

基于引文内容分析的高被引论文主题识别研究

祝青松 冷伏海

摘 要 基于被引次数的引文分析无法直接揭示论文的研究内容,利用关键词或从标题、摘要和全文中抽取的主题词很难客观反映论文的被引原因。本文以碳纳米管纤维研究领域的高被引论文为研究对象进行引文内容抽取和主题识别,经人工判读验证:基于引文内容分析的高被引论文识别的核心主题能够较好地揭示高被引论文的被引原因(引用动机),而且与论文的研究内容相符合;与基于全文、基于标题和摘要的主题识别相比,在引文内容分析基础上识别的主题具有更好的主题代表性,能够有效揭示被引文献的研究内容,是对原文相关信息的重要补充。本文的实验表明基于引文内容分析的高被引论文主题识别是可行而且有效的。图4。表4。参考文献31。

关键词 引文内容分析 主题识别 高被引论文 引用动机

分类号 G350

Topic Identification of Highly Cited Papers Based on Citation Content Analysis

Zhu Qingsong & Leng Fuhai

ABSTRACT Citation analysis based on citation frequency fails to directly reveal research contents of papers, neither can it objectively reflect the reason for citation with keywords or topic words extracted from titles, abstracts and full-texts. Taking highly-cited papers of carbon nanotube fiber field as examples, this paper extracts citation contents and identifies the topics. Through human interpretation, it verifies that the core topics of identifying highly-cited papers based on citation content analysis can better reveal the reason for citation (i. e., motivation for citation) of highly-cited papers and accord with research contents of papers. Compared with the topic identification based on titles, abstracts and full-texts, the topics identified on the basis of citation content analysis have better representativeness and can effectively reveal research contents of cited papers, and are important supplement to related information in original texts. The experiment results of this paper prove the feasibility and validity of the topic identification through citation content analysis on highly cited papers. 4 figs. 4 tabs. 31 refs.

KEY WORDS Citation content analysis. Topic identification. Highly cited papers. Citation motivation.

1 引言

高被引论文指在某个统计时间段内,被引用次数排在学科前列的论文,在一定程度上代表了学科的研究进展,具有重要的研究意义。汤森路透集团^[1]文献评价分析工具 ESI 中将高被引论文(Most Cited Papers)定义为过去十年被引用次数排在各学

科前1%的论文。一方面,科技论文的被引用次数及其改进指标如影响因子^[2]、h 指数^[3]等,作为重要的计量学指标,已被广泛应用于研究水平测度、科研绩效评价、学术期刊评价等方面^[4-5];另一方面,基于科技论文被引用次数的引文分析及共引分析、耦合分析等作为重要的情报研究方法,也广泛应用于学科结构研究、领域研究前沿和热点探测等领域^[6-7]。

科技论文的引用行为呈现出高度的复杂性,如 Garfield^[8] 在 1964 年提出 15 种引用原因,且被引用次数无法揭示出作者的引用动机。以被引用次数为基础的传统引文分析将所有的引文同等看待,施引文献和被引文献之间的关联性也通常不加以区分。随着全文本文献可获取性的不断提高和文本挖掘技术的持续发展,引文类型识别将引用动机的研究推进到一个新阶段^[9]。引文类型识别是通过施引文献引用被引文献的文本内容进行分析来对引文进行分类,如 Pham^[10] 将引文分为基础、支持、局限和比较,Le^[11] 将引文分为基于被引文献、被引文献一部分、支持施引文献、指出被引文献的问题或差距、比较当前工作和其他类型。引文类型识别主要从引用功能和观点倾向两种角度来对引文进行分类。

为了进一步揭示高被引论文的研究内容,通常利用其关键词、主题词(标题、摘要或全文抽取所得)来表征高被引论文的研究主题,如侯跃芳等^[12] 将引文共引聚类与内容词分析法相结合揭示专题发展,通过引文共引聚类描述妊娠糖尿病专题研究的发展历史,通过高频引文的重要来源文献的内容词分析描述该专题研究现状。然而,高被引论文的主题词只能反映该论文自身的研究内容,不能揭示其被引的原因和内容,因而需要深入施引文献的全文来挖掘其引用的被引文献的内容,即引文内容,在引文内容分析的基础上抽取代表被引文献被引原因的主题词。另外,本文中的主题 Topic 指论文中用于表征相关研究内容的词汇或短语,相对于一般具有严格规范控制和复杂概念体系的主题 Subject(如受控词表和本体等),Topic 更符合从引文内容抽取相关词汇或短语来表征高被引论文主题的研究方法。

科技论文中大部分的创新并不是完全创新,而是在以往创新基础上的再创新,既有继承性,又有变化性。科技论文的继承性通过引用和被引用来实现,并以参考文献的形式出现在科技论文的正文后面,在正文中以特定的形式进行标记,本文称为引用标记,如 3, [19, 24], (15 - 17) 或 (Teufel 1999) 等。同时,将引用标记所在的句子称为引用

句,与引用句在内容上相关联的上下文句子称为引用上下文。引文内容分析指对引用句或引文上下文的分析,引用句是施引文献和被引文献的直接关联,引文上下文蕴含更加丰富的语义信息,本文中引文内容分析指对从引用句中抽取的引文内容进行分析。

引文内容分析是对传统以被引用次数为基础的引文分析的重要补充,能较好地揭示被引文献和施引文献之间在语义内容上的关联,基于引文内容分析的论文主题识别从施引文献的角度出发,识别的主题词能更好地表征被引文献的研究主题和主要贡献,有利于共引、耦合等进一步分析中的语义揭示和内容挖掘。

2 相关研究

Small^[13] 将引文内容分析分为两种:①引文上下文分析(citation context analysis),主要是面向引用功能或观点倾向的引文类型识别;②引文上下文的内容分析(content analysis of citation contexts),主要是面向主题词或短语的语义内容挖掘。第一种更多关注引用功能或观点倾向的分类,重视外部特征,忽略引文的内容分析,在一定程度上仍然是外在层面的分析;第二种是对第一种方法的重要补充,重视内部特征,深入语义内容进行分析,有更好的应用价值。Ding^[14] 认为基于内容的引文分析是下一代引文分析的方向,并将其分为两个层面:一个是语法层面,指引文分布在文献中的不同语法结构中(出现在文中不同章节位置);另一个是语义层面,指引文具有不同的语义贡献(比如重要或不太重要的贡献、肯定或否定型贡献)。

