DOI:10.16353/j.cnki.1000-7490.2013.03.018

● 赵悦阳[1]，崔　雷[2]
（1. 中国医科大学　附属盛京医院图书馆，辽宁　沈阳　110004；2. 中国医科大学　信息管理与信息系统（医学）系，辽宁　沈阳　110001）

HITS 算法在文本聚类结果类别描述中的应用尝试*

摘　要：文章基于 HITS 算法，提取出聚类结果中每一类的特征词，客观地描述聚类分析结果，排除分析者的阅读能力、理解能力和归纳能力等主观性，不受所研究的文本量大小的限制，使科研人员更准确更容易分析聚类结果，为进一步研究提供服务。

关键词：算法；文本聚类；类别描述；同被引聚类分析

Abstract: Based on HITS algorithm, this paper extracts each type of characteristic word from clustering results, describes the clustering analysis results objectively, excludes the subjectiveness of the analyzers in reading capability, understanding capability and induction capability, and ignores the limitation of the size of the text volume to be studied to make the scientific research personnel analyze the clustering results more accurately and more easily so as to provide service for further research.

Keywords: algorithm; text clustering; category description; co-citation clustering analysis

聚类分析是数据挖掘中一项重要的研究方法，在解决医学、心理学、社会学以及模式识别、图像处理问题中都有着重要的作用。同被引聚类分析是文献计量学的主要方法之一，可以用来表示某一学科或专题的研究结构和状况。对同被引聚类结果的解释和分析，客观、准确地阐述每一类代表的意义或说明的内容，则是进行研究时非常重要的环节，也是需要迫切解决的问题。

通常高频词能够揭示主题，但是却无法描述类别特有的详细信息。对于聚类分析，归纳出的几类是通过计算高频引文的相似度而得到的，而每篇论文的重要性在决定它所在类的类别描述中也是起到一定作用的。

因此，发现一种方法应用到聚类结果的类别描述中，使得类别描述更加准确并且更具有该类的代表性是值得探讨的问题。

本研究基于 HITS（Hypertext-Induced Topic Search）算法，提取出聚类结果中每一类的特征词，客观地描述聚类分析结果。排除分析者的阅读能力、理解能力和归纳能力等主观性，不受所研究的文本量大小的限制，使科研人员更准确更容易地分析聚类结果，为进一步研究提供服务。

* 本文为中国图书馆学会医院图书馆委员会科学研究基金项目“HITS 算法在文本聚类结果特征提取中的应用”（项目编号：Ytwjj11002），中国医科大学第二临床学院科学研究基金项目“HITS 算法在文本聚类结果特征提取中的应用”（项目编号：MA17）的研究成果。

1　方法

1.1　HITS 算法

HITS 算法是由康奈尔大学（Cornell University）的 J. Kleinberg 博士于 1998 年首先提出的用于计算网页重要性的算法[1]。

HITS 算法包括两个指标：权威值（Authority）和中心值（Hub）。权威值用来衡量网页内容，即对于一个特定的检索，权威网页提供最好的相关信息。中心值用来衡量一个网页链接到其他网页的数量，即中心网页提供很多指向其他高质量权威型网页的超链。网页的权威值越高，表示这个网页越重要；中心值越高，表示这个网页被链接的次数越多。

HITS 算法的基本思想是：好的 Hub 型网页指向好的 Authority 网页，好的 Authority 网页是由好的 Hub 型网页所指向的网页。

执行算法：

1）将查询 q 提交给基于关键词查询的检索系统，从返回结果页面的集合中取前 n 个网页（如 $n=200$），作为根集合（Root Set），记为 S，则 S 满足：①S 中的网页数量较少；②S 中的网页是与查询 q 相关的网页；③S 中的网页包含较多的 Authority 网页。

