热点.通过知网学术趋势(http://trend.cnki.net/TrendSearch/),查询各热点主题发展趋势,以学术关注度最大的年份作为该热点主题的学术关注时间;发现共词分析法挖掘的热点“决策支持”、“行为研究”、“推理”和“认知科学理论”最大学术关注度的时间偏离目前时间,说明其研究已久,故不能作为该领域的前沿热点;总体而言,主题模型挖掘的热点较共词分析法挖掘的热点更准确.

表 6　两种媒体类型挖掘的热点主题对比

Tab. 6　Hot topic results comparison of two media types

社会网络媒体挖掘热点		传统媒体挖掘热点	
热点主题	热点时间/年	热点主题	热点时间/年
情感分析	2016	量子计算机	2015
社会信息安全	2016	认知科学理论	2014
文本挖掘	2016	虚拟现实	2014
图像识别	2015	情感分析	2014
可视化	2015	蒙特卡洛树搜索	2014
自然语言处理	2014	医学成像	2011
算法优化	2014	自然语言学习	2010
人工智能理论	2013	图像处理	2007
大数据应用	2012	逻辑推理	2006
深度学习	2012	人类行为研究	2000

表 7　两种方法热点主题挖掘结果对比

Tab. 7　Hotspot mining results comparison of two methods

主题模型挖掘的热点		共词分析法挖掘的热点	
热点主题	学术关注时间/年	热点主题	学术关注时间/年
图像识别	2015	人机交互	2015
文本挖掘	2015	知识演化	2015
可视化	2015	决策支持	2014
深度学习	2015	感知	2015
算法优化	2015	行为研究	2011
社会信息安全	2013	推理	2009
人工智能理论	2013	认知科学理论	2008
自然语言处理	2013	自然语言处理	2013
情感分析	2015	情感分析	2015
大数据应用	2015	大数据应用	2015

3　总结

针对以往挖掘学科前沿热点存在时滞过长等不足,本文提出基于社会网络关注度的学科领域文献热度评价模型挖掘学科前沿热点.通过对各指标进行相关性获取相关属性,采用主成分分析划分社会与传统属性,构建具有社会网络关注度的文献热度评价模型.以“artificial intelligence and image processing”学科文献记录数据为实验数据,利用构建的文献热度评价模型识别该学科有影响力和热度的文献,由于文献内容冗余和有噪声,故本文采用在文本抽取中效果较好的 LDA 模型,通过两组对比实验,得出自然语言处理、算法优化、情感分析、深度学习等热点主题在人工智能和图像处理领域较其他主题更为热门,图像识别,大数据应用、可视化、人工智能理论、信息安全等热点发展趋势较均衡的结论,同时也验证了本文挖掘的学科领域前沿热点知识的前沿性、时效性和准确性.

参考文献:

[1]　王慧. 社会网络分析在学科热点分析中的实证研究[D]. 镇江:江苏大学, 2010.

[2]　LUO Y,ZHAO S L,LI X C,et al. Text keyword extraction method based on word frequency statistics[J]. Journal of computer applications, 2016, 36(3): 718 – 725.

[3]　XIE P. Study of international anticancer research trends via co-word and document co-citation visualization analysis[M]. New York: Springer-Verlag, 2015.

[4]　XIE Y F. Hotspots of ecological and environmental risk research in China based on multidimensional scaling analysis and cluster analysis[J]. Advanced materials research, 2013, 807(2):641 – 646.

[5]　XIE Y F. Hotspots and trend of water transfer research in China based on clustered analysis and social network analysis[J]. International journal of digital content technology and its applications, 2013, 7(3):534 – 541.

[6]　江贺, 陈信, 张静宣,等. 软件仓库挖掘领域:贡献者和研究热点[J]. 计算机研究与发展, 2016, 52(12):2768 – 2782.

[7]　XU T Y, QU H N, ZHAO S S, et al. The visualization analysis of research hotspot and frontier technology of the smart power distribution and utilization based on the cite space[J]. Energy and power engineering, 2017, 9(4):515 – 524.

[8]　ZHAO R, WEI M. Impact evaluation of open source software: an Altmetrics perspective[J]. Scientometrics, 2017, 110(2): 1 – 17.

[9]　YU D, LI D F, MERIGO J M, et al. Mapping development of linguistic decision making studies[J]. Journal of intelligent and fuzzy systems, 2016, 30(5):2727 – 2736.

[10]　PENG Y L, LIN A W, WANG K, et al. Global trends in DEM-related research from 1994 to 2013: a bibliometric analysis[J]. Scientometrics, 2015, 105(1):347 – 366.

[11]　YIN Z Y, CHEN D Y, LI B F. Global regulatory T-cell research from 2000 to 2015: a bibliometric analysis[J]. Plos one,

2016, 11(9):0162099.
[12] WU Y, DUAN Z G. Visualization analysis of author collaborations in schizophrenia research[J]. BMC psychiatry, 2015, 15(1):1-8.
[13] 徐佳俊, 杨飚, 姚天昉,等. 基于LDA模型的论坛热点话题识别和追踪[J]. 中文信息学报, 2016,30(1):43-49.
[14] DING W Y, CHEN C M. Dynamic topic detection and tracking: a comparison of HDP, C-word, and cocitation methods[J]. Journal of the association for information science and technology, 2014, 65(10):2084-2097.
[15] 吴岘辉, 张晖, 杨春明,等. 一种话题相关的微博意见领袖挖掘算法[J]. 小型微型计算机系统, 2014, 35(10):2296-2301.
[16] 由庆斌, 韦博, 汤珊红. 基于补充计量学的论文影响力评价模型构建[J]. 图书情报工作, 2014, 58(22):5-11.
[17] HOLSAPPLE C W,JOHNSON L E,MANAKYAN H et al. Business computing research journals: a normalized citation analysis[J]. Journal of management information systems, 2015, 11(1):131-140.

Subject Frontiers Hot Spots Mining Based on Social Network Attention

ZHANG Hui[1], YANG Xiaoyan[2], ZHAO Xujian[2], YANG Chunming[2], LI Bo[2]
(1. *School of Science, Southwest University of Science and Technology, Mianyang* 621010, *China*;
2. *School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang* 621010, *China*)

Abstract: Mining subject hotspots from scientific literature data was an urgent problem in the academic and industry field. The existing methods couldn't consider the popularity of a scientific paper as a propagating object in the social network, which reflected the practical popularity of information diffusion timely. A hotspot mining method was proposed. Firstly, the literature popularity evaluation model was constructed for social network attention scoring. The model consisted of data correlation analyzing, correlated attributes partitioning and social network attention degree factor mining. Meanwhile, LDA model was adopted to discover the hotspots from the scientific papers. Finally, contrastive experiments were performed on the topic of artificial intelligence and image processing. The results showed that the method could timely discover recent hotspots, so it was effective in hot frontiers mining.

Key words: literature evaluation model; social network attention; correlation analysis; LDA model; factor mining

(责任编辑:方惠敏)