引文内容分析的早期研究主要是人工对引文文本内容进行判读和总结,如 Small^[15] 将引用内容作为观点表达的概念符号,认为将共被引聚类和引文内容分析结合起来能更好地揭示研究领域的知识基础。Small^[16] 利用这种方法对重组 DNA 领域进行了分析,首先利用共被引聚类方法追溯重组 DNA 领域的演化历史,然后利用引文内容分析揭示聚类之间主题的变化,并将其引申到共被引内容

分析,进一步揭示文献概念之间的关系。该方法的关键是用引文文本内容中出现频次最高的词或短语来表示引文,将引文标签化,在一定程度上对演变进程有了更好的解释。

随着文本挖掘技术的提升以及全文本获取的可行性,对文献全文的挖掘和分析越来越多,引文内容分析也在其中。Nakov 等^[17]指出引文内容分析的一系列潜在应用,如可比语料库构建、实体识别、关系抽取、自动文摘、同义词识别和消歧、文献检索等,这些应用都是建立在对引文内容分析的基础之上,其出发点都是引文内容相比被引文献的摘要和全文等包含更丰富的语义信息。Elkiss 等^[18]通过多个实验发现引文内容和被引文献的摘要有一定程度的重合,但是针对同一篇被引文献的施引文献往往关注被引文献的不同方面,不会对其贡献进行完全的描述,并且发现引文内容比摘要具有更好的统一性,包含摘要所不具有的额外语义信息,是对摘要的重要补充。

目前,引文内容分析的研究主要集中于主题识别、自动文摘、信息检索、文本分类和聚类等方面。①主题识别:引文分布在文献的不同位置,引文位置相邻文本的主题词在一定程度上表示了引文的主题分布。文献可以用很多主题词来表示,但是引文并不跟文献所有的主题词都相关,通过引文分布可以增强引文信息,提高引文分析的精度。如 Liu 等^[19]在此假设基础上使用有指导主题模型算法 LLDA 来表示文献和引文主题的分布。②自动文摘:引文内容从施引文献的角度反映其所起的作用,通过对引文内容的抽取和分析,可用于文摘的自动生成。如 Teufel^[20]认为引文中可能包含很多主观内容,并提出 Argumentative Zoning 的方法,利用这些内容来生成摘要。Mohammad 等^[21]阐述了引文内容在多文档自动文摘方面的优势。Qazvinian 等^[22]提出一个基于图的自动文摘模型 C-LexRank,主要利用引文内容来生成技术文摘。③信息检索:利用引文内容可以很好地提高检索结果的相关性。如 Bradshaw^[23]提出的文档索引模型 (Reference Directed Indexing, RDI),利用引用句中

抽取的词语对文档进行标引,并用于检索系统,检索结果的相关性明显改善。Ritchie 等^[24]通过实验证明利用引文内容中抽取的词语可提高信息检索的性能。④文本分类和聚类:文本分类主要是利用引文内容中包含的很多原文中未出现的重要词语来对文献进行标引,如 Aljaber 等^[25]利用引文内容来提高生物医学论文的 MeSH 分类效果。文本聚类同样是利用引文内容中包含的同义词、上下位词等相关词汇来提高聚类效果,如 Aljaber 等^[26]证明结合使用引文内容抽取的主题词和原文全文抽取的主题词可以较好地识别研究主题并用于文献的聚类。

综上所述,引文内容是施引文献对被引文献知识创新内容的重新组织,揭示了被引文献对后续研究的主要贡献,相比于被引文献摘要和全文能够提供更加客观和丰富的语义信息,并且引文内容可以通过主题词的形式进行表征,可以有效提高信息检索等实际应用的效果。

3 研究设计

3.1 数据来源

本文以“碳纳米管纤维”研究领域为例,选取该领域的高被引论文作为研究对象进行引文内容分析和主题识别研究,同时选取 ISI Web of Science 数据库中的子库 SCI-EXPANDED、SSCI、CP-CI-S、CCR-EXPANDED、IC 来构建领域数据集,检索式为 $\{ \text{TI} = ((\text{"carbon nanotube *"} \text{ or } \text{"carbon-nanotube *"} \text{ or } \text{"carbon nano-tube *"} \text{ or } \text{CNT} \text{ or } \text{SWNT} \text{ or } \text{MWNT} \text{ or } \text{DWNT} \text{ or } \text{SWCNT} \text{ or } \text{MWCNT} \text{ or } \text{DWCNT}) \text{ and } (\text{fiber * } \text{ or } \text{fibre * } \text{ or } \text{yarn * })) \}$, 检索时间为 2013 年 3 月 11 日,文献类型限定为 Article 和 Proceedings Paper,共返回有效检索结果 991 条,构成实证研究的数据集。该领域的科技论文分布在 1995 年到 2013 年,而且逐年增加,呈现出良好的发展态势(见图 1)。2013 年出现急剧下降的原因是数据不全,因而不能代表领域的发展趋势。

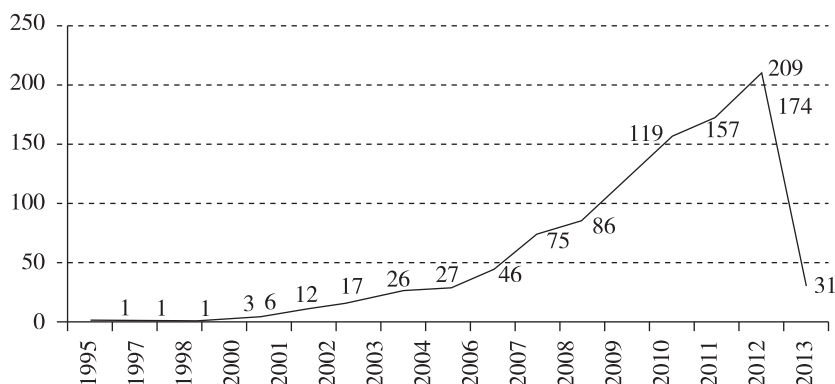


图1 碳纳米管纤维领域科技论文年度分布曲线

3.2 研究方法

本文以碳纳米管纤维领域的高被引论文为例进行实证研究,验证利用引文内容分析来进行高被引论文主题识别的可行性和有效性,具体研究方法如下:

(1) 被引文献集和施引文献集构建

利用 HistCite 对检索到的 991 条数据进行初步分析,利用本地被引用次数(Local Citation Score, LCS)对数据进行排序,选取高被引论文,并构建被引文献集和施引文献集,其中,被引文献集由高被引论文组成,施引文献集由数据集内引用被引文献集的论文组成。

