View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by National Science Library,Chinese Academy of Sciences

自动术语识别——对科技文献进行文本挖掘的重要技术方法*

刘建华[1,2]　张智雄[1]　徐　健[1,2,3]　许雁冬[1]
[1](中国科学院国家科学图书馆　北京 100190)
[2](中国科学院研究生院　北京 100049)
[3](中山大学资讯管理系　广州 510275)

【摘要】自动术语识别是知识抽取和文本挖掘等信息技术中的关键步骤。研究现有自动术语识别的主要思路,明确其中的关键问题,研究已有的相关项目和系统的术语识别方法,并分析现有的一些术语资源。借此丰富基于术语识别的文本挖掘理论和方法,为进一步构建相关试验系统提供良好借鉴。
【关键词】自动术语识别　术语变体　术语歧义
【分类号】G250.73

Automatic Term Recognition——An Important Method for Text Mining on Scientific Literature

Liu Jianhua[1,2]　Zhang Zhixiong[1]　Xu Jian[1,2,3]　Xu Yandong[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](Graduate University of the Chinese Academy of Sciences, Beijing 100049, China)
[3](Department of Information Management, Sun Yat-Sen University, Guangzhou 510275, China)

【Abstract】Automatic Term Recognition(ATR) is a key process of knowledge technology such as knowledge extraction and text mining. To enrich the text mining theories and methods based on term recognition, support constructing related systems, it refers to some main existing methods for ATR, find key problems of the process. Through researches on related programs and systems, existing term resources, we could choose the best one for ourselves' ATR system.
【Keywords】Automatic term recognition　Term variation　Term ambiguity

1　引　言

为快速准确地从急剧增长的科技文献等自由文本中获取知识,知识抽取应运而生[1]。抽取对象不仅包含时间、人物等实体,对特殊领域而言,更重要的是集中体现学科领域核心知识的术语和术语间关系[2,3]。基于术语和术语间关系构建领域术语库,可为知识抽取、文本挖掘、链接分析等提供结构化知识单元,实现领域新兴研究探测等[4,5]。

鉴于此,自动扫描自由文本,识别文中指代概念的词串的自动术语识别(Automatic Term Recognition, ATR)[6]吸引了越来越多研究者。国内外不少研究机构,如斯坦福大学、英国曼彻斯特国家文本挖掘中心(National Center for Text Mining, NaCTeM)[7]、北京大学计算语言学研究所等纷纷开展研究,开发出 TerMine、ATRACT 等 ATR 工具。

收稿日期: 2008-06-16
*本文系国家社会科学基金项目"从数字信息资源中实现知识抽取的理论和方法研究"(项目编号:05BTQ006)的研究成果之一。

作为本体学习任务之一，术语识别在 Text2Onto、OntoLiFT 等本体学习工具中也得到不同程度的实现[8]。这些研究为 ATR 提供了理论参考和实践模型。

ATR 作为知识抽取第一步，也是文本挖掘等知识技术的关键步骤[5,6]。研究现有 ATR 的主要思路和关键问题，分析相关项目和系统的方法，可为构建 ATR 系统提供借鉴，同时也为探讨基于术语识别的文本挖掘理论和方法提供基础。

2　术语识别的主要思路

术语作为特殊主题领域某概念的指定名称，可能是单词、符号、化学式或数学公式、组织或管理部门的正式名称（如图 1）等[9]。它们与一般词或短语的区别在于，术语与指定概念间有单一意义关系（单义性），在文中表达概念时形式和内容具有稳固性。对特定领域而言，术语较其它一般词具有频繁使用、相对固定的上下文环境（即共现）、特定排版（如斜体）等特点。一般情况下，术语是名词或名词短语，但某些时候动词、形容词、动词短语或形容词短语都可能成为术语[9,10]。

Profits at Canada's six big banks topped C$6 billion ($ 4.4 billion) in 1996, smashing last year's C$5.2 billion ($3.8 billion) record as Canadian Imperial Bank of Commerce and Nation Bank of Canada wrapped up the earning season Thursday. The six banks each reported a double-digit jump in net income for a combined profit of C$6.25 billion ($ 4.6 billion) in fiscal 1996 ended Oct 31.

But a third straight year of record profits came aimed growing public anger over perceived high service charges and credit card rates, and tight lending policies.

Bank officials defended the group's performance, saying that million of Canadians owned bank shares through mutual funds and pension plans.

