

学术文献的文体特征及其检索意义

——计算机科学文献与相关新闻报道文体的比较研究*

邹永利 冯静娴 郑 荟

摘 要 计算机科学学术文献具有独特的文体特征,且部分特征能够用于计算机科学网络中文学术文献的自动识别与检索,提高计算机科学学术文献的相对检准率。本文通过建立计算机学术文献语料库与 IT 新闻语料库,对计算机科学学术文献和 IT 新闻在特有表述、平均句长、中西文比例三方面的文体特征进行提取与加权,应用于基于百度的网络中文学术检索系统(NSIRS)中。结果显示,本研究提取并设定的文体特征能有效提高 NSIRS 对计算机科学学术文献的检索效率,从而证明发掘并应用特定学科文献独特的文体特征的意义。图 2。表 4。参考文献 14。

关键词 文体特征 计算机科学 学术文献 IT 新闻 NSIRS 系统

分类号 G354

Stylistic Characteristics and Retrieval of Chinese Academic Literatures: A Comparative Study on the Stylistic Characteristics between Academic Papers and Press Reports of Computer Science

Zou Yongli, Feng JingXian & Zheng Hui

ABSTRACT Computer science academic documents have distinctive stylistic features that can be explored to facilitate the automatic identification and retrieval of Chinese computer science academic papers on the web. This paper establishes a computer academic literature corpus and an IT news corpus to find the different typical expressions, average length of sentences, ratio of Chinese characters to Roman alphabets, and gives them different weights. Finally, this paper applies the results to Baidu-based NSIRS system. Precision evaluation over the NSIRS shows significant advantage of our approach over previous study using the same system. 2 figs. 4 tabs. 14 refs.

KEY WORDS Stylistic characteristics. Computer science. Academic literature. IT news. NSIRS system.

互联网上存在大量 HTML 格式的免费中文学术文献,但缺失实用的检索工具或系统。为了在海量互联网信息中自动识别网络中文学术文献,排除干扰文献,我们借鉴日本学者安形辉、石田荣美、池内淳^[1-2]等人的研究思路,采用文体计量学方法,对比学术文献与新闻报道、网络日志之间的文体差

异,开发出基于中文学术文献文体特征的网页检索系统(NSIRS)^[3],验证了某些文体特征对于提高中文网络学术文献检准率的影响。研究发现,中文学术文献文体特征存在学科差异^[4],但尚未就这些差异对特定学科文献检索效率的影响作深入研究。计算机科学是与互联网联系最为紧密的学科,其网

* 本文系国家社会科学基金项目“网络中文学术文献的自动识别与检索研究——基于学术文献文体、链接及图文相关度的研究与系统开发”(编号:10BTQ049)的研究成果之一。

通讯作者:邹永利, Email: isszyl@mail.sysu.edu.cn.

络文献具有数量多、时效性强等特点,挖掘其特有的文体特征不仅能够提高该学科网络学术文献的检索效率,而且能为信息检索对不同学科文体差异的利用提供依据。

1 文体概念及相关研究

文体是人们以特定的方式,在不同的交际环境下,同特定的交际对象,围绕特定目的进行语言交际时所使用的不同语言形式^[5]。根据语篇的表层语言语法描写和深层语篇语义结构,可以对应地将文体特征分为语言描述特征和语篇结构特征^[4],学术文献的文体主要包括层次结构、逻辑结构、特有表述或典型句型、所含语种数量及比例、平均句长等。特有表述是文献中非主题性的高频语词,分为结构性表述与修辞性表述两类^[6]。