2）将 S 扩展为基本集合（Base Set）T，T 包含由 S 指出或指向 S 的网页。可以设定一个上限如 1 000～5 000 个网页。

3）开始权重传播。这是一个递归的过程，用于决定 Hub 与权威权重的值。具体操作如下：①为基本集中的每个页面赋予一个非负的权威权重 a_p 和非负的 Hub 权重 h_p，并将所有的 a 和 h 值初始化为同一个常数，如 $a_p=1$，$h_p=1$。②Hub 与 Authority 的权重可按如下公式进行迭代计算。

$$a_p=\sum_{q|q\to p}h_q \tag{1}$$

$$h_p=\sum_{q|q\to p}a_q \tag{2}$$

公式（1）反映了若一个页面由许多好的 Hub 所指，则其权威权重会相应增加（即增加为所有指向它的页面的现有 Hub 权重之和）。公式（2）反映了若一个页面指向许多好的权威页，则 Hub 权重也会相应增加（即权重增加为该页面链接的所有页面的权威权重之和）。③每次迭代后使用公式（3）和公式（4）进行规范化处理，保证不变性。④当 a 和 h 值没有收敛时，转向②。

$$\sum_p(a_p)^2=1 \tag{3}$$

$$\sum_p(h_p)^2=1 \tag{4}$$

实验证明，经过大约 10～15 次迭代计算，a 和 h 值将趋于稳定，迭代结束。此时可设置阈值 T，将所有 a 和 h 大于 T 的网页挑选出来，排序输出查询结果。实践证明，该算法对于许多查询具有良好的查准率和查全率。

1.2 利用 HITS 算法进行聚类结果类别描述

1.2.1 数据预处理 ①截词根。在英语词汇中，一个词可能有多种形态，如词的单、复数形式的不同，英美拼写方法不同、词性不同等。本次试验将摘要以每个单词为个体，利用 Porter Stemming 算法还原英文单词的词性、词形变化，去掉前缀、后缀等。链接地址：http://tartarus.org/martin/PorterStemmer/index-old.html。这个源程序有很多语言版本，包括 C，Java，Perl，Python，CJHJ，Ruby，VB，Javascript，PHP，Delphi，Lisp，等等，这次试验使用 Java 版本。②去停词。去停词依照的是从摘要中依次读取每个单词然后和停词列表中的词对比，如果存在就去掉。停词表通过汇总网上现有的几个停词表生成，并随时添加新的停用词。

1.2.2 TF-IDF 算法为词赋权重 TF-IDF（Term Frequency & Inverse Documentation Frequency）算法是由 Salton 首次提出的[2]，是单词权重最为有效的实现方法。该算法的主要思想是：一个词在特定的文档中出现的频率越高，说明它在区分该文档内容属性方面的能力越强（TF）；一个词在文档中出现的范围越广，说明它区分文档内容的属性越低（IDF）。经过 Salton 的多次论证，信息检索领域广泛使用 TF-IDF 算法计算权重，其经典计算公式为：

$$W_{ij}=\mathrm{tf}_{ij}\times\mathrm{idf}_j=\mathrm{tf}_{ij}\times\log(N/n_j)$$

其中 tf_{ij} 指特征项 t_j 在文档 di 中出现的次数；idf_j 指出现特征项 t_j 的文档的倒数。N 表示总文档数，n_j 指出现特征项 t_j 的文档数。

TF-IDF 算法是一种统计方法，用以评估一个词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF 加权的各种形式常被搜索引擎应用，作为文件与用户查询之间相关程度的度量或评级。

这里依据 TF-IDF 算法为每个词赋予其在每一类论文集的权重，论文集则可以认为是一个 N 维的 TF-IDF 向量，N 代表所有类别论文集合的单词个数，建立论文与单词的矩阵，然后生成标准化矩阵。矩阵行代表词的节点，矩阵列代表论文节点，矩阵中的数值是 TF-IDF 值，再将矩阵导入，通过 HITS 算法提取出“权威”的词。

