情报科学
第35卷 第7期 2017年7月

# 基于改进的TF*IDF方法分析学科研究热点
## ——以情报学为例

刘小慧,李长玲,冯志刚

(山东理工大学 科技信息研究所,山东 淄博 255049)

**摘　要:**【目的/意义】提出一种TF*IDF改进算法,用于全文分词后的语词权重计算,提取高权重语词,分析学科研究热点。【方法/过程】以万方数据库中2015年《情报学报》的载文为例,对每篇文章全文分词,用改进的TF*IDF方法计算语词权重。【结果/结论】发现该改进算法准确可行,且运用该方法分析得到,用户研究、大数据、情报学、社交网络、技术领域、文献作者、突发事件、零被引等,是2015年情报学的研究热点。

**关键词:** 研究热点;TF*IDF;全文分词;情报学

**中图分类号:**G250.2　　**文献标识码:**A　　**文章编号:**1007-7634(2017)07-82-06

DOI:10.13833/j.cnki.is.2017.07.015

# Analyze of Subject Research Hot Spots Based on An Improved Algorithm of TF*IDF
## ——Taking Information Science for Example

*LIU Xiao-hui, LI Chang-ling, FENG Zhi-gang*

(*Institute of Scientific and Technological Information, Shandong University of Technology, Zibo* 255049, *China*)

**Abstract:**【Purpose/significance】This paper comes up with an improved algorithm of TF*IDF to calculate the weight of words from papers after the word segmentation, extract high weight words and determine the research hot spots.【Method/process】Taking the articles published in Journal of the China Society for Scientific and Technical Information which from Wan Fang data platform in 2015 as an example, segment the word of every article, then calculate the weight of words by the improved algorithm of TF*IDF.【Result/conclusion】This improved algorithm is accurate and feasible. And through analyzing, it is found that the research hot spots of information science in 2015 are user researches,big data,informatics,social network,technical field,the author,emergency and non-citation and so on.

**Keywords:** research hot spots; TF*IDF; word segmentation for the full text; information science

## 1 引　言

掌握学科前沿研究热点,有助于把握科学研究的动态变化、学科核心研究领域和发展主导趋势,为学者的研究工作提供借鉴,促进知识的创新。

词语是信息传递的最小单位,通过文章中有代表性的词语便可大致分析出文章的研究主题,而关键词是被普遍认为能代表文章主题的词语,所以很多学者基于文章关键词分析某阶段的学科研究热点。苏新宁、肖明分别对图情档学和情报学的高频关键词进行统计分析,总结该学科的研究热点,并通过比较不同年份各关键词在数量上的变化,对该学科的研究热点变化趋势进行分析[1-2]。李长玲等用共词聚类方法,分析2002-2006年的硕士学位论文中的高频关键词,得出当时的研究热点[3]。付鑫金等对博硕学位论文的高频关键词进行统计,构造共词矩阵,通过多维尺度分析等方法探析当时的研究热点[4]。柯平等借助Ucinet等工具对SSCI收录的22种国外期刊的论文进行关键词统计分析,把国外图情学研究热点分成三类[5]。赵蓉英基于关键词词频分析,绘制热点词汇分布图,展示情报学的研究热点[6]。Huiling

**收稿日期:**2017-01-04

**基金项目:**国家社科基金项目(16BTQ078)

**作者简介:**刘小慧(1990-),女,山东济南人,硕士研究生,主要从事信息计量与科学评价研究;通讯作者:李长玲.

Chen和Guoqing Zhao基于引文分析理论，利用CitespaceⅡ构建“知识可视化”领域的引文网络，用文章的标题、关键词做节点标注，分析该领域的研究热点与前沿[7]。隋鑫等用内容分析法与社会网络分析法对论文的主题、关键词共现频次与中心性布局进行分析，通过对比三者的结论，得到2009–2013年图情领域的研究热点[8]。吴德志通过分析科研立项的相关数据，统计其关键词，总结图书情报学领域的研究热点[9]。

通过以上分析可以看出，学者们多是基于文章的关键词探析研究热点。但是，关键词的选取带有很强的主观色彩，概括文章的准确性不能保证，甚至某些文章的关键词并不能概括文章的研究内容；同时，文章的关键词数量在3–8之间不等，显然，关键词数量过多或过少均意味着这些关键词对于概括文章研究内容的重要性各不相同，若简单地将某阶段文献中某个关键词的频次累加，作为该词在此阶段的研究热度，显然是不准确的。