(2) 基于规则的引文内容抽取

从施引文献集中分别抽取被引文献集的引文内容,包括全文预处理、引用句识别和引文内容抽取三个主要步骤。其中,全文预处理包括利用 EndNote 来获得被引文献集和施引文献集中所有文献的 PDF 全文,以及利用 ABBYY FineReader 将 PDF 全文转换成易于处理的 TXT 格式,并进行适当的人工清洗;引用句识别是从施引文献集中识别包含被引文献所对应引用标记的句子;引文内容抽取是在引用句识别的基础上去除噪音并抽取被引文献所对应的引文内容。

(3) 基于 C-value 的高被引论文主题识别

本文采用 C-value 算法^[27-28]进行高被引论文的主题识别,该算法是对纯粹短语词频统计的改

进,加强了多词术语和嵌套术语的识别,并成功应用到英国曼彻斯特大学国家文本挖掘中心的 Ter-Mine 系统中,用于主题词汇抽取和术语识别等。C-value 算法具有领域无关性,是语言学分析方法和统计分析方法相结合的多词术语识别算法,着重强调统计分析方法。语言学分析部分是对文本进行词性标注、形容词/名词序列抽取、停用词表过滤等处理方法获取候选术语。统计分析部分是根据候选术语的四个特征进行统计,构建术语模型判定术语。四个特征包括候选术语的词频、词长、候选术语被其他更长候选术语包含的频次以及这些更长候选术语的数目。

4 结果与讨论

4.1 被引文献集和施引文献集构建

对碳纳米管纤维领域的 991 条检索数据进行统计,其中具有被引用次数的论文 412 篇(占 41.57%),被引用次数大于等于 50 次的 11 篇(占 1.11%),本文选取被引用次数大于等于 20 次的 37 篇(占 3.73%)论文作为该领域数据集内的高被引论文,限于篇幅本文仅给出被引用次数排名前十的高被引论文(见表1)。从表1可以看出,10 篇高被引论文的被引用次数都在 50 次以上,发表时间分布在 2000 年到 2007 年,且来源刊中有 5 篇来自 Science,2 篇来自 Nature。

表1 TOP 10 高被引论文

编号	题名	第一作者	来源刊	时间	LCS
6	Macroscopic fibers and ribbons of oriented carbon nanotubes	Vigolo B	SCIENCE	2000	146
66	Multifunctional carbon nanotube yarns by downsizing an ancient technology	Zhang M	SCIENCE	2004	143
56	Direct spinning of carbon nanotube fibers from chemical vapor deposition synthesis	Li Y L	SCIENCE	2004	98
33	Super-tough carbon-nanotube fibres	Dalton A B	NATURE	2003	96
22	Nanotechnology: Spinning continuous carbon nanotube yarns	Jiang K L	NATURE	2002	94
60	Macroscopic, neat, single-walled carbon nanotube fibers	Ericson L M	SCIENCE	2007	77
18	Carbon nanotube/carbon fiber hybrid multiscale composites	Thostenson E T	JOURNAL OF APPLIED PHYSICS	2002	72
215	High-performance carbon nanotube fiber	Koziol	SCIENCE	2007	68
123	Spinning and processing continuous yarns from 4-inch wafer scale super-aligned carbon nanotube arrays	Zhang X B	ADVANCED MATERIALS	2006	55
57	Saturable absorbers incorporating carbon nanotubes directly synthesized onto substrates and fibers and their application to mode-locked fiber lasers	Yamashita S	OPTICS LETTERS	2004	52

将碳纳米管纤维领域的 37 篇高被引论文作为被引文献集,利用 HistCite 软件得到被引文献集在实验数据集中的施引情况,获得每篇被引文献对应的施引文献编号(见图2)。其中 Num 指论文编号,

TI 指被引文献的标题,LCS 指被引文献的被引用次数,Citing 指该被引文献所对应的施引文献集,施引文献以#来进行间隔,Citing 中施引文献的数量与引用的次数 LCS 一致。