图 1　术语标注示例[1]

综合目前国内外各 ATR 研究，主要有三种思路，基于语言学的方法、基于统计的方法和混合方法[5,10,11]。

2.1　基于语言学的方法

术语一般以名词或名词短语出现，而对特定领域的术语而言，往往具有特殊的词缀（如以“hypen”为前缀的词常为生物医学术语）和特定的组成模式（如很多术语首字母大写）[12]。

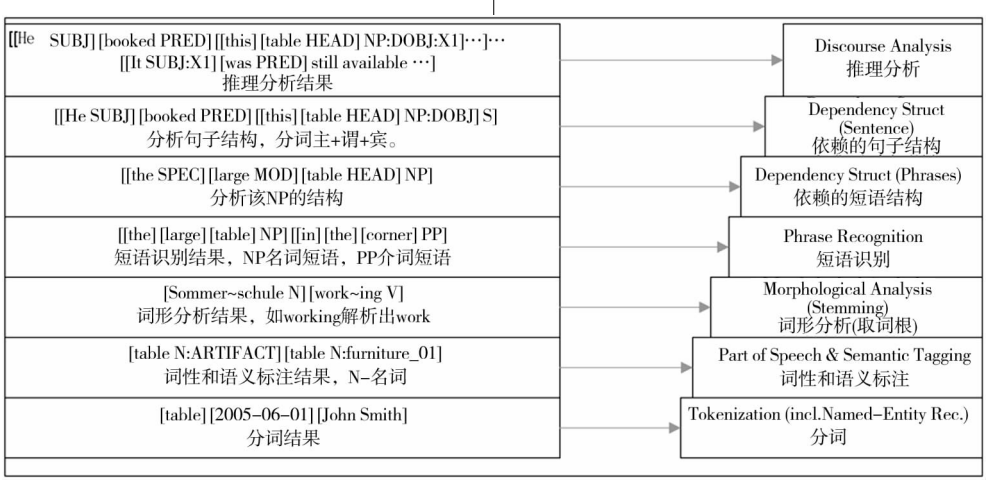

图 2　语言学方法分析术语的层级图

鉴于术语这些词性、词缀和词形等语言学方面的特点，可以利用自然语言处理方法，重复使用术语构成的语法和词形模式判断词串是否符合为术语。如 Paul 等提出利用语言学方法分析术语的层级图（见图 2）[13]。其他如 Ananiadou 等提出基于通用语法的方法，使用统一词形语法、特定词缀字典、拉丁文/希腊文的新古典组合方式表等实现了医学术语识别[12]。但是，这种方法对设定的术语构成模式依赖较大，识别效率有限，在词间关系的识别上尚缺少有力试验的验证。

2.2　基于统计的方法

术语除语言学特征外，还具有一些显著的统计特征，如共现、逆文档词频、熵、互信息等[13]。统计方法判断术语主要依据特征值构建统计模型，查看看词串指定特征值是否符合该模型阈值。例如，根据逆文档词频公式计算词语的权重，将词串按权重结果排序，根

据需要抽取的术语个数等参数选取阈值超过一定术语度①的词作为术语，如 KEA 即以此为构建统计模型的指标之一[14]，当用户需选用 10 个术语时，即取术语度排列前 10 位的短语为最终结果。这种方法不依赖于术语的领域特征，能够较方便地在不同领域使用。但其主要关注多词术语，容易忽略有意义的单词术语，识别效果不尽如人意[10]。

2.3　混合方法

由于上述的两种方法各有缺陷，研究者们通常将两种方法结合使用。有研究者利用统计方法获取初步结果后，再基于语言学方法利用语法过滤器处理统计结果。Samdja 等人通过试验验证过这种方法的可操作性[10]。而更多研究者首先使用语言法方法处理文本获取候选术语，再利用统计模型判断词串是否为术语。这种方法在术语自动识别和聚类（Automatic Term Recognition and Clustering for Terms，ATRACT）[7]、TerMine[15]、关键短语抽取算法（Keyphrase Extraction Algorithm，KEA）等系统中被证实是一种较为有效的思路。

除这三种方法外，也有研究者在探索其它思路，如有作者提出从术语定义入手，利用规则识别出定义后，再定义匹配模板中术语的位置，通过统计方法来判断术语的方法[16]。

3　术语识别的主要困难

ATR 相关研究中，研究者们普遍认为变体术语（Term Variation）、同义术语和多义术语（Term Ambiguity）是影响识别效率的主要困难[6,7,12]。