传统的TF*IDF是用于计算文档关键字权值的一种重要方法[10]，该方法有自己的优点，即其所处理的对象可以是基于文章分词得到的来自全文的所有语词，通过计算所有语词的TF*IDF值评价各语词的重要性，这样便可以把文中的每一个词语都考虑在内。但该方法也有一定的不足，不少学者针对其不足之处，结合要实现的目的，提出了不同的改进方法。张瑜基于传统TF*IDF方法引入类间偏斜度、类内离散度和权重调整因子，得到了一种效果更好的文本分类方法[11]。常凯对传统的TF*IDF方法做了改进，使之用于垃圾邮件过滤，增加了有区分度的特征词的权值，实现了准确度更高的过滤[10]。Benatallah B等人基于传统TF*IDF方法提出了一种新的评估语义相似度的计算方法[12]。Yong Zhuang提出了一种基于文本位置的TF*IDF改进算法，用于电子信息特征提取[13]。

传统的TF*IDF方法经过不同的改进，可以实现不同的功能。基于以上分析，本文提出了一种基于传统TF*IDF的改进方法，使之适用于学科研究热点的提取。这种基于全文分词的评价方法，可能会比仅用文章关键词判断学科研究热点更准确。

## 2 研究方法

### 2.1 传统TF*IDF算法

用于计算文档关键字权值的传统的TF*IDF，主要考虑文本特征词的频次（Term Frequency，TF）和倒排文档频率（Inverse Document Frequency，IDF）。TF是指特征词$term_k$在给定文档中出现的次数，TF越大，表明该特征词对该文档越重要。IDF是指在一个文档集中，特征词$term_k$按文档统计出现的频繁程度。其公式为：

$$IDF(term_k)=\log\frac{N}{n_k+\beta}$$

式中，N为文档集中文档总数，$n_k$为出现过特征词$term_k$的文档数，即$term_k$的逆文本频数。b是为保证分母有意义而设置的常数，通常取0.01、0.1或者1。该算法的主要思想是：如果某特征词在越多的文档中出现过，则该特征词的区别性就越低，重要性越小；反之，则更具有区别性，重要性也就越高[14]。

对于提取研究热点而言，传统TF*IDF方法有一定的借鉴意义：在该方法中，使用的词语不是仅仅来源于文章已经给定的关键词，而是来源于文章全文，统计全文中特征词的词频，更具有客观性。同时，传统TF*IDF算法也有缺陷：按照传统TF*IDF算法的核心思想，包含某特征词的文章越多，该词就越不重要，对应的IDF值就越小，导致该词由TF*IDF方法算得的权值越小。但对于研究热点的提取而言，在一定数量范围之内，随着逆文本数量的增多，特征词是越来越有意义的，但超过一定数量范围，特征词会随着逆文本数量的增多，意义越来越小。基于此，本文提出了适用于寻找研究热点的TF*IDF改进算法，并界定了逆文本数量的区间范围。

### 2.2 TF*IDF的改进算法

（1）改进TF。

将搜集的测试文档集中所有文档看做一个整体，记为I，文档总量为N。将文档集I中所有文档利用分词软件进行全文切分，得到若干语词。本文将传统TF*IDF中的TF定义为特征词$term_k$在文档集I中出现的总频率，记为$TF_k$，即

$$TF_k=\frac{\sum_I tf_{term_k}}{M_I}$$

式中，分子为特征词$term_k$在文档集I中出现的总频次，分母$M_I$为文档集I中所有特征词的总数（所有语词的总词频数，含重复语词）。把所有语词都放在一个集合中考察，且通过分母统一了各语词的量纲，如此取得的权值不仅更具客观性，也不会致使最终算得的权值数据偏大。

（2）改进IDF。

根据1.1的分析可知，为了识别学科研究热点，TF*IDF公式中，在一定数量范围内，因子IDF应与文档集I中包含特征词$term_k$的文档数$n_k$成正比，但传统TF*IDF中二者却是成反比。为了调节这个关系，本文把$IDF(term_k)$取倒数，即变为DF，为突出本方法的思想基础，本文记为$IDF_k$，即：

$$IDF_k=\frac{1}{IDF(term_k)}=\frac{1}{\log\frac{N}{n_k+\beta}}$$

此处，$IDF_k$可以看作为一个调节因子，主要解决这样一个问题：如果特征词$term_k$只在一篇很长的文档中出现过，由于所在的文档比较长，难免会使该特征词的频次和计算得到的频率偏高，但是它毕竟只出现在一篇文档中，所以其重要性理应不是很大。但由于$IDF_k$与包含特征词$term_k$的文档数$n_k$恰好成正比，所以在因文档过长导致的特征词频次和频率偏高这一问题中，$IDF_k$便起到了调控作用，使特征词$term_k$在一定区间内出现的文章数越多，权值就越大。如此便可以更加有效地识别学科研究热点。