1	Num	TI	LCS	Citing
2	6	Macroscopic fibers and ribbons of oriented carbon nanotubes	146	10#11#13#14#15#17#19#21#22#26#28#31#33#35#40#41#44#45#46#47#48#49#50#51#52#53#54#55#56#57#58#59#60#61#62#63#64#65#66#67#68#69#70#71#72#73#74#75#76#77#78#79#80#81#82#83#84#85#86#87#88#89#90#91#92#93#94#95#96#97#98#99#100#101#102#103#104#105#106#107#108#109#110#111#112#113#114#115#116#117#118#119#120#121#122#123#124#125#126#127#128#129#130#131#132#133#134#135#136#137#138#139#140#141#142#143#144#145#146#147#148#149#150#151#152#153#154#155#156#157#158#159#160#161#162#163#164#165#166#167#168#169#170#171#172#173#174#175#176#177#178#179#180#181#182#183#184#185#186#187#188#189#190#191#192#193#194#195#196#197#198#199#200#201#202#203#204#205#206#207#208#209#210#211#212#213#214#215#216#217#218#219#220#221#222#223#224#225#226#227#228#229#230#231#232#233#234#235#236#237#238#239#240#241#242#243#244#245#246#247#248#249#250#251#252#253#254#255#256#257#258#259#260#261#262#263#264#265#266#267#268#269#270#271#272#273#274#275#276#277#278#279#280#281#282#283#284#285#286#287#288#289#290#291#292#293#294#295#296#297#298#299#300#301#302#303#304#305#306#307#308#309#310#311#312#313#314#315#316#317#318#319#320#321#322#323#324#325#326#327#328#329#330#331#332#333#334#335#336#337#338#339#340#341#342#343#344#345#346#347#348#349#350#351#352#353#354#355#356#357#358#359#360#361#362#363#364#365#366#367#368#369#370#371#372#373#374#375#376#377#378#379#380#381#382#383#384#385#386#387#388#389#390#391#392#393#394#395#396#397#398#399#400#401#402#403#404#405#406#407#408#409#410#411#412#413#414#415#416#417#418#419#420#421#422#423#424#425#426#427#428#429#430#431#432#433#434#435#436#437#438#439#440#441#442#443#444#445#446#447#448#449#450#451#452#453#454#455#456#457#458#459#460#461#462#463#464#465#466#467#468#469#470#471#472#473#474#475#476#477#478#479#480#481#482#483#484#485#486#487#488#489#490#491#492#493#494#495#496#497#498#499#500#501#502#503#504#505#506#507#508#509#510#511#512#513#514#515#516#517#518#519#520#521#522#523#524#525#526#527#528#529#530#531#532#533#534#535#536#537#538#539#540#541#542#543#544#545#546#547#548#549#550#551#552#553#554#555#556#557#558#559#560#561#562#563#564#565#566#567#568#569#570#571#572#573#574#575#576#577#578#579#580#581#582#583#584#585#586#587#588#589#590#591#592#593#594#595#596#597#598#599#600#601#602#603#604#605#606#607#608#609#610#611#612#613#614#615#616#617#618#619#620#621#622#623#624#625#626#627#628#629#630#631#632#633#634#635#636#637#638#639#640#641#642#643#644#645#646#647#648#649#650#651#652#653#654#655#656#657#658#659#660#661#662#663#664#665#666#667#668#669#670#671#672#673#674#675#676#677#678#679#680#681#682#683#684#685#686#687#688#689#690#691#692#693#694#695#696#697#698#699#700#701#702#703#704#705#706#707#708#709#710#711#712#713#714#715#716#717#718#719#720#721#722#723#724#725#726#727#728#729#730#731#732#733#734#735#736#737#738#739#740#741#742#743#744#745#746#747#748#749#750#751#752#753#754#755#756#757#758#759#760#761#762#763#764#765#766#767#768#769#770#771#772#773#774#775#776#777#778#779#780#781#782#783#784#785#786#787#788#789#790#791#792#793#794#795#796#797#798#799#800#801#802#803#804#805#806#807#808#809#810#811#812#813#814#815#816#817#818#819#820#821#822#823#824#825#826#827#828#829#830#831#832#833#834#835#836#837#838#839#840#841#842#843#844#845#846#847#848#849#850#851#852#853#854#855#856#857#858#859#860#861#862#863#864#865#866#867#868#869#870#871#872#873#874#875#876#877#878#879#880#881#882#883#884#885#886#887#888#889#890#891#892#893#894#895#896#897#898#899#900#901#902#903#904#905#906#907#908#909#910#911#912#913#914#915#916#917#918#919#920#921#922#923#924#925#926#927#928#929#930#931#932#933#934#935#936#937#938#939#940#941#942#943#944#945#946#947#948#949#950#951#952#953#954#955#956#957#958#959#960#961#962#963#964#965#966#967#968#969#970#971#972#973#974#975#976#977#978#979#980#981#982#983#984#985#986#987#988#989#990#991#992#993#994#995#996#997#998#999#1000
3	66	Multifunctional carbon nanotube yarns by downsizing an ancient technology	143	121#123#125#139#147#151#156#162#175#177#187#193#198#202#204#206#208#210#212#214#216#218#220#222#224#226#228#230#232#234#236#238#240#242#244#246#248#250#252#254#256#258#260#262#264#266#268#270#272#274#276#278#280#282#284#286#288#290#292#294#296#298#300#302#304#306#308#310#312#314#316#318#320#322#324#326#328#330#332#334#336#338#340#342#344#346#348#350#352#354#356#358#360#362#364#366#368#370#372#374#376#378#380#382#384#386#388#390#392#394#396#398#400#402#404#406#408#410#412#414#416#418#420#422#424#426#428#430#432#434#436#438#440#442#444#446#448#450#452#454#456#458#460#462#464#466#468#470#472#474#476#478#480#482#484#486#488#490#492#494#496#498#500#502#504#506#508#510#512#514#516#518#520#522#524#526#528#530#532#534#536#538#540#542#544#546#548#550#552#554#556#558#560#562#564#566#568#570#572#574#576#578#580#582#584#586#588#590#592#594#596#598#600#602#604#606#608#610#612#614#616#618#620#622#624#626#628#630#632#634#636#638#640#642#644#646#648#650#652#654#656#658#660#662#664#666#668#670#672#674#676#678#680#682#684#686#688#690#692#694#696#698#700#702#704#706#708#710#712#714#716#718#720#722#724#726#728#730#732#734#736#738#740#742#744#746#748#750#752#754#756#758#760#762#764#766#768#770#772#774#776#778#780#782#784#786#788#790#792#794#796#798#800#802#804#806#808#810#812#814#816#818#820#822#824#826#828#830#832#834#836#838#840#842#844#846#848#850#852#854#856#858#860#862#864#866#868#870#872#874#876#878#880#882#884#886#888#890#892#894#896#898#900#902#904#906#908#910#912#914#916#918#920#922#924#926#928#930#932#934#936#938#940#942#944#946#948#950#952#954#956#958#960#962#964#966#968#970#972#974#976#978#980#982#984#986#988#990#992#994#996#998#1000
4	56	Direct spinning of carbon nanotube fibers from chemical vapor deposition synthesis	98	60#66#75#77#87#89#92#94#123#125#156#162#168#198#202#204#206#208#210#212#214#216#218#220#222#224#226#228#230#232#234#236#238#240#242#244#246#248#250#252#254#256#258#260#262#264#266#268#270#272#274#276#278#280#282#284#286#288#290#292#294#296#298#300#302#304#306#308#310#312#314#316#318#320#322#324#326#328#330#332#334#336#338#340#342#344#346#348#350#352#354#356#358#360#362#364#366#368#370#372#374#376#378#380#382#384#386#388#390#392#394#396#398#400#402#404#406#408#410#412#414#416#418#420#422#424#426#428#430#432#434#436#438#440#442#444#446#448#450#452#454#456#458#460#462#464#466#468#470#472#474#476#478#480#482#484#486#488#490#492#494#496#498#500#502#504#506#508#510#512#514#516#518#520#522#524#526#528#530#532#534#536#538#540#542#544#546#548#550#552#554#556#558#560#562#564#566#568#570#572#574#576#578#580#582#584#586#588#590#592#594#596#598#600#602#604#606#608#610#612#614#616#618#620#622#624#626#628#630#632#634#636#638#640#642#644#646#648#650#652#654#656#658#660#662#664#666#668#670#672#674#676#678#680#682#684#686#688#690#692#694#696#698#700#702#704#706#708#710#712#714#716#718#720#722#724#726#728#730#732#734#736#738#740#742#744#746#748#750#752#754#756#758#760#762#764#766#768#770#772#774#776#778#780#782#784#786#788#790#792#794#796#798#800#802#804#806#808#810#812#814#816#818#820#822#824#826#828#830#832#834#836#838#840#842#844#846#848#850#852#854#856#858#860#862#864#866#868#870#872#874#876#878#880#882#884#886#888#890#892#894#896#898#900#902#904#906#908#910#912#914#916#918#920#922#924#926#928#930#932#934#936#938#940#942#944#946#948#950#952#954#956#958#960#962#964#966#968#970#972#974#976#978#980#982#984#986#988#990#992#994#996#998#1000
5	33	Super-tough carbon-nanotube fibres	96	45#53#56#60#62#66#67#70#75#79#81#82#86#89#92#94#125#137#139#141#143#145#147#149#151#153#155#157#159#161#163#165#167#169#171#173#175#177#179#181#183#185#187#189#191#193#195#197#199#201#203#205#207#209#211#213#215#217#219#221#223#225#227#229#231#233#235#237#239#241#243#245#247#249#251#253#255#257#259#261#263#265#267#269#271#273#275#277#279#281#283#285#287#289#291#293#295#297#299#301#303#305#307#309#311#313#315#317#319#321#323#325#327#329#331#333#335#337#339#341#343#345#347#349#351#353#355#357#359#361#363#365#367#369#371#373#375#377#379#381#383#385#387#389#391#393#395#397#399#401#403#405#407#409#411#413#415#417#419#421#423#425#427#429#431#433#435#437#439#441#443#445#447#449#451#453#455#457#459#461#463#465#467#469#471#473#475#477#479#481#483#485#487#489#491#493#495#497#499#501#503#505#507#509#511#513#515#517#519#521#523#525#527#529#531#533#535#537#539#541#543#545#547#549#551#553#555#557#559#561#563#565#567#569#571#573#575#577#579#581#583#585#587#589#591#593#595#597#599#601#603#605#607#609#611#613#615#617#619#621#623#625#627#629#631#633#635#637#639#641#643#645#647#649#651#653#655#657#659#661#663#665#667#669#671#673#675#677#679#681#683#685#687#689#691#693#695#697#699#701#703#705#707#709#711#713#715#717#719#721#723#725#727#729#731#733#735#737#739#741#743#745#747#749#751#753#755#757#759#761#763#765#767#769#771#773#775#777#779#781#783#785#787#789#791#793#795#797#799#801#803#805#807#809#811#813#815#817#819#821#823#825#827#829#831#833#835#837#839#841#843#845#847#849#851#853#855#857#859#861#863#865#867#869#871#873#875#877#879#881#883#885#887#889#891#893#895#897#899#901#903#905#907#909#911#913#915#917#919#921#923#925#927#929#931#933#935#937#939#941#943#945#947#949#951#953#955#957#959#961#963#965#967#969#971#973#975#977#979#981#983#985#987#989#991#993#995#997#999#1000
6	22	Nanotechnology: Spinning continuous carbon nanotube yarns	94	33#35#56#60#66#75#92#93#121#123#125#128#139#156#162#187#189#191#193#195#197#199#201#203#205#207#209#211#213#215#217#219#221#223#225#227#229#231#233#235#237#239#241#243#245#247#249#251#253#255#257#259#261#263#265#267#269#271#273#275#277#279#281#283#285#287#289#291#293#295#297#299#301#303#305#307#309#311#313#315#317#319#321#323#325#327#329#331#333#335#337#339#341#343#345#347#349#351#353#355#357#359#361#363#365#367#369#371#373#375#377#379#381#383#385#387#389#391#393#395#397#399#401#403#405#407#409#411#413#415#417#419#421#423#425#427#429#431#433#435#437#439#441#443#445#447#449#451#453#455#457#459#461#463#465#467#469#471#473#475#477#479#481#483#485#487#489#491#493#495#497#499#501#503#505#507#509#511#513#515#517#519#521#523#525#527#529#531#533#535#537#539#541#543#545#547#549#551#553#555#557#559#561#563#565#567#569#571#573#575#577#579#581#583#585#587#589#591#593#595#597#599#601#603#605#607#609#611#613#615#617#619#621#623#625#627#629#631#633#635#637#639#641#643#645#647#649#651#653#655#657#659#661#663#665#667#669#671#673#675#677#679#681#683#685#687#689#691#693#695#697#699#701#703#705#707#709#711#713#715#717#719#721#723#725#727#729#731#733#735#737#739#741#743#745#747#749#751#753#755#757#759#761#763#765#767#769#771#773#775#777#779#781#783#785#787#789#791#793#795#797#799#801#803#805#807#809#811#813#815#817#819#821#823#825#827#829#831#833#835#837#839#841#843#845#847#849#851#853#855#857#859#861#863#865#867#869#871#873#875#877#879#881#883#885#887#889#891#893#895#897#899#901#903#905#907#909#911#913#915#917#919#921#923#925#927#929#931#933#935#937#939#941#943#945#947#949#951#953#955#957#959#961#963#965#967#969#971#973#975#977#979#981#983#985#987#989#991#993#995#997#999#1000
7	60	Macroscopic, neat, single-walled carbon nanotube fibers	77	66#81#86#87#89#92#94#111#123#125#138#139#149#162#187#191#193#195#197#199#201#203#205#207#209#211#213#215#217#219#221#223#225#227#229#231#233#235#237#239#241#243#245#247#249#251#253#255#257#259#261#263#265#267#269#271#273#275#277#279#281#283#285#287#289#291#293#295#297#299#301#303#305#307#309#311#313#315#317#319#321#323#325#327#329#331#333#335#337#339#341#343#345#347#349#351#353#355#357#359#361#363#365#367#369#371#373#375#