现有文体特征的研究主要针对英文文献、新闻报道、广告和文学作品,对学术文献文体特征的研究较为罕见。国内对于计算机科学学术文献文体研究主要从语言学角度出发,研究文体特征应用于写作指导,且以英文居多。高时阔等人^[7]利用艺术分析的方法,分析科技期刊学术论文的文体结构——符号结构、层次结构和认知结构的特征,即文字直白、专业,图表简明、写实,层次结构规范等。袁月杨^[8]通过对比本族语为英语和本族语为汉语的作者所著计算机科学科技文献的英语摘要(总计120条),运用认知语言学中的原型理论统计语料库中标题和摘要的平均长度、单个句子平均长度、常用词汇和词频等,发现中国作者所著的计算机科技论文标题和摘要僵化,用词简单重复,其研究目的是提高科技论文英语标题的选择以及摘要的写作质量。笔者^[3,4,6,9-11]曾通过建立语料库,尝试系统地探讨学术文献整体的文体特征——特有表述、平均句长和中西文比例等,以及学术文献文首文尾的特征^[11],但没有针对某一具体学科文献的文体特征进行专门研究。

2 研究方法

本文选取计算机科学学术文献作为研究对

象,建立计算机学术文献语料库及其主要干扰文献IT新闻语料库,采取词频和文档频率相结合的方法选取计算机科学学术文献和相关新闻的特有表述、平均句长和中西文比例三方面的文体特征,同时采用类间和类内信息分布熵的特征加权法,将研究结果应用于NSIRS系统。最后,我们利用此前实验验证中^[4]选取的属于计算机科学的三个词语作为检索词进行效果测试,分别运用于百度、原NSIRS系统和本文研究改良后的NSIRS系统,计算各自的检准率,考查其变化情况。

语料主要来自两个方面:首先是复旦大学李荣陆建立的中文分类语料库。我们将其中computer(计算机)类中的文献按照学术、新闻、其他等进行人工分类,选取学术文献1000篇作为研究语料库;其次,参照库选取搜狗文本分类语料库中的IT类共2000篇,该类数据为搜狐新闻网站IT类收录,作为新闻语料库。

3 文体特征的提取

学术文献与新闻报道由于各自写作目的、读者对象不同,在特有表述、平均句子长度和中西文字符比例三方面,呈现显著不同的统计特征。

3.1 特有表述

文献的特有表述可分为结构性表述和修辞性表述^[12]。学术文献常见的结构性表述有绪论、前言/前沿、引言、序言、导论、研究背景、研究目的、研究方法、研究意义、参考文献、关键词、摘要、中图分类号/中图法、文献标识码、文章编号、责任编辑等;修辞性表述通常有在什么领域……、本文拟探讨……、本文旨在……、作者希望……、对……的研究分析表明等。

新闻的特有表述一般有记者、编辑、摄/摄影、稿源、文章来源、独家报道、发言人、目击者、据悉、声称、声明、独家声明、本报讯、专电、专稿等;有些表述类似学术文献中的修辞性表述,如据……报道、……摄(摄影)、某网某月某日电、……网(社)发、记者采访到……、发表……言论、……日报某月

某日报道、网友建议……、……回应等。

本文首先使用 Myzicifreq 词频统计软件获取高频词,但由于其在具体词频数统计方面存在误差,又使用 Microsoft Word 的查找功能统计出高频词的准确词频。基于上述词的词频统计,并统计其社会新闻语料库中出现的频率,最终特有表述的选取要求满足两方面的条件:一是仅出现在学术文献中的特有表述,二是在学术文献中出现频率较高

而在新闻中出现频率较低。本文继续选用以往研究^[3]选取时的统计方法:特有表述出现次数除以文档数,即表述在每篇文献平均出现的次数,规定特有表述在新闻中出现频率不高于 5%,当特有表述出现在新闻中时,其在学术文献中的频率不低于 15%。将所有学术文献特有表述进行统计,最终筛选出图 1 列出的特有表述,结果与以往研究^[3]以同样方法获得的学术文献的特有表述有所不同^①。