在统计标引法理论研究中，帕欧(M.L.Pao)经研究发现，文献中有效词的词频应在 $n=[(1+8I_1)^{1/2}-1]/2$ 附近，即 $(n-a,n+a)$ 之内。其中，$I_1$ 是出现频次为1的词的数量，a值的大小由经验确定，但不宜过大[15]。

参考帕欧的统计标引法理论确定特征词 $term_k$ 的文档分布频数区间为：$n=[1+8N^{1/2}-1]/2$ 附近，取 $a=\sqrt{n}$，即 $(n-\sqrt{n},n+\sqrt{n})$ 之内。

(3)改进后的TF*IDF计算公式。

基于以上分析和界定，特征词 $term_k$ 改进后的权重计算公式为：

$$\omega_{term_k}=TF_k*IDF_k=\frac{\sum_l tf_{term_k}}{M_l}\times\frac{1}{\log\frac{N}{n_k+\beta}}$$

其中，$n_k$ 的定义域为 $(n-\sqrt{n},n+\sqrt{n})$，$n=[1+8N^{1/2}-1]/2$。

式中，频率($TF_k$)反映特征词 $term_k$ 在所有文档中的重要性；$IDF_k$ 使该词在一定文档数量范围内，随逆文本频数越大越重要。二者结合，使学科研究热点的识别方法更有效。

## 3 实证分析—以2015年《情报学报》载文为例分析情报学研究热点

《情报学报》作为中国科技情报学会的会刊，不仅是我国情报学研究的主要情报源，还是反映我国情报学研究成果的主要窗口，在较大程度上代表着我国情报学理论研究和实践研究的进展及今后发展趋势[16]。因此，本文选取2015年《情报学报》载文作为实证研究的样本数据。

### 3.1 数据来源与处理

本文选择万方知识服务平台作为检索数据库，以“刊名=情报学报 并且 年份=2015”作为检索途径，检索日期为2016年3月20日。共检索到109篇论文，去掉《征文通知》、《投稿须知》等非学术性文章，余97篇。下载全文并做如下处理：

(1)文本处理。

在万方知识服务平台下载得到的文章格式为.PDF，为方便后续分词，使用Adobe Reader将其转化为.TXT文本文档；去掉每篇文章的页眉标注、页脚标注、作者简介和参考文献；使用文本文档中的替换功能，去掉文件中的空格和换行符，以免分词处理时因空格或换行符影响分词的准确性。

(2)全文分词。

对比多种分词系统的分词效果，本文最终选择使用NLPIR/ ICTCLAS2014中文分词系统进行全文分词，该系统由北京理工大学张华平副教授研制开发。为提高分词准确性，本文添加了分词词典，其内容为2012年发布的《中文核心期刊要目总览》中情报学的9种主要期刊(情报学报、情报科学、情报理论与实践、情报杂志、情报资料工作、图书情报工作、图书情报知识、图书与情报和现代图书情报技术)中2015年所有文章的关键词。对样本数据中97篇文章分别做全文分词，其中某一篇文章的部分分词结果如图1。

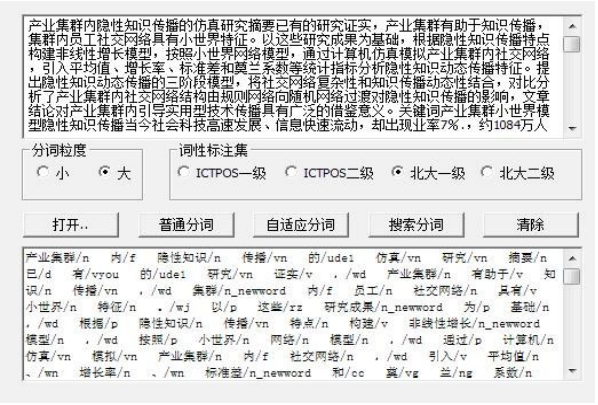


图1 NLPIR/ ICTCLAS2014分词结果示例

由图1可以看出，该分词系统有以下几个优点：①把文章分成了单个的语词，没有文字重复，也没有文字遗漏；②可以利用词典功能，使分词结果更加准确，不会把比较长的词语分的太过琐碎，比如“社交网络”，而不是“社交”和“网络”；③每个词语之后都紧跟着相应的词性，恰好可以满足后续处理中筛选动词词语和名词词语的需求。