4.2 基于规则的引文内容抽取

基于规则的引文内容抽取是从施引文献集中抽取被引文献对应的引文内容,抽取对象是 PDF 全文转换成的 TXT 文本,抽取过程主要包括引用句识别和引文内容抽取。

(1) 引用句识别

参考文献记录分为有序号和无序号两种形式,其中有序号形式指“序号+被引文献记录”,常见序号格式有:数字、[数字]、(数字)等,无序号形式是直接给出被引文献记录,而没有序号。依次对每篇被引文献所对应的施引文献进行处理,对于有序号形式的利用第一作者、来源、年、卷等进行匹配,获得被引文献在施引文献全文文本中所对应的参考文献记录,抽取参考文献记录所对应的序号。如图 3 中可利用第一作者 Dalton、来源 *Nature*、年 2003、卷 423 等获得 33 号被引文献在 125 号施引文献参考文献记录中的序号为[7]。对于无序号形式的不进行这个处理,直接进行引用句的抽取。

- [7] A. B. Dalton, S. Collins, E. Muñoz, J. M. Razal, V. H. Ebron, J. P. Ferraris, J. N. Coleman, B. G. Kim, R. H. Baughman, *Nature* **2003**, 423, 703.
[8] M. Zhang, K. R. Atkinson, R. H. Baughman, *Science* **2004**, 306, 1358.