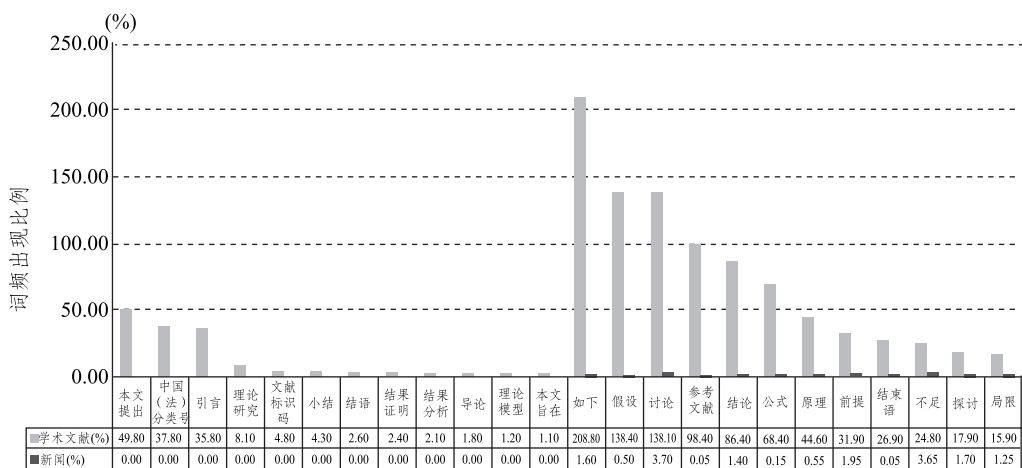


图 1 计算机科学学术文献特有表述总体分布图

选取网络新闻的特有表述,满足其在新闻中出现频率高于在学术文献中出现频率的条件。最终获得的特有表述如图 2 所示,结果也与以往研究^[3]获取的特有表述有所不同^②。

新闻语料库中出现频率最高的特有表述是“本报讯”,且这一表述不存在于学术文献语料库中,是很好的新闻特有表述,可运用于系统作为滤词。尽管“记者”同样大量存在于新闻之中,但在学术文献中仍占有一定比例。其次是出现频率较低的“据悉”“业内人士”“媒体报道”“发言人”,其出现频率高于 2%,但不存在于学术文献中,因而也可以作为新闻特有表述成为系统滤词。

3.2 平均句长

研究^[3]表明,学术文献和新闻在平均句长上存在很大差异,将其作为检索学术文献的判断标准,可以达到过滤相关新闻的效果。句长的统计方式如下:一是以标点符号“。”“?”“!”作为一个句子结束的标志,统计语料库中包含的句子数量,然后再以总字符数除以句子数量,得出计算机科学学术文献和新闻报道的平均句长;二是将“,”“;”“:”也作为句子分隔符,排除有连接意味的短语,如“其中”“研究表明”等这类可单独成句的衔接短语,发掘学术文献和新闻在句长方面更具体的差异。利用 Microsoft Word 查找功能分别统计两个语料

① 参考文献[3]选取的学术文献特有表述有:文献标识码、文章编号、中图分类号、引言、结束语、理论研究、导论、参考文献、摘要、公式、结论、实验、如下、前提、综述、探讨。

② 参考文献[3]选取新闻特有表述时参考的特有表述有:免责声明、摄影、热词搜索、记者、网络编辑、本报讯、据报道、目击者、稿源、发言人、声称、独家报道、据悉、声明、发布时间、相关人士、消息人士、文章来源。