(3)语词筛选与统计。

由于介词、形容词等意义不大，因此去掉97篇文章分词结果中这类不能说明研究内容的语词，筛选保留文档的动词和名词，并设置要求保留语词的长度大于等于3个字；使用Bibexcel分别统计每篇文档中各语词的词频，用VBA自编程序统计各语词的逆文本频数，即各语词在文档集I中出现的文档数($n_k$)，以便下一步确定有效的文档数取值范围(即 $IDF_k$ 中 $n_k$ 的取值范围)及计算各语词的 $IDF_k$。部分VBA自编程序对应的伪代码如下：

```
For 文章1 to 文章97
For 语词i
If  在当前文章中出现过
then  逆文本频数= 逆文本频数+1
End If
Next
Next
```

(4)限制有效词文本频数。

为筛选有效词，需计算 $n_k$ 定义域，以限定语词出现的文档频数不太高或太低。由1.2部分知，$n_k$ 的定义域为 $(n-\sqrt{n},n+\sqrt{n})$，$n=[1+8N^{1/2}-1]/2$。由N=97，得 $n_k$ 的取值为(13.44-9.85,13.44+9.85)，通过取整函数，得[3,23]，即文档数在[3,23]的语词为有效词。

(5)同义词合并。

经过全文分词处理，并有效词筛选后，对语义相同或相似的语词进行合并。如“用户需求”、“用户心理”、“用户行为”等，均用“用户研究”代替，“零被引”、“零被引率”、“未被引”等，均用“零被引”代替。

以其中某篇文章《基于LDA的社会化标签综合聚类方法》为例展示筛选并统计的结果，鉴于篇幅限制，仅展示该文

章前16个高频语词的词频及逆文本频数，统计结果见表1。

表1 词频及逆文本频数统计结果(部分数据)

| 语词 | 词频 | 逆文本频数($n_k$) | 语词 | 词频 | 逆文本频数($n_k$) |
|---|---|---|---|---|---|
| 标签语料库 | 19 | 1 | 混合主题 | 9 | 1 |
| 标签聚类 | 16 | 1 | 语义相似度 | 7 | 9 |
| 聚类结果 | 16 | 6 | 语料库 | 7 | 10 |
| 社会化标注系统 | 14 | 2 | 隐含主题 | 5 | 3 |
| 标注信息 | 14 | 3 | 整体语义 | 5 | 1 |
| 潜在主题 | 13 | 2 | 语义分析 | 4 | 6 |
| 概率分布 | 13 | 5 | 主题模型 | 3 | 5 |
| 综合聚类方法 | 13 | 1 | 向量空间模型 | 2 | 10 |
| … | … | … | … | … | … |

在该文章中，作者标注的关键词有：社会化标注系统、标签聚类、主题模型和潜在语义。结合文章标题和原文中作者给出的关键词，对比上表，可以发现，由文章全文分词得出的有效高频语词不仅涵盖了文章的关键词，而且能更加全面的反映文章内容。这说明，基于全文分词得到的词语，用改进的TF*IDF方法获取学科研究热点，应该更有实际意义。

### 3.2 $TF_k$*$IDF_k$ 值的计算

97篇文章经过以上五步处理后，将所有如表1的处理结果汇总到同一个Excel文件中，使用VBA语言自编程序统一计算每个语词的$TF_k$和$IDF_k$，并计算$TF_k$*$IDF_k$值，计算所用的部分VBA语言伪代码如下：

```
For  t= 文章1 to 文章97
For  语词i
If   文章1中的语词i = 文章t中的语词i
then 文章1中语词i的TFk= 文章1中语词i的TFk+文章t中的语词i的TFk
End If
Next
Next
```

为便于观察与比较，对上述计算结果做归一化处理。本文采用Frobenius范数(也称2-范数)做归一化处理，该归一化的表达公式为：

$$Y=\left(\frac{x_1}{\sqrt{\sum_{i=1}^{N}x_i^2}},\cdots,\frac{x_N}{\sqrt{\sum_{i=1}^{N}x_i^2}}\right)$$