3.1　术语变体

术语变体即除标准名称外，尚有很多同义词、变体词等指称同一个概念。这些变体的产生与研究者不完全遵守领域术语命名规范有关，同时不断产生的新词和语言的多样性也是产生术语变体的重要原因。具体而言，术语变体主要包括拼写（使用连字号斜线、大小写、不同的拉丁文或希腊语翻译等）、形态（单复数）、词形（同义词）、结构（词语位置变换、介词使用等）及首字母缩略词等。有研究表明，文献中大概有三分之一的术语是术语变体[12]。这些变体影响了识别效率，因此要借助上下文信息等，从语义、语用层，将表达同一概念的术语与指定术语联系起来。

3.2　术语歧义

除术语变体外，术语歧义是术语识别中的又一大挑战。通常情况下，研究者们会利用术语的多面性，用同一术语表达一个概念的不同方面。而有些术语常常在同一个领域的不同研究中表达几个独立的概念。此外，原本为了简化表达的首字母缩写术语在同一领域的不同研究中也经常出现雷同，同一个缩写术语可能表达完全无关的两个概念。这些都向 ATR 提出挑战，ATR 需要利用上下文信息确定术语的真实语义。

实际 ATR 中，这三个问题也是研究者们关注的重点。有系统专门针对单个问题探究解决方案，如 MetaMap 项目利用 UMLS 叙词表概念，有效提高术语变体识别效率[7]。但更多的系统如 ATRACT 针对这三个问题，综合选择适当的方法，提高识别效率。

4　ATR 相关项目与系统分析

在理论研究基础上，研究者们积极试验，开发出众多实用的 ATR 工具，验证各种方法的可操作性和有效性，促进 ATR 理论的不断完善。

4.1　ATRACT

ATRACT[2,7,12] 是生物化学数据库系统项目 Biopath② 的术语管理平台。该系统将术语看作用户信息检索的知识单元，集成信息抽取、分类和知识管理等技术，着重解决术语和术语关系识别两个任务，智能地指导用户在多知识源中发掘知识。

目前，ATRACT 主要处理 HTML 和 XML 文档，其核心模块包括 3 部分：自动术语识别（ATR，又称 C/NC 值模块），缩写词识别（Acronym Recognition Module，ARM）和术语聚类（Automatic Term Clustering Module，ATC）。

ATR 混合语言学和统计学方法，在词性标注等自然语言处理结果基础上，利用 C/NC 值算法，从文本中抽取出多词术语及长术语中内嵌的短术语。依据 C/NC 值，ATC 充分利用上下文信息和统计特征进行术语消歧。考虑到术语一般倾向于亲近相同语义族术语，如果背景词对术语的鉴定有贡献，则认为该背景词与

① 很多研究者探究术语排名近似于术语度。

② 全称：Biochemical Pathways，该项目主要利用公共数据源构建各种生化反应、等级分类和反应网络的数据库，在这些数据基础上自动生成生化反应的反应网络和可视化路径，协助研究者理解反应物间相互关系。

术语含义之间存在显著关联。因此可比较背景词和术语间语义相似度,基于平均互信息的层级聚类算法,实现术语聚类。

整个识别过程中,ATRACT 通过一系列影响术语抽取和结构化的参数实现术语识别。实际使用时,用户可以根据自身需求定制术语识别过程。通过外围检索工具,选用不同参数,用户可以控制检索术语的类型和术语识别的噪音。另外,如果用户在识别过程中指定一个参照术语列表,ATRACT 可仅高亮显示出文档中的新术语,有效连接用户已有知识和新的可获取知识。

ATRACT 良好的模块设计和系统效率在 NaCTeM 很多试验中得到验证。为进一步提高术语识别效率,研究者们仍在探求利用语义、句法结构等信息实现术语识别相关任务。

4.2　TerMine

TerMine 是 NacTeM 开发的术语抽取工具,它将领域无关的 C - value 算法和 NacTeM 开发的 AcroMine 术语缩写识别系统融合在一起,从英文文本中抽取候选术语,且更主要是抽取多词的复合术语[15,17]。

TerMine 选用的 C - value 算法集成了语义分析和统计分析两种方法,尤其侧重统计分析。系统通过词性标注、形容词/名词序列抽取、停用词表过滤等语言学处理方法获取候选术语,进一步统计候选术语词频、更长候选术语的数目等特征,构建术语模型判定术语。根据用户不同需求,TerMine 提供了三种使用途径:网络示范(适用于轻量级的集成系统),批处理服务(处理大于 2MB 的文档),SOAP 服务(在其它应用中集成 TerMine)。通过试验,该系统在处理语料的规模和速度两方面都表现较好。