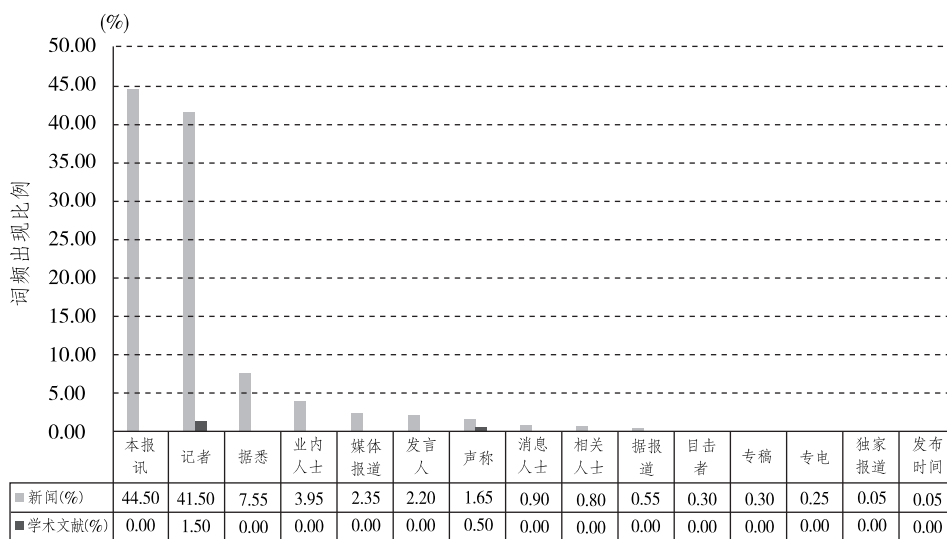


图2 IT新闻特有表述总体分布

库中结束符号的数量,算出句子总数,再用中、西文字符总数除以句子总数,得到各语料库的平均句长(见表1)。

表1的统计数据表明,语料库中学术文献与新闻报道在平均句长上存在较为明显的差异,两种计算方式中新闻报道的平均句长均小于学术文献,其中平均句长二的统计数据更清晰地显示二者之间的差异,这是由新闻文体简洁、精炼的特征所导致的。有研究显示,新闻报道要具可读性,关键之一是写短句子,目标是让初中二年级文化程度以上的人都能看得懂^[13],导致新闻报道的平均句长远低于学术文献。

表1 平均句长统计

	计算机学术文献	计算机新闻
平均句长一	63	47
平均句长二	22	17

3.3 中西文比例

中西文字符比例是指文中所含中文字符数与西文字符数的比例,计算方法是:除符号以外的全部中文字数/除符号以外的全部英文字母数。研究^[3]表明,中西文比例越低,属于学术文献的几率越大。统计结果如表2所示。

表2 中西文比例统计

	中文字数	英文字母数	中西文比例
计算机学术文献	3,484,494	2,194,804	1.58761
计算机新闻	1,454,836	118,878	12.23806

表2中的数据清楚地显示,计算机科学学术文献与相关新闻报道的中西文比例差异显著,计算机科学学术文献中西文比例很低,英文大多存在于摘要以及参考文献之中,而计算机科学兴起于国外,中文文献多借鉴外文研究成果,专业术语、数学公式等表述多使用英文字符,因而明显偏低于新闻文

体的中西文比例。

4 权重设定

NSIRS 系统是通过赋予文体特征一定权重,对 Google 检索结果页面中各网页的内容长度权值、文

体特征权值相加,得到各网页的总权值。根据各网页的总权值,在 Google 检索结果页面的基础上对各条检索记录进行重新排序。因此,根据以上统计结果,赋予计算机学术文献各文体特征相应的权值是改进 NSIRS 系统的重要环节^[4]。

4.1 特有表述

颜秀飞^[4]对学术文献特有表述按照词频分为三级加权,未能充分反映特征项对区分各类别学术文献的贡献差异。本文认为不应单单考虑词频,还应考虑文档频率。在学术文献和新闻报道两类文献中分布越不均匀、而在学术文献中分布越均匀的特有表述,权值应该越高。张保富、施化吉等^[14]提出结合特征项的类间和类内信息分布熵的概念,能够更好地反映本文所需的加权要求,其加权公式如式 1:

$$W_{ik}(d) = \frac{tf_{ik}(d) \times \log\left(\frac{N}{n_k} + 0.1\right)}{\sqrt{\sum_{i=1}^n (tf_{ik}(d))^2 \times \log\left(\frac{N}{n_k} + 0.1\right)^2}} \times \alpha(H_{ac}) \times H_{ic} \quad (\text{式 } 1)$$