经过上述程序计算、归一化处理后的$TF_k$*$IDF_k$原始值及标准值，按照由高到低排序，标准权值大于0.1的语词及相关数据见表2。

鉴于目前许多学者通过统计词频的方法寻找研究热点，为便于两种方法结论的对比，本文在表2中一并列出权值较高的语词在全文中出现的频次和其逆文本频数。

改进的$TF_k$*$IDF_k$方法不仅考虑了语词的词频，还考虑到语词出现的文档数(即表2中的逆文本频数)，这比仅依靠频次判定学科的研究热点更有效。表2中，“大数据”的词频低于“零被引”，但其逆文本频数高于“零被引”，综合衡量其研究热度$TF_k$*$IDF_k$值大于“零被引”。就目前的研究现状，“大数据”的研究热度高于“零被引”，应该是客观事实。所以，本文提出的$TF_k$*$IDF_k$改进方法在词频和语词的逆文本频数之间起到了较好的调节作用，对于分析学科研究热点是有效的。

表2 语词$TF_k$★ $IDF_k$值及其他值比对表

| 语词 | 原始$TF_k$*$IDF_k$ | 标准$TF_k$*$IDF_k$ | 总词频 | 逆文本频数 |
|---|---|---|---|---|
| 技术领域 | 0.022 | 0.535 | 857 | 19 |
| 用户研究 | 0.020 | 0.499 | 702 | 13 |
| 文献作者 | 0.009 | 0.248 | 719 | 10 |
| 大数据 | 0.009 | 0.242 | 331 | 11 |
| 情报学 | 0.008 | 0.221 | 233 | 22 |
| 社交网络 | 0.008 | 0.219 | 244 | 15 |
| 零被引 | 0.008 | 0.214 | 417 | 4 |
| 突发事件 | 0.007 | 0.183 | 299 | 6 |
| … | … | … | … | … |

注：★所示的权值均为四舍五入的估计值

表2中，$TF_k$*$IDF_k$标准权值大于0.1的8个语词：技术领域、用户研究、文献作者、大数据、情报学、社交网络、零被引、突发事件，综合考量了出现频次与文档数，因此它们应该是该年情报学学科的主要研究热点。那么，这些研究热点的主要研究内容是什么，需要进一步细化处理与分析。

### 3.3 学科研究热点分析

为细化分析8个情报学研究热点语词的相关研究内容，下载获取样本数据《情报学报》97篇文章作者给定的关键词，在NetDraw2.0中构建关键词共现网络。

不同于全文分词得到的语词，关键词是作者选定概括载文内容的词语，为详细展现2015年的研究内容，且考虑到97篇文章的关键词数量不多，所以不做文章给定关键词的同义合并处理。为重点突出地分析8个语词的相关研究内容，通过“Layout-Ego Networks(new)”命令，在该网络中筛选8个语词的相关关键词，呈现这些语词的关键词子网络，阈值设定为大于1，如图2所示。

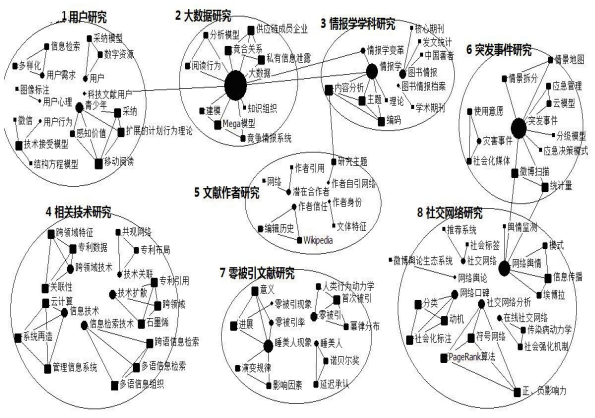


图2 高权重词语与相关关键词的共现网络

图2中，圆形节点为与8个高权重语词语义相关的关键词，方形节点为与之出现在共现网络中的其他关键词。

结合表2与图2，可以分析得出，基于《情报学报》的样本

数据,2015年情报学的前沿研究主题主要在以下几个方面:

(1)用户研究。与“用户”相关的关键词主要有青少年、用户心理、用户行为、用户需求、科技文献用户等。研究领域主要包括基于技术接受模型的微信用户信息发布行为研究、用户图像标注心理与行为、青少年用户移动阅读采纳行为及大数据环境下科技文献用户的知识组织模型等。此类研究之所以成为热点,主要是因为人工智能技术日趋成熟,人们的学习、阅读和娱乐越来越数字化、智能化,为了提供更加人性化的服务,此类关于用户需求和使用习惯的研究成为了一种趋势。

(2)大数据研究。“大数据”是指以多元形式、多种来源搜集的庞大数据组,是人们获得新的认知、创造新的价值的源泉;大数据还是改变市场、组织机构,以及政府与公民关系的方法[17]。大数据时代的到来,给企业发展带来便利的同时,也带来了一定的威胁,私有信息泄露直接影响了企业之间的竞争策略,大数据研究在竞争情报研究领域已引起了学者的重视。同时大数据时代的到来也对知识组织方式产生了影响。因此,2015年情报学关于大数据的研究内容主要包括:大数据环境下企业竞合关系的变化研究、大数据环境下基于Mega模型的竞争情报系统研究、大数据环境下读者阅读行为的分析模型研究和大数据环境下知识组织的变化研究等。

(3)情报学学科研究。与“情报学”相关的关键词主要有情报学、图书情报档案、图书情报、情报学变革等。关于情报学学科研究的热点可以归结为两个方面。一方面,情报学是一个新兴学科,学科基础理论并不完善[6],为了明确情报学的学科研究对象、研究内容以及丰富发展情报学,情报学基础理论一直是学者们关注的焦点。2015年相关的研究主要有基于内容分析的情报学理论及主题研究、通过统计中国著者在国际图书情报领域核心期刊的发文量分析我国情报学研究进展等。另一方面,主要是关于大数据时代到来引起的情报学学科变革研究。大数据环境下,数据量剧增,情报学面临的环境发生了巨大变化,研究内容、方法、技术等需要改进,必然导致情报学核心内涵发生变革。

(4)相关技术研究。与“技术”相关的关键词主要有跨领域技术、技术接受模型、技术扩散、技术关联、信息技术、信息检索技术等。研究内容主要包括跨领域技术特征研究、跨领域视角的技术扩散特征研究、基于专利的相关技术研究、多语信息检索技术研究和云计算技术研究等。

(5)文献作者研究。与“作者”相关的关键词主要有作者身份、作者信任、作者自引网络、潜在合作者等。研究内容主要包括:作者研究主题的确定、潜在合作者的发现和作者信任度研究;同时,博客中作者身份的识别研究也成为网络舆情的新兴研究领域。

(6)突发事件。与“突发事件”相关的关键词有灾害事件等。这类研究主要涉及危机管理研究领域,包括当灾害事件发生时,公众对社会化媒体的使用意愿研究、微博用户在突发事件中的扫描量研究、基于云模型的突发事件应急管理研究、应对突发事件情景地图的设计研究,以及突发事件发生时的应急决策模式研究等。

(7)零被引文献研究。与“零被引”相关的关键词主要有睡美人、睡美人现象、零被引、零被引现象、零被引率等。主要研究内容包括:以诺贝尔奖为例基于被引速率识别“睡美人”文献、图情档领域零被引现象分析、通过研究文献首次被引的规律解决零被引文献中的迟滞承认、零被引文献综述研究等。近年来,关于零被引文献的研究越来越多,据统计,在情报学领域,关于零被引文献的研究自2011年的1篇相关文章到2015年已累计上升到45篇。

(8)社交网络研究。与“社交网络”相关的关键词有网络舆情、社交网络分析、网络口碑、在线社交网络、网络舆论等。研究内容主要包括:社交网络中节点的正负影响力计算方法研究、基于微博舆论生态系统的网络舆论、网络口碑传播动机识别研究、网络舆情信息传播模式研究、在线社交网络中谣言传播与社会强化管理机制研究、基于社交网络互联网推荐系统研究综述等。

本文以《情报学报》2015年97篇研究论文为样本,在基于全文分词的基础上,用改进的$TF_k$*$IDF_k$方法,分析得到情报学研究热点,包括技术领域、用户研究、文献作者、大数据、情报学、社交网络、零被引、突发事件等。其中情报学、网络舆情等主题与王知津等人在《我国情报学研究热点及问题分析—基于2010–2014年情报学核心期刊》一文[18]中基于关键词频次和布拉德福理论思想得到的我国2010–2014年研究热点结论相似;用户研究、大数据研究、相关技术研究等主题,与赵蓉英等人在《2010–2015年国内外情报学研究热点可视化比较》一文[19]中,通过词频统计方法发现的2010–2015年国外研究热点结论基本一致。这一方面说明《情报学报》具有引领或跟踪情报学国际前沿的作用;另一方面说明用改进的$TF_k$*$IDF_k$方法识别学科热点是有效的。