目前,NaCTeM 已集成了 TerMine 与加州大学伯克利分校开发的 XML 搜索引擎 Cheshire3[18],形成 Cheshire3 - TerMine 系统。用户通过 Cheshir3 可在 MEDLINE 中搜索文摘,找到与检索词相关的术语,根据术语显著程度排列文章文章,再使用 TerMine 标出文中相关术语,提高获取信息效率。另外,NaCTeM 还实现了 Protégé 与 TerMine 的融合,在 Protégé TerMine 插件中,TerMine 工具从文本中抽取出的术语可通过接口快速转换为 OWL 本体,大大简化了本体开发过程。

4.3　KEA

KEA[14]是新西兰 Waikato 大学开发的关键短语抽取开源工具。该系统采用机器学习方法,用户可根据需要在识别过程中选择是否使用词典,识别文中关键短语。

与前两者相似,KEA 同样融合语言学和统计两种方法,在分词、去词根、停用词处理等预处理之后,利用设定规则(包括术语的最大词长、术语首尾词不含停用词表中的词等)筛选出候选术语,再选用逆文档词频 TF × IDF、First Occurrence(该短语在文档中首次出现的位置)、Length(单词个数)、Node Degree(关联词个数)4 个特征,生成统计模型,判断词串是否为关键短语。

4.4　信息技术领域术语辅助提取项目

信息技术领域术语辅助提取项目[19]由北京大学计算语言学研究所和中国标准研究中心合作,主要目的是制定信息科技领域的术语库建设标准,建立领域术语库,开发领域 ATR 软件。

项目组利用信息科学技术语料,提取词串领域特征、语法结构等语言学特征,结合词汇关联度,训练出特征模型,提取出特征相符的候选术语。在此基础上,进一步利用局部上下文、篇章结构等信息,过滤和筛选术语候选词。经过人工校对后将筛选出的术语加入领域词库,结合新的领域词库和通用词库对语料库进一步细切分,循环执行过滤和筛选,提取更高一层结合紧密的语言单位。

实验证明,该方法对发现大量新术语,更新领域术语库有明显帮助。在试验的基础上,该项目组将进一步完善识别策略,提高识别效率。同时将该系统应用于术语信息服务系统中,用户浏览网页时,借助该识别插件,自动获取该网页上所有信息科技领域的术语和术语有关的超链接,帮助用户快速获取相关知识。

5　ATR 相关资源建设

经过数十年研究,研究者们不仅形成了各种理论,构建了众多识别系统,同时还建设了众多可公共获取的术语资源。

5.1　UMLS

UMLS[7,20]是由美国国立卫生研究院开发的术语系统,目前包含 100 多万个概念和 280 多万术语,这几乎包含了所有生物医学词汇。这些词汇经过有效组织,组成 135 个等级类目,其中包含 54 种关系。借助 UMLS 可以将自由词与唯一的概念词相对应,解决术语

变体(UMLS 中主要指缩写、大小写变体)、同义词判别的问题。如 S. B. Johnson 等借助 UMLS 的概念关系利用机器学习方法在消除术语歧义中取得了较好的效果[21]。

5.2 Termino

Termino[7,22]是在 e - Science 框架下的临床医学和 myGrid 两个项目推动下建立起来的。这两个项目都涉及术语识别和分类。由于 UMLS 固定的结构体系和有限的词汇来源限制了术语与多个本体间链接的抽取,项目组促成建立 Termino。该系统由一个装载术语信息的数据库和一个从数据库中识别术语的编译工具组成,术语识别由预处理模块、形态分析、术语记录效用分析 3 个模块组成,预处理部分主要为形态分析提供分词和词性标注结果,效用分析部分则由术语专家参与选择经过形态分析生成的术语集合。该系统着力于建立和维持多种类型资源(本体概念、术语、受控词表、分类描述符等)之间的链接,提供基于词典的标注机制,为大规模、各种来源的生物医学词汇提供弹性的数据库模式。