其中, H_{ac} 为特征项的类间信息分布熵, H_{ic} 为特征项的类内信息分布熵,能反映出特征项在类间和类内的分布比例情况,将二者结合起来能够反映特征项对各类别贡献。NSIRS 将特有表述作为加权依据的算法是:遍历系统处理后的 Google 检索页面返回的当页结果,按照设定的文体特征关键词权值对每条结果中包含的关键词词频进行权值计算,然后叠加得出最终权值。由于 NSIRS 系统已经考虑了词频因素,因而本文在权值设定时未以词频作为设定依据,仅考虑类间信息熵与类内信息熵,权值算法为类间信息熵因子与类内信息熵的乘积(式 2):

$$W_i = \alpha(H_{ac}) \times H_{ic} \quad (\text{式 } 2)$$

表 3 是权值设定结果,由统计特有表述在学术文献和新闻报道中的文档频率、在每篇文档中的出现次数得出。

以往研究表明,在 NSIRS 系统中加入新闻报道的特有表述,对于过滤相关新闻这类干扰文献有显著作用,因此我们根据新闻特有表述词频的统计数据,将仅出现于新闻报道中的高频词“本报讯”赋予负权值-2,将次之的“记者”“业内人士”“媒体

报道”“发言人”等赋予负权值-1。

4.2 文献句长

前文以两种方式统计了计算机科学学术文献与相关新闻报道的句长,二者存在较显著差异,可作为学术文献与新闻的区分标准。由于第二种句长统计方法暂未在 NSIRS 系统中实现,因而权值设定仅参照第一种方法即以“。”“?”“!”作为句子分隔符统计得出的数据。统计结果显示,计算机科学学术文献语料库的平均句长为 63,变化范围是 37~110,而新闻报道的平均句长为 47,与陈亮对《中国经营报》、《21 世纪经济报道》、《经济观察报》平均句长的统计结果基本一致^[13]。平均句长越长,文献为学术文献的可能性越大,因此 NSIRS 系统中平均句长权值的设定为:平均句长为 0~15 的文献权值设为 0,16~30 的文献设为 1,31~47 时设为 2,48~62 时设为 3,63 以上时设为 4。

4.3 中西文比例

统计结果显示,计算机科学学术文献语料库中文献的中西文比例,变化幅度较大,约为 0.3~12.9,平均约为 1.6;而新闻报道的中西文比例平均值约为 12.2,最小值为 3.1。新闻报道中存在完全不包含任何西文字母的情况,而计算机学术文献中不存在这种情形。因此,在 NSIRS 系统中中西文比例权值可设定为:比例值为 0 时权值设为 0,21 以上时设为 0,16~20 时设置为 1,12~15 时设置为 2,3~11 时设置为 3,比例值为 1~2 时权值设为 4。

5 系统验证

我们将本文提取的文体特征权值应用于 NSIRS 系统中,考察检索效果的变化情况。由于 Google 服务器访问不稳定,我们将网页来源换成百度搜索引擎,并基于百度检索结果进行加权、排序、输出,页面总权值依旧为网页内容长度权值和文体特征权值之和。由于本研究是对 NSIRS 系统的扩展和完善,因而在系统验证方面沿用以往使用的“相关性范畴”界定和“前 X 命中记录查准率”^[14]

表3 学术文献特有表述权值设定

	H_{ac}	$\alpha(H_{ac})$	H_{ic}	W_i
本文提出	0	1	5.525972	5.525972
中图(法)分类号	0	1	5.934894	5.934894
引言	0	1	5.853427	5.853427
理论研究	0	1	3.902317	3.902317
文献标识码	0	1	3.871201	3.871201
小结	0	1	3.761200	3.761200
结语	0	1	3.258097	3.258097
结果证明	0	1	2.730114	2.730114
结果分析	0	1	2.912494	2.912494
导论	0	1	2.813355	2.813355
理论模型	0	1	1.098612	1.098612
如下	0.165358731	0.764830556	6.308024	4.824570
假设	0.092557825	0.868366356	2.549596	2.213983
讨论	0.375525165	0.465936613	5.905504	2.751591
参考文献	0.008307572	0.988185159	6.831293	6.750582
结论	0.243584909	0.653579057	5.350154	3.496749
公式	0.077570038	0.889681649	4.739232	4.216408
原理	0.171057603	0.75672575	5.248451	3.971638
前提	0.431860776	0.385817383	4.547435	1.754479
结束语	0.024742911	0.964811193	5.574097	5.377951
不足	0.653759651	0.070237826	4.994987	0.350837
探讨	0.459003296	0.347215905	4.773844	1.657555
局限	0.457551374	0.349280795	4.733257	1.653236