1.2.3 HITS 算法提取关键词 为了提取出有效的关键词，可以建立一个关于单词和论文的二维图（见图 1 中 C 部分）。

在图 1 右侧的单词节点（Term Nodes）代表所有的关键词，它们与左侧的论文节点（Document Nodes）组成网络链接。

单词和论文之间的数值是通过以下步骤得来的：首先，基于 TF-IDF 算法建立一个单词与论文的矩阵，每一行代表一个词（t_1，…，t_m），每一列代表一篇论文（d_1，…，d_n）。矩阵中的数值是 TF-IDF 值。如果某个单词没有出现在某篇论文中，那么矩阵中的值用 0 表示（见图 1 中 A 部分）。然后，应用 Cosine 系数将矩阵标准化（见图 1 中 B 部分），这样得出的单词与论文之间对应的数值就是图 1 中 C 部分的数值。

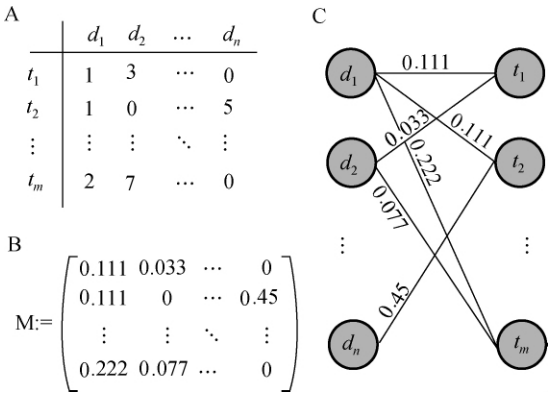

图 1 单词与论文之间的表示图

通过 HITS 算法可以发现“权威”的单词和“中心”

的论文。原理可以引申为：一个单词如果在中心论文里出现许多次，那么它具有较高的权威性；而如果一篇论文中包含许多的权威单词，则它具有较高的中心性。因此可以推断，含有许多“权威”单词的论文一定是“中心”的核心论文，而在许多“中心”的核心论文中出现的单词一定是“权威”的关键词。对于聚类结果中的每一类，论文的主题是相似的而且是某领域的核心论文，HITS 算法能够有效地发现关键词和更为核心的论文。

1.3 评价 HITS 算法提取的关键词

为了方便对算法提取出的关键词进行评价，选用笔者于 2005 年发表在《情报学报》上的论文《专题文献的同被引聚类分析在表现学科专业发展历史的可靠性评价》中的样本数据和聚类分析结果[3]。

将 HITS 算法提取出的关键词按照 Authority 值由高到低排序，每一类截取排名靠前的 10 个词用于评价 HITS 算法提取的关键词。选取前 10 名的关键词的原因是因为在信息检索中，关键词作为一个揭示文本主题的单位，标引的关键词数适合定在 9 个词以内。首先，根据“7±2”认知规则，“9”是一般用户不需要特别努力就能够记住的词条个数；其次，文献的关键词手工标引词一般为 3～5 个，最多小于 10 个[4]。

以往描述聚类结果每一类的类别时，是通过阅读全文和参考 PUBMED 标引的主题词来归类。这里我们就比较算法提取的关键词和 PUBMED 标引的主题词，一组是关键词，一组是主题词，将它们分别与参照论文中的聚类分析结果比较。主要比较两个指标：①二者提取的关键词的准确性。②二者提取的关键词区分聚类结果各类别的能力。对于第一个指标可以根据查准率的原理引申出：

准确率 = 算法（PUBMED）提到的关键词（主题词）数量/参照论文提到的关键词数量 × 100%

确定 PUBMED 标引的主题词是按照下面的方法进行的：从 PUBMED 数据库中下载 70 篇高频论文的主题词，按照聚类结果归纳的四类，通过书目信息共现挖掘系统提取主题词，对每一类均按照词频排序，选取排名靠前的 10 个词作为比较对象。