图 3 参考文献列表

对于有序号形式的根据参考文献序号定位被引文献在施引文献全文文本正文中的引用位置,抽取引用句。参考文献序号在正文中出现的形式主要有:单序号(如 3、(9)、[6]);双序号(如 12,13、[19,24]、[23,24]);连续多序号(如 2-4、(15-17)、[1-3]);不连续多序号(如[15,16,24]、[23,27-29]、[16]、[18]-[23])等,其中多序号是指包含三个及以上序号。对于单序号和双序号可直接进行序号匹配,对于多序号形式需要判定定位的序号是否在其范围内。对于无序号形式在正文中一般以作者和年的组合出现,常见的格式有:(作者年)、作者(年)、(作者,年)等,直接利用第一作者和年的匹配来定位引用句的位置,进行引用句的抽取。

(2) 引文内容抽取

由于抽取的引用句可能包含多个被引文献,

因此需要对引用句进行处理,获得所需被引文献对应的引文内容。引文内容抽取是根据参考文献序号的位置来判定的,主要分为以下几种情况:

① 如果引用句中只有一处含有参考文献序号且包含被引文献对应的参考文献序号,那么整个引用句作为引文内容。

Here we report a simple and versatile approach that can create rigid fibers and ribbons of preferentially oriented SWNTs (11).

When nanotubes are randomly oriented in the matrix, polarized Raman spectroscopy selects out the signal from those nanotubes lying in the polarization direction [14-17].

② 如果引用句中有多处含有参考文献序号且有标点符号或并列、转折、比较等连接词(and、or、but、though、either...or、both...and、not only...but also、as、whereas、as well as、in contrast to、similar to 等)将引用句分隔成若干部分,那么保留包含被引文献对应的参考文献序号所在的部分。

Carbon nanotube (CNT) fiber performance has improved significantly since 2000 [1] , now reaching strengths (9 GPa) and moduli (350 GPa) comparable or greater than those reported for carbon fibers [2,3].

Applications of CNT yarns as high strength conductive fibers [10,15,16] and thermionic and field emission electron sources [12-14] have been explored .

In addition , these carbon nanotube fibres are good electrical [8,9,11] and heat conductors [12,13] , have a high thermal and chemical stability and are very resilient to maltreatment due to their yarn - like character [14] .

③ 如果引用句中有多处含有参考文献序号且作为句子的同一成分并列出现,通常以冒号或列举词(for instance、such as、including、by means of、because of、using、through、about、from、by、for、e. g. 等)开始,那么保留包含被引文献对应的参考文献序号所在的部分。

CNT fibers can be fabricated through coagula-

tion-based wet spinning[2], direct synthesis from CNT aerogel[3], and a dry-state spinning out of vertically aligned CNT arrays[4].

Recent progress have shown the possibility to make optimized materials: films of aligned nanotubes by using high magnetic fields 2-4 and fibers of aligned nanotubes by using an electrophoretic process 5 or spinning aqueous suspensions. 6

④ 如果引用句中有多处含有参考文献序号且分布包含多种情况,并将引用句分割成若干部分,

那么以保留包含被引文献对应的参考文献序号所在的若干部分为首要原则。

Compared with other methods [1-7], it simplifies processing [1-6], avoid the toxic acyl chlorination processing[5] or catalyst contamination[1-4] and increases controllability and effectivity[7].

通过以上过程可以获得每篇高被引论文在各自施引文献集中的引文内容,抽取结果如图4所示。最后将引文内容按照施引文献进行汇总,得到高被引论文的引文内容。

	A	B	C	D	E
1	被引文献	施引文献	引文内容	引文句	参考文献
2		6	13 The achievement of such i	The achievement of such ropes [31]	Vigolo B, Penicaud A, Coulon C
3		6	19 Recent progress have shown th	Recent progress have shown th	B. Vigolo, B. Penicaud, C. Coulon
4		6	21 Several techniques have b	Several techniques have been d[1]	B. Vigolo, A. Penicaud, C. Coulon
5		6	22 The creation of continuo	The creation of continuous yal.	Vigolo, B. et al. Science 290, 1
6		6	28 A simple particle-coagula	A simple particle-coagulation (1)	Vigolo, B.; Penicaud, A.; Coulon
7		6	31 No density correction wa	No density correction was app	(23) Vigolo, B.; Penicaud, A.; Coulon
8		6	33 In the original process ,	In the original process 4 , s4.	Vigolo, B. et al. Science 290, 1
9		6	35 In order to translate th	In order to translate the sup	[8] B. Vigolo, A. Penicaud, C. Coulon
10		6	40 Recent progress has show	Recent progress has shown the	B. Vigolo, A. Penicaud, C. Coulon
11		6	41 CNT fibers , developed r	CNT fibers , developed recent	(11) Vigolo, B.; Penicaud, A.; Coulon
12		6	45 Drawing these as-spun ge	Drawing these as-spun gel fib	19 B. Vigolo, A. Penicaud, C. Coulon
13		6	53 An interesting new proces	An interesting new process fo	[23] B. Vigolo, A. Penicaud, C. Coulon
14		6	56 Similarly , nanotube-pol	Similarly , fibers of nanotub	11. B. Vigolo et al. , Science 290 ,
15		6	60 To date , SWNT composite	To date , SWNT composite fibe	12. B. Vigolo et al. , Science 290 ,
16		6	66 Breakthroughs have been	Breakthroughs have been made	[3. B. Vigolo et al. , Science 290 , 1

图4 引文内容抽取结果(部分)

4.3 基于 C-value 的高被引论文主题识别

利用 C-value 算法分别对 37 篇高被引论文的引文内容进行候选主题抽取,得到 37 组候选主题列表,表 2 给出了第 1 篇高被引论文的部分候选主题。

表 2 第 1 篇高被引论文的部分候选主题

序号	候选主题	C-value 值
1	cnt fiber	26.00
2	carbon nanotube	24.71
3	composite fiber	16.82
4	wet spinning	12.33
5	coagulation spinning	9.63
6	continuous fiber	9.00
7	carbon nanotube fiber	7.92
8	cnt dispersion	7.00
9	sodium dodecyl sulfate	6.34
10	vinyl alcohol	6.00
11	simple particle-coagulation spinning	5.34
12	cnt yarn	5.00