的方法,以观察检索效果的变化情况。

本文选取此前研究^[3]在系统验证时所使用的属于计算机科学范畴的3个检索词(“网格计算”、“数据挖掘”、“自主计算”),将其分别置于百度、以往研究设定的系统和本研究设定的系统中进行检索试验^①,选取前50条命中记录,计算并对比三者的相对查准率。结果如表4所示,其中NSIRS1表

示以往研究设定的NSIRS系统,NSIRS2表示按照本文研究结果设定的NSIRS系统。

表4表明,应用学术文献的文体特征可明显提高百度的学术文献检索效率。在本次实验中,百度的平均检准率是48.61%,NSIRS1的平均检准率是52.89%,比百度学术文献相对检准率提高了4.28%;而按照本文研究结果设定的NSIRS2,进一步提高计

① 检索时间是2012年4月19日至4月21日

表 4 相对检准率

	网格计算 (%)	数据挖掘 (%)	自主计算 (%)
百 度	55.56	40.28	50.00
NSIRS1	61.80	44.10	52.78
NSIRS2	62.15	47.57	55.56

算机科学学术文献的检准率,相比百度平均提高了 6.48%,比以往研究设定的 NSIRS1 平均提高了 2.2%。证明本研究对于计算机科学学术文献检索以及 NSIRS 系统的改进均具有实际意义。

6 结论

本文通过建立语料库的方法,分析了计算机科学学术文献及其主要干扰文献——IT 新闻报道,在特有表述、平均句长、中西文比例三个方面存在的差异,揭示了文体特征用于识别学术文献的可行性以及文体特征所具有的学科差异性。实验部分证明了文体特征能够有效提高计算机科学学术文献的检准率,而且对特定学科领域学术文献文体特征的应用能够更好地提高该学科文献的检准率,

显示分学科研究文体特征的意义所在。此外,本研究改进的加权方法可应用于实现不同语料库研究特有表述加权的自动批量生成,这为扩大语料库研究并实现自动加权提供了一种较为科学可行的方法。

尽管本研究扩大了样本量,但限于人工提取效率的局限,样本量仍不充足。此外,NSIRS 系统检索生成的分析对象文档存在内容不完整、格式不规范等缺陷,从而限制了文体特征的应用,致使本研究的应用效果未能完全体现。

今后我们会在理论上进行更深入、系统的研究,补充,还将尝试更全面地检验其实践意义,改进加权方法,在实现 NSIRS 系统对于单一文体特征效果测试的基础上,尝试找出各项文体特征之间存在的关系,寻求三者的最优组合。

参考文献

- [1] 安形輝,池内淳,石田栄美.日本語学術論文 PDF ファイルの自動判定[J].Library and Information Science, 2006,56(2): 43-63.(Teru AGATA, Atsushi IKEUCHI, Emi ISHIDA. Automatic identification of academic articles in Japanese PDF files [J]. Library and Information Science. 2006,56(2): 43-63.)
- [2] 池内淳,安形輝,石田栄美.プーリング手法を用いた学術論文の自動判別実験[C]//情報処理学会研究報告. 東京: 日本情報処理学会,2007,34;33-40.(Atsushi IKEUCHI, Teru AGATA, Emi ISHIDA. Automatic detection for academic articles using pooling method[C].IPSJ SIG Technical Report. Tokyo: Information Processing Society of Japan, 2007,34;33-40.)
- [3] 邹永利,颜秀飞.文体特征与网络中文学术文献的自动识别——学术文献文体与新闻报道文体的比较研究[J].情报学报,2012,31(2):160-165.(Zou Yongli, Yan Xiufei. Identification and retrieval of Chinese academic papers from the web: A comparative study on the stylistic characteristics between academic papers and news reports[J].Journal of the China Society for Scientific and Technical Information, 2012,31(2):160-165.)
- [4] 颜秀飞.网络学术文献文体特征的获取与检索研究[D].广州:中山大学,2010.(Yan Xiufei. Application of stylistic characteristics in acquisition and retrieval of Chinese academic articles on the web[D].Guangzhou: Sun Yat-sen University,2010.)
- [5] 张武德.科技英语文体的主要特点[J].湖北函授大学学报,2004,17(4): 56-57.(Zhang Wude. Main stylistic characteristics of technical English[J].Journal of Hubei Correspondence University,2004,17(4):56-57.)