2 结果与分析

2.1 算法提取的关键词和 PUBMED 标引的关键词准确性比较

表 1 列出了算法提取的关键词以及 Authority 值和 PUBMED 标引的主题词以及频次（均取排名前 10 位）。表 2 列出了算法、PUBMED 和论文提到的关键词比较（表 2 中相同的词出现在同一行）。对于第一类，算法中有 3 个词：Tocopherol，Posttraumatic 和 Lipid Peroxidation 与对照论文相同；主题词中有 4 个词：Methylprednisolone，Methylprednisolone Hemisuccinate，Spinal Cord Injuries 和 Lipid Peroxides 与对照论文相同。这里，Spinal Cord Injuries 可以认为与 Posttraumatic 是相同的。对于第二类，算法中有 5 个词与对照论文相同，分别为：Angiotensin，Adrenaline，Nimodipine，Posttraumatic 和 Spinal Cord Blood Flow；主题词中也有 5 个词相同，分别为：Methylprednisolone，Pregnatrienes，Lipid Peroxides，Lipid Peroxides 和 Spinal Cord。其中算法中的前 3 个和主题词中的中间 3 个都是治疗用的药物或激素，所以可以认为与对照论文中的 Thyrotropin-releasing hormone，Naloxone 和 Dexamethasone 相同。对于第三类，算法中只有一个 Axon 词相同；主题词中有 Nerve Regeneration 和 Axons 相同。对于第四类，算法中虽然表面上看没有相同的词，可是 physical，Activities of Daily Living 是指身体、锻炼、能力、每日活动，都可以和康复、恢复联系起来，所以看作与对照论文中的 Recovery，Rehabilitation 相同，认为有两个词相同；主题词中没有词与对照论文相同。分别按照各类计算准确率：

第一类：算法准确率 = 3/10 = 30%，PUBMED 准确率 = 4/10 = 40%

第二类：算法准确率 = 5/10 = 50%，PUBMED 准确率 = 5/10 = 50%

第三类：算法准确率 = 1/10 = 10%，PUBMED 准确率 = 2/10 = 20%

第四类：算法准确率 = 2/10 = 20%，PUBMED 准确率 = 0/10 = 0

计算结果表明，对于每一类算法提取出的关键词和 PUBMED 标引的主题词的准确率相差不大。

2.2 算法提取的关键词和 PUBMED 标引的关键词区分各类别的能力比较

表 3 列出了算法提取的关键词和 PUBMED 标引的关键词。从表 3 中可以清楚的发现，每一类中的词与其他类中的词均不相同。这说明，HITS 算法提取出的词能够区分各类别的特点。而对于 PUBMED 标引的关键词，排名前 10 位的各组词里有 6 个词是四类中都相同的，有一个词 Pregnatrienes 是三类中都相同的，有两个词 Axons 和 Dose-Response Relationship 是两类中都相同的。这样，第一类中只有 Double-Blind Method 和 Methylprednisolone Hemisuccinate 可用来描述第一类；第二类中只有 Follow-Up Studies，Injections，Intravenous 和 Regional Blood Flow 可用来描述；第三类中只有 Nerve Regeneration 可用来描述；第四类中有 Paraplegia，Middle Aged 和 Quadriplegia 可用于描述。除了第三类的 Nerve Regeneration 能够代表该类以外，另外三类都不能与其他两类更好地区分。