由表 2 可知,每篇高被引论文可由一系列候选主题共同来表征,但是每个候选主题所占的权重(C-value 值)并不相同,权重大的候选主题能够更好地揭示高被引论文,本文将权重较大的候选主题定义为核心主题。对于抽取出的候选主题,需要设定阈值来从中选取代表高被引论文的核心主题。本文通过实验发现,C-value 值大于 5.00 的候选主题能够很好地表征相关内容,大于 10.00 的候选主题具有强相关性,能够表征论文的核心研究内容。因此,本文将 C-value 值的阈值设置为 10.00,大于阈值的候选主题作为该论文的核心主题,限于篇幅,本文只给出 TOP10 高被引论文的核心主题识别结果(见表 3),核心主题之间用#间隔,圆括号内数值为主题词对应的 C-value 值。

为了验证实验结果,本文通过人工判读的方法进行分析,分别对 37 篇高被引论文的标题、摘要以及引文内容进行判读,结果发现识别的核心主题能

表 3 TOP 10 高被引论文核心主题

编号	核心主题	主题数
6	cnt fiber(26.00)#carbon nanotube(24.71)#composite fiber(16.82)#wet spinning(12.33)	4
66	cnt fiber(46.80)#carbon nanotube(41.41)#cnt yarn(29.76)#cnt array(27.00)#tensile strength(18.67)#aligned cnt array(12.68)#dry spinning(10.67)#aligned cnt(10.00)	8
56	cnt fiber(34.86)#chemical vapor deposition(30.62)#carbon nanotube(21.30)#direct spinning(13.00)	4
33	composite fiber(25.65)#carbon nanotube(18.50)#cnt fiber(13.00)	3
22	cnt array(21.13)#carbon nanotube(20.43)#cnt yarn(20.40)#cnt fiber(16.00)#dry spinning(10.67)	5
60	cnt fiber(26.33)#wet spinning(15.75)#carbon nanotube(13.00)	3
18	carbon fiber(38.75)#chemical vapor deposition(24.36)#carbon nanotube(19.00)#interfacial shear strength(17.43)#interfacial load transfer(11.68)	5
215	cnt fiber(28.83)#carbon nanotube(14.40)#chemical vapor deposition(10.09)	3
123	cnt yarn(15.00)#cnt fiber(11.67)	2
57	carbon nanotube(10.40)	1

够较好地揭示高被引论文被引的原因,而且与论文的研究内容相符合。例如,第 1 篇高被引论文“Macroscopic fibers and ribbons of oriented carbon nanotubes”^[29]的研究内容是提出了一种将单壁碳纳米管装配成较长丝带和纤维的方法,该方法首先将纳米管分散在表面活性剂溶液中,然后在聚合物溶液的流动中浓缩纳米管形成纳米管网格,最后将网格整理成纳米管纤维。通过引文内容抽取的核心主题较好地反映了该研究内容,即将碳纳米管(carbon nanotube)利用凝固纺丝法(coagulation spinning)来制备碳纳米管纤维(cnt fiber),其中,凝固纺丝法属于湿法纺丝(wet spinning),其生成的纤维属于复合纤维(composite fiber)。而且抽取的核心主题来自施引文献集,更加具有客观性,能更好地反映出该论文被引的原因,如制备方法 coagulation spinning 和 wet spinning 在原文中并没有出现,是后续研究引用该论文时对其研究方法的总结,具有更好的主题特征,是对原文内容的重要语义补充。

为了进一步验证结果的有效性,本文分别利用第 1 篇高被引论文的标题和摘要、全文进行主题识别,同样利用 C-value 算法,实验结果对比如表 4 所示,每个结果给出排在前五位的主题词。从形式上,基于引文内容主题识别的 C-value 值明显高于基于全文、基于标题和摘要的值,因而具有更好的主题代表性,基于全文的值又稍好于基于标题和摘要的值;从内容上,基于标题和摘要、基于全文抽取的主题词都来自原文,且有一定的重复性,虽然抽取的主题词也都与论文研究内容相关,但是不能够直接反映出论文最主要的贡献和研究内容,而基于引文内容主题抽取效果要明显好于前两者,主要原因是引文内容来自多篇施引文献,且施引文献都对引文研究内容进行了总结和归纳,主题更加集中。通过实验结果对比发现,在引文内容分析基础上识别的主题能够有效揭示被引文献的研究内容,是对原文相关信息的重要补充,为领域主题演化、主题聚类等进一步分析奠定了基础。

表 4 实验结果对比

引文内容		标题和摘要		全文	
cnt fiber	26.00	nanotube fiber	2.00	swnt fiber	6.00
carbon nanotube	24.71	classical carbon fiber	1.58	van der waal attraction	6.00
composite fiber	16.82	single-walled carbon nanotube	1.58	nanotube bundle	5.00
wet spinning	12.33	high-quality bucky paper	1.58	high-quality bucky paper	3.17
coagulation spinning	9.63	oriented carbon nanotube	1.58	classical carbon fiber	3.17

5 结语

引用行为具有复杂的动机,而目前的引文分析较少考虑引用动机,且将所有的引文同等看待,被引用次数只能从外在计量论文的影响力,无法揭示论文的研究内容;同时,基于引文的共引、耦合聚类分析等情报研究方法通常选用关键词或从标题、摘要、全文中抽取主题词来表征论文的研究主题,仍然是从论文自身的内容出发,无法揭示论文的被引原因。本文提出基于引文内容分析来进行高被引论文主题识别的研究思路,在一定程度上解决了以上问题,对传统引文分析方法具有很好的借鉴意义,而且通过碳纳米管纤维领域的实证研究验证了该方法的有效性和实用性。

为了进一步提高该方法的应用性,本文认为该方法还需要从三个方面进行完善。

(1)引文内容方面:本文中的引文内容是从引用句中抽取的,没有涉及引文上下文,而引文上下文通常是对引用句的进一步阐述,蕴含着丰富的语义信息,为揭示施引文献和被引文献的关联提供了

更多线索,进一步的工作需要加强这些内容的抽取和分析。

(2)主题识别方面:本文选取 C-value 算法来进行主题识别,该算法对于多词术语和嵌套术语的抽取效果很好,但该方法存在不能识别单词术语等缺陷,与其他方法结合使用效果会更好,比如陈仕吉等^[30]提出的基于 C-value 与 TF-IDF 的文献簇主题识别方法。另外,概率主题模型在主题演化等方面具有一定的优势,如崔凯^[31]基于 LDA 的主题演化研究与实现,进一步尝试将 LDA 等概率主题模型应用到基于引文内容的主题识别和主题演化分析。

(3)方法应用方面:本文通过实证研究已经初步发现引文内容相对于被引文献原文的优势,引文内容和原文内容的结合使用能更好地表征被引文献的研究内容。另外,主题识别是引文内容分析的基础,下一步研究将在主题识别的基础上进行面向问题和应用的研究,比如前面相关研究中提到国外学者针对引文内容分析开展的面向自动文摘、信息检索等应用的研究。

参考文献

[1] The Thomson Corporation. Essentialscience indicators [EB/OL]. [2013 - 05 - 27]. <http://esi.webofknowledge.com/home.cgi>.
[2] Garfield E. Citation analysis as a tool in journal evaluation[J]. Science, 1972(178): 471 - 479.
[3] Hirsch J E. An index to quantify an individual's scientific research output[C]//Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(46): 16569 - 16572.