5.3 CCD

CCD[23]是由北京大学计算语言所开发的与 WordNet 兼容的汉语语义词典。由中英文中的同义词集合定义概念节点,这些节点连通概念之间的上下位关系和一些附加关系等组成了 CCD 的概念网络。目前 CCD 中共包含 25 个名词文档、15 个动词文档及很多形容词和副词文档。这些文档中的各概念及概念间关系的演绎规则都事先严格形式化了,可应用于中文的语义分析中。Hongying Zan 等人借助 CCD 实现了基于词典的单个词语的术语抽取[24]。

6 结　语

尽管目前中英文的自动术语识别在生物医学领域取得了较大进展,但变体术语识别、术语消歧等依旧面临着较多问题。不少方法只是在小的测试集上获得较好性能,大规模的测试仍有待开展。其他领域的 ATR 系统也亟待开发,虽然在信息科技领域已经取得一定成绩,但人工参与的工作还是很多,进一步提高自动化效率也是需要思考的问题。此外,后续的术语管理和维护等工作也需要同步跟进,以确保基于术语的多种应用的有序进行。术语识别作为知识抽取、信息检索、文本挖掘等信息技术的关键环节,要进一步充分汲取这些技术中较为成熟的方法,提高 ATR 效率,促进知识抽取的发展。

参考文献:

[1] Feldman R, Fresko M, Kinar Y, et al. Text Mining at the Term Level[J]. *Lecture Notes In Computer Science*, 1998:65 - 73.

[2] Mima H, Ananiadou S, Nenadic G. The ATRACT Workbench:Automatic Term Recognition and Clustering for Terms[J]. *Lecture Notes in Computer Science*, 2001,2166:126 - 133.

[3] Milios E, Zhang Y, et al. Automatic Term Extraction and Document Similarity in Special Text Corpora[C]. In: *Proceeding of the 6th conference of the Pacific Association for Computational Linguistics*, New York:ACM, 2003:275 - 284.

[4] Love S. Benchmarking the Performance of Two Automated Term - Extraction Systems:LOGOS and ATAO[EB/OL]. [2008 - 04 - 03]. http://www. olst. umontreal. ca/pdf/memoirelove. pdf.

[5] Kajikawa Y, Sugiyama Y. Causal Knowledge Extraction by Natural Language Processing in Material Science:A Case Study in Chemical Vapor Deposition[J]. *Data Science Journal*, 2006,5:108 - 118.

[6] Jensen L J, Saric J, Bork P. Literature Mining for the Biologist: from Information Retrieval to Biological Discovery[J]. *Nature Reviews (Genetics)*, 2006,7:119 - 129.

[7] Krauthammer M, Nenadic G. Term Identification in the Biomedical Literature[J]. *Journal of Biomedical Informatics*, 2004,37(6): 512 - 526.

[8] Asunción Gómez - Pérez, David Manzano - MachoA Survey of Ontology Learning Methods and Techniques [EB/OL]. [2008 - 06 - 05]. http://www. sti - innsbruck. at/fileadmin/documents/deliverables/Ontoweb/D1. 5. pdf.

[9] Term versus Word[EB/OL]. [2008 - 02 - 24]. http://www. termiumplus. gc. ca/didacticiel_tutorial/english/lesson1/page1_2_4_e. html.

[10] Alegria I, Arregi O, Balza I. Linguistic and Statistical Approaches to Basque Term Extraction[EB/OL]. [2008 - 2 - 24]. http://ixa. is. ehu. es.

[11] 于卫. 自动中文术语识别若干方法研究[D]. 哈尔滨:哈尔滨工业大学,2004.

[12] Ananiadou S, Nenadic G. Automatic Terminology Management in Biomedicine[M]. *Text Mining for Biology and Biomedicine*, UK: Artech House Publishers, 2006.

[13] Buitelaar P, Cimiano P, Grobelnik M. Ontology Learning from Text [C]. In:*the ECML/PKDD* 2005 *Workshop on*:*Knowledge Discovery and Ontologies*, Porto, Protugal, 2005.

[14] Olena Medelyna. Automatic Keyphrase Indexing with a Domain - Specific Thesaurus[D]. Germany:University of Freiburg, 2005.

[15] TerMine Plugin for Protege 4 [EB/OL]. [2008 - 4 - 3]. http://www. co - ode. org/downloads/protege - x/plugins/termine - docs. pdf.

[16] 张榕. 术语定义抽取、聚类与术语识别研究[D]. 北京:北京语言大学,2006.

[17] TerMine[EB/OL]. [2008 - 04 - 03]. http://www. nactem. ac. uk/software/termine/

[18] Cheshire3 - Termine Demonstration using Medline Abstracts[EB/OL]. [2008 - 04 - 03]. http://www. nactem. ac. uk/software/ctermine/.