表 1　算法提取和 PUBMED 标引的关键词列表

第一类				第二类				第三类				第四类			
算法提取的关键词		PUBMED 标引的主题词		算法提取的关键词		PUBMED 标引的主题词		算法提取的关键词		PUBMED 标引的主题词		算法提取的关键词		PUBMED 标引的主题词	
关键词	Authority 值	主题词	词频	关键词	Authority 值	主题词	词频	关键词	Authority 值	主题词	词频	关键词	Authority 值	主题词	词频
lipid peroxidation	0. 99400497	Spinal Cord Injuries	24	mean systemic arterial blood pressure	0. 994004975	Spinal Cord Injuries	26	spinal cord injury	0. 994004958	Spinal Cord Injuries	38	heart rate reserve	0. 99400497	Spinal Cord Injuries	51
posttraumatic	0. 29084309	Methylprednisolone	17	pressor	0. 994004975	Methylprednisolone	17	channel	0. 994004958	Spinal Cord	22	physical	0. 84970341	Spinal Cord	22
degeneratsion	0. 2516396	Spinal Cord	9	adrenaline	0. 71743934	Spinal Cord	12	axon	0. 389781027	Methylprednisolone	20	strain	0. 77254757	Methylprednisolone	20
tocopherol	0. 2516396	Lipid Peroxides	9	nimodipine	0. 676074154	Lipid Peroxides	9	methylprednisolone sodium succinate	0. 347622011	Time Factors	12	subject	0. 2567272	Time Factors	18
calcium	0. 20069564	Time Factors	8	spinal cord blood flow	0. 675934268	Time Factors	8	myelin	0. 334654358	Lipid Peroxides	10	task	0. 23935762	Lipid Peroxides	10
role	0. 20068024	Ischemia	4	agent	0. 527282785	Ischemia	7	purify	0. 251256751	Ischemia	8	transfer	0. 23516024	Middle Aged	10
progress	0. 1767051	Methylprednisolone Hemisuccinate	4	angiotensin	0. 287575736	Regional Blood Flow	5	rostral	0. 251256751	Axons	8	activities of daily living	0. 23245641	Paraplegia	9
support	0. 17508742	Pregnatrienes	4	infuse	0. 287575736	Pregnatrienes	4	extend	0. 251256751	Nerve Regeneration	7	heart	0. 15530055	Ischemia	8
ascorbate	0. 16804224	Double-Blind	4	investigate	0. 287575736	Follow-Up Studies	4	fewer	0. 251256751	Pregnatrienes	6	wash	0. 15530055	Axons	8
dalpha	0. 16804224	Dose-Response Relationship, Drug	4	posttraumatic	0. 222222288	Injections, Intravenous	4	prelabel	0. 251256751	Dose-Response Relationship, Drug	6	ascend	0. 15530055	Quadriplegia	8

表 2　算法、PUBMED 和论文提到的关键词比较

第一类			第二类			第三类			第四类		
关键词	主题词	论文	关键词	主题词	论文	关键词	主题词	论文	关键词	主题词	论文
tocopherol		tocopherol			calcium channel blocker		Nerve Regeneration	regenera tion	physical		recovery, rehabilitation
	Methylprednisolone	methylprednisolone	angiotensin	Methylprednisolone	thyrotropinreleasing hormone	axon	Axons	axon	activities of daily living		
	Methylprednisolone Hemisuccinate		adrenaline	Pregnatrienes	naloxone	spinal cord injury	Spinal Cord Injuries		heart rate reserve	Spinal Cord Injuries	
posttraumati	Spinal Cord Injuries	posttraumatic	nimodipine	Lipid Peroxides	dexamethasone	channel	Spinal Cord	strain	Spinal Cord		
lipid peroxidation	Lipid Peroxides	U74006f (lipid peroxidation)	posttraumatic	Lipid Peroxides	posttraumatic	methylprednisolone sodium succinate	Methylprednisolone		subject	Methylprednisolone	
degeneratsio	Spinal Cord		spinal cord blood flow	Spinal Cord	spinal cord	myelin	Time Factors		task	Time Factors	
calcium	Time Factors		agent	Ischemia		purify	Lipid Peroxides		transfer	Lipid Peroxides	
role	Ischemia		infuse	Regional Blood Flow		rostral	Ischemia		heart	Middle Aged	
progress	Pregnatrienes		investigate	Follow-Up Studies extend	Pregnatrienes		wash	Paraplegia			
support	Double-Blind Method	mean systemic arterial blood pressure	Injections, Intravenous	fewer	Dose-Response Relationship, Drug	ascend	Ischemia				
ascorbate	Dose-Response Relationship, Drug		pressor	Time Factors		prelabel				Axons	
dalpha										Quadriplegia	