- [4] 苏成,潘云涛,马峥,等. 基于 PrestigeRank 算法与同行评议的科技论文评价研究[J]. 情报学报, 2012, 31(2): 180-188. (Su Cheng, Pan Yuntao, Ma Zheng, et al. Prestigerank and peer review for evaluation of scientific papers [J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(2): 180-188.)
- [5] 许海云,方曙. 基于引文分布和引文网络的学术期刊评价指标研究[J]. 情报科学, 2013, 31(2): 130-139. (Xu Haiyun, Fang Shu. Study of evaluation index based on distribution and networks of citation for academic journals[J]. Information Science, 2013, 31(2): 130-139.)
- [6] Small H, Sweeney E, Greenlee E. Clustering the science citation index using co-citations. II. Mapping science[J]. Scientometrics, 1985, 8(5-6): 321-340.
- [7] Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the American Society for Information Science and Technology, 2006, 57(3): 359-377.
- [8] Garfield E. Can citation indexing be automated?[C]//Statistical Association Methods For Mechanized Documentation, Symposium Proceedings, 1964: 189-192.
- [9] Teufel S. Argumentative zoning; Information extraction from scientific text[D]. University of Edinburgh, 1999.
- [10] Pham S B, Hoffmann A. A new approach for scientific citation classification using cue phrases[C]//Proceedings of Australian Joint Conference in Artificial Intelligence, 2003: 759-771.
- [11] Le Minh-Hoang, HoTu-Bao, Nakamori Yoshiteru. Detecting emerging trends from scientific corpora[J]. International Journal of Knowledge and Systems Sciences, 2005, 2(2): 53-59.
- [12] 侯跃芳,崔雷,吴迪. 应用引文共引聚类——内容词分析法对学科发展的研究[J]. 情报学报, 2007, 26(2): 309-314. (Hou Yuefang, Cui Lei, Wu Di. Study on the history and the current work of a subject by combined co-citation cluster and content words analysis[J]. Journal of the China Society for Scientific and Technical Information, 2007, 26(2): 309-314.)
- [13] Small H. Citation context analysis[C]//Progress in Communication Sciences, 1982: 287-310.
- [14] Ding Y. Content-based citation analysis: The next generation in citation analysis [EB/OL]. [2012-11-14]. <http://www.lis.illinois.edu/events/2012/09/26/content-based-citation-analysis-next-generation-citation-analysis>.
- [15] Small H G. Cited documents as concept symbols[J]. Social Studies of Science, 1978, 8(3): 327-340.
- [16] Small H G, Greenlee E. Citation context analysis of a co-citation cluster: Recombinant-DNA[J]. Scientometrics, 1980, 2(4): 277-301.
- [17] Nakov P, Schwartz A, Hearst M. Citances: Citation sentences for semantic analysis of bioscience text[C]//Proceedings of the SIGIR'04 Workshop on Search and Discovery in Bioinformatics, 2004: 81-88.
- [18] Elkiss A, Shen S, Fader A, et al. Blind men and elephants: What do citation summaries tell us about a research article?[J]. Journal of the American Society for Information Science and Technology, 2008, 59(1): 51-62.
- [19] Liu X, Zhang J, Guo C. Full-text citation analysis: A new method to enhance scholarly networks[J]. Journal of the American Society for Information Science and Technology, 2013, 64(9): 1852-1863.
- [20] Teufel S. Argumentative zoning for improved citation indexing[M]//Computing Attitude and Affect in Text: Theory and Applications, 2006(20): 159-169.
- [21] Mohammad S, Dorr B, Egan M, et al. Using citations to generate surveys of scientific paradigms[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies, 2009: 584-592.
- [22] Qazvinian V, Radev D R, Mohammad S M, et al. Generating extractive summaries of scientific paradigms[J]. Journal of Artificial Intelligence Research, 2013(46): 165-201.

- [23] Bradshaw S. Reference directed indexing: Indexing scientific literature in the context of its use[D]. Northwestern University, 2002.
- [24] Ritchie A, Teufel S, Robertson S. Using terms from citations for IR: Some first results[C]//Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval (ECIR), 2008: 211–221.
- [25] Aljaber B, Martinez D, Stokes N, et al. Improving MeSH classification of biomedical articles using citation contexts[J]. Journal of Biomedical Informatics, 2011(44): 881–896.
- [26] Aljaber B, Stokes N, Bailey J, et al. Document clustering of scientific texts using citation contexts[J]. Information Retrieval, 2010, 13(2): 101–131.
- [27] Frantzi K T, Ananiadou S, Tsujii J. The C-value/NC-value method of automatic recognition for multi-word terms[C]//Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries, 1998: 585–604.
- [28] Frantzi K T, Ananiadou S, Mima H. Automatic recognition of multi-word terms: The C-value/NC-value method[J]. International Journal on Digital Libraries, 2000, 3(2): 1–17.
- [29] Vigolo B, Penicaud A, Coulon C, et al. Macroscopic fibers and ribbons of oriented carbon nanotubes[J]. Science, 2000, 290(5495): 1331–1334.
- [30] 陈仕吉,王小梅. 基于 C-value 与 TF-IDF 的文献簇主题识别研究[J]. 情报学报, 2009, 28(6): 821–826. (Chen Shiji, Wang Xiaomei. The topic recognition research of paper cluster by combining C-value and TF-IDF[J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(6): 821–826.)
- [31] 崔凯. 基于 LDA 的主题演化研究与实现[D]. 长沙: 国防科学技术大学, 2010. (Cui Kai. The research and implementation of topic evolution based on LDA[D]. Changsha: National University of Defense Technology, 2010.)

祝清松 中国科学院国家科学图书馆博士研究生。

通讯地址: 北京海淀区北四环西路 33 号中国科学院国家科学图书馆。邮编: 100190。

冷伏海 中国科学院国家科学图书馆研究员。通讯地址同上。

(收稿日期: 2013-05-29; 修回日期: 2013-07-26)