## 4 结 语

学科研究热点是信息计量学的重要研究领域,对揭示学科的发展方向和发展规律有重要意义。本文基于传统TF*IDF提出一种适合提取学科研究热点的改进方法,记为$TF_k$*$IDF_k$。将万方知识服务平台2015年《情报学报》中的载文作为样本数据,经过全文分词,用$TF_k$*$IDF_k$计算所有语词的权值,以表示其在2015年情报学研究中所做的贡献,从而发现本学科研究热点。

本文的不足之处在于:由于万方数据库数据录入的迟滞性,在本文检索样本数据时,2015年的文献未被收录完整,因此导致本文中的实证研究数据并不是2015年《情报学报》的全部数据,所以可能使结果有一定的局限性,但这并不影响该方法的验证。

## 参考文献

1 苏新宁. 图书馆、情报与文献学研究热点与趋势分析(2000~2004)——基于CSSCI的分析[J]. 情报学报,

2007, 26(3): 373-383.
2 肖 明, 李国俊, 杨 楠. 基于词频分析的国内情报学研究热点(1998～2007)[J]. 情报杂志, 2009 ,(8): 21-25.
3 李长玲, 翟雪梅. 基于硕士学位论文的我国图书馆学与情报学研究热点分析[J]. 情报科学, 2008,(7): 1056-1060.
4 付鑫金, 方 曙, 庞弘燊. 基于共词分析的我国情报学博硕士学位论文研究热点分析[J]. 情报科学, 2011,(11): 1722-1725.
5 柯 平, 贾东琴, 李廷翰. 2010年国外图书馆学情报学研究热点分析[J]. 情报科学, 2011,(9): 1281-1288.
6 赵蓉英, 马丽娜. 国际情报学核心期刊与研究热点的可视化分析[J]. 情报科学, 2011,(8): 1238-1243.
7 Cheung S K S, Fong J, Kwok L, et al. The Analysis of Research Hotspots and Fronts of Know- ledge Visualization Based on CitespaceII [M]. Berlin: Heidelberg, 2012: 57-68.
8 隋 鑫, 王念祖. 2009-2013年国内图书馆学情报学研究热点分析[J]. 情报科学, 2015,(10): 61-65.
9 吴德志, 董 颖, 刘长清, 等. 基于科研立项的图书情报学研究热点分析[J]. 情报科学, 2016 ,(6): 121-124.
10 常 凯. 基于TF*IDF垃圾邮件过滤改进算法的研究[J]. 电脑知识与技术, 2010,(25): 6928-6930.
11 张 瑜, 张德贤. 一种改进的特征权重算法[J]. 计算机工程, 2011,(5): 210-212.
12 Benatallah B, Bestavros A, Manolopoulos Y, et al. Coupled Item-Based Matrix Factorization [M]. Berlin: Heidelberg, 2014: 1-14.
13 Jin D, Lin S, Zhuang Y. An Improved TFIDF Algorithm in Electronic Information Feature Extraction Based on Document Position [M]. Berlin: Heidelberg, 2012: 449-454.
14 沈志斌, 白清源. 文本分类中特征权重算法的改进[J]. 南京师范大学学报(工程技术版), 2008,(4): 95-98.
15 Pao M L. Automatic text analysis based on transition phenomena of word occurrences [J].Journal of the American Society for Information Science, 1978, 29(3): 121-124.
16 张学福.《情报学报》被引分析与研究[J]. 情报学报, 1998, (5): 387.
17 Cukier K, Mayer-Scho nberger V. Big data: a revolution that will transform how we live, work, and think [M]. Boston: Houghton Mifflin Harcourt, 2013:49.
18 王知津, 李博雅. 我国情报学研究热点及问题分析——基于2010-2014年情报学核心期刊[J]. 情报理论与实践, 2016,(9): 7-13.
19 赵蓉英, 魏明坤. 2010-2015年国内外情报学研究热点可视化比较[J]. 图书馆杂志, 2016,(8): 15-22.

（责任编辑:徐 波）