表3　算法提取的关键词和 PUBMED 标引的关键词区分各类别的能力比较

算法提取的关键词				PUBMED 标引的主题词			
第一类	第二类	第三类	第四类	第一类	第二类	第三类	第四类
lipid peroxidation	mean systemic arterial blood pressure	spinal cord injury	heart rate reserve	Spinal Cord Injuries	Spinal Cord Injuries	Spinal Cord Injuries	Spinal Cord Injuries
posttraumatic	pressor	channel	physical	Spinal Cord	Spinal Cord	Spinal Cord	Spinal Cord
degeneratsion	adrenaline	axon	strain	Methylprednisolone	Methylprednisolone	Methylprednisolone	Methylprednisolone
tocopherol	nimodipine	methylprednisolone sodium succinate	subject	Time Factors	Time Factors	Time Factors	Time Factors
calcium	spinal cord blood flow	myelin	task	Lipid Peroxides	Lipid Peroxides	Lipid Peroxides	Lipid Peroxides
role	agent	purify	transfer	Ischemia	Ischemia	Ischemia	Ischemia
progress	angiotensin	rostral	activities of daily living	Pregnatrienes	Pregnatrienes	Pregnatrienes	Paraplegia
support	infuse	extend	heart	Double-Blind Method	Follow-Up Studies	Nerve Regeneration	Middle Aged
ascorbate	investigate	fewer	wash	Dose-Response Relationship，Drug	Injections，Intravenous	Axons	Axons
dalpha	posttraumatic	prelabel	ascend	Methylprednisolone Hemisuccinate	Regional Blood Flow	Dose-Response Relationship，Drug	Quadriplegia

3　分析

3.1　对算法提取结果的分析

将 HITS 算法提取的关键词和 PUBMED 标引的主题词分别与对照论文中所描述的类别信息相比较。由于对照论文中为了简明扼要，对类别的描述是通过一句话概括的，所以在还原时会得到较少数量的描述词，因而与二者比较时相同的词自然会减少。因此分别比较每一类，准确率都不高。对于第四类，PUBMED 标引的排名靠前的主题词没有能够说明脊髓损伤的恢复研究的，不能达到要求。相比较之下，HITS 算法提取的关键词既有较高的准确率，又有很强的代表性。因此，HITS 算法提取的特征词是比较理想的。

通过算法提取出来的关键词数量很多，选取排名靠前的一定数量的词作为特征词，归纳起来类似摘要一样，而摘要中的每个词分量都很重，都能代表这一类说明的含义，这些词之间具有语义上的关联。相比较之下，如果某一类中含有 20 多篇摘要，逐一阅读既需要时间，又需要精力，那么通过计算机自动提取出一个全是关键词的摘要，阅读起来自然要比阅读 20 篇摘要更省时省力。

3.2　对 HITS 算法可行性的分析

HITS 算法的主要思想是：网页的重要程度是与所查询的主题相关的。相对于不同主题，同一网页的重要程度也是不同的。可以引申为：一个单词如果在中心论文里出现许多次，那么它具有较高的权威性；而如果一篇论文中包含许多的权威单词，则它具有较高的中心性。因此，可以推断出，含有许多“权威”单词的论文一定是“中心”的核心的论文，而在许多“中心”的核心论文中所出现的单词一定是“权威”的关键词。对于聚类分析结果来说，每一类的文章都是相似的，主题也是相关的。因此，将 HITS 算法用于聚类结果的关键词提取是非常适合的。

4　结束语

HITS 算法用于聚类结果的特征提取是可以实现的，其能有效地发现关键词，很好地区分每一类的特征，提取出聚类结果中每一类的特征词，客观地描述聚类分析结果，使科研人员能够更准确更容易地分析聚类结果，为进一步研究提供服务。□

参考文献

[1] KLEINBERG J M. Authoritative sources in a hyperlinked environment [C] //Paper presented at: 9th Annual ACM-SIAM Symposium on Discrete Algorithms，1998：668-677.

[2] SALTON G，CLEMENT T Y. On the construction of effective vocabularies for information retrieval [C] // Proceedings of the 1973 Meeting on Programming Languages and Information Retrieval. ACM New York，NY，USA，1973：48-60.

[3] 赵悦阳．专题文献的同被引聚类分析在表现学科专业发展历史的可靠性评价 [J]．情报学报，2005，24（4）.

[4] 章成志．自动标引研究的回顾与展望 [J]．现代图书情报技术，2007（11）：33-39.

作者简介：赵悦阳，女，硕士。

崔雷，男，教授。

收稿日期：2012－09－12