- [6] 邹永利. 学术文献的非主题特征及其意义[J]. 中国图书馆学报, 2010, 37(3): 100-107. (Zou Yongli. Non-topical attributes of academic documents and the implication[J]. Journal of Library Science in China, 2010, 37(3): 100-107.)
- [7] 高时阔, 宇文高峰. 科技期刊学术论文文体结构特点分析[J]. 中国科技期刊研究, 2004, 15(1): 19-21. (Gao Shikuo, Yuwen Gaofeng. Analysis on characteristics of sci-tech papers' literary structure[J]. Chinese Journal of Scientific and Technical Periodicals, 2004, 15(1): 19-21)
- [8] 袁月杨. 基于原型理论的计算机科学科技论文英文标题及摘要的对比研究[D]. 长沙: 国防科学技术大学, 2009. (Yuan Yueyang. A comparative study of titles and abstracts of research papers in computer science: A prototypical perspective[D]. Changsha: National University of Defence Technology, 2009.)
- [9] 何侃. 基于文体特征的网络学术文献检索研究[D]. 广州: 中山大学, 2007. (He Kan. The retrieval of online academic papers based on stylistic characteristics[D]. Guangzhou: Sun Yat-sen University, 2007.)
- [10] 邹永利, 何侃, 徐健. 文体特征在网络学术文献检索中的意义与应用[J]. 情报理论与实践, 2008, 31(4): 594-597. (Zou Yongli, He Kan, Xu Jian. The implication and application of stylistic characteristics in retrieving academic literatures from the web[J]. Information Studies: Theory & Application, 2008, 31(4): 594-597)
- [11] 林吉鹏. 学术论文文首、文尾的文体特征研究[D]. 广州: 中山大学, 2010. (Lin Jipeng. A study on stylistic characteristics of the beginning and ending of academic papers[D]. Guangzhou: Sun Yat-sen University, 2010.)
- [12] Kircz J G. Rhetorical structure of scientific articles; The case for argumentational analysis in information retrieval[J]. Journal of Documentation, 1991, 47(4): 354-372.
- [13] 陈亮. 专业性的迷失——初探中国财经类报纸报道方式的误区[EB/OL]. [2011-01-17]. http://www.cnr.cn/zggb/lt/200801/t20080130_504692882.html. (Chen Liang. The lost of expertise: A study on misunderstandings in reporting Chinese financial newspaper[EB/OL]. [2011-01-17]. http://www.cnr.cn/zggb/lt/200801/t20080130_504692882.html.)
- [14] 张保富, 施化吉, 马素琴. 基于 TFIDF 文本特征加权方法的改进研究[J]. 计算机应用与软件, 2011, 28(2): 17-20. (Zhang Baofu, Shi Huaji, Ma Suqin. An improved text feature weighting algorithm based on TFIDF[J]. Computer Applications and Software, 2011, 28(2): 17-20.)

邹永利 中山大学资讯管理学院教授。

通讯地址: 广州大学城外环东路 132 号中山大学北学院楼 B 栋资讯管理学院。邮编: 510006。

冯静娴 中山大学资讯管理学院硕士研究生。通讯地址同上。

郑 荟 中山大学软件学院硕士研究生。

通讯地址: 广州大学城外环东路 132 号中山大学软件学院。邮编: 510006。

(收稿日期: 2013-09-01)