[19] 穗志方等. 信息科学与技术领域术语自动提取研究[C]. 见:第五届东亚术语论坛,2002.

[20] UMLS[EB/OL]. [2008 - 04 - 03]. http://www. nlm. nih. gov/research/umls/.

[21] Liu H, Johnson S B, Friedman C. Automatic Resolution of Ambiguous Terms Based on Machine Learning and Conceptual Relations in the UMLS [J]. *Journal of the American Medical Associations*, 2002, 9(6): 621 - 636.

[22] Harkema H, Gaizauskas R, Mark H, et al. A Large Scale Terminology Resource for Biomedical Text Processing. Linking Biological Literature[J], *Ontologies and Databases*, 2004(6): 53 - 60.

[23] 俞士汶,于江生. 中文概念词典的结构[J]. *中文信息学报*, 2002, 16(4): 12 - 20.

[24] Zan H, Duan G, Fan M. Single World Term Extraction Using a Bilingual Semantic Lexicon - based Approach[C]. In: *Third International Conference on Natural Computation*, ICNC: IEEE Computer Society, 2007: 451 - 456.

（作者 E - mail: liujh@ mail. las. ac. cn）

《现代图书情报技术》特邀专栏组稿

《现代图书情报技术》是中国科学院主管、中国科学院国家科学图书馆主办的计算机信息管理技术方面的学术性刊物。刊物拥有清晰的定位,即以跟踪技术的研究、应用、交流为主体,服务于广大信息技术人员。

本刊从 2004 年起开设不定期栏目——《特邀专栏》,每一期专栏集中发表关于某个特定方面的技术研发与应用的研究型文章,汇集科研成果、聚焦研究前沿。

1　《特邀专栏》目的与定位

对于学术期刊而言,高质量的稿件始终是刊物发展的关键所在。因此,编辑部在广泛组稿的同时,也希望透过业界专家的支持,合作策划重大选题,集中组织优秀稿件,系统深入进行报道。

2　《特邀专栏》操作办法及流程

(1)本栏目特邀国内外知名专家、学者、教授担任专栏主编,专栏的设立一般由期刊的策划编辑和特邀专栏主编沟通,根据国内外图书情报技术学科的发展需要提出选题。

(2)选题一旦确定后,由特邀专栏主编承担稿件的组织,审核并撰写前言。一期特邀专栏一般为 3 - 5 篇文章为宜。稿件组织过程中,策划编辑将与特邀专栏主编进行定期的沟通,及时掌握稿件的撰写情况,并对稿件的撰写提出适当的建议和意见。

(3)稿件经特邀专栏主编审核通过,提交给编辑部。后期由策划编辑负责与作者的联系沟通及安排出版等事宜。

(4)专栏的选题一旦确定后,将确定基本时间表。一般的操作周期为 3 - 5 个月。以正式确定特邀专栏题目为起始点,在 1 个月内确定约请论文的作者和题目,3 个月内确定初稿,5 个月内确定采用稿。

(5)对于拟定录用的特邀专栏稿件,本刊将减免发表费,并支付稿费。稿件一旦发表,编辑部将及时赠与样刊。

3　《特邀专栏》稿件内容要求

(1)深入反映本专栏选题方向的前沿研究成果或重大应用成果,侧重理论研究、技术分析、系统论证或设计等,注意理论与实践相结合。

(2)特邀专栏稿件应该主要是原始性和原创性研究论文,也可以有一篇综述性论文,但综述性论文必须可靠地覆盖该方向的原始核心文献。

(3)文章按照严谨的学术文章体例写作,即明确扼要地界定研究问题,简要说明研究方法,系统精炼地描述国际国内发展状况,进而详细地描述作者自身研究工作的技术线路及研究结果。

(4)特邀专栏的一系列文章应注意覆盖专栏选题所涉及的各个研究方向和多个研究单位,充分覆盖可能存在的多种观点和技术线路。

(5)充分承认前人/别人的工作,充分引证所参考引用的文献(尤其是本研究工作中的原始核心文献和国内最先出现的研究文献),严格遵守著录规范。

4　《特邀专栏》稿件格式要求

(1)论文版式请参照本刊网站"下载专区"中"论文模板"。

(2)多个作者时,请注明通信作者,并注明各个作者的单位。

(3)每篇稿件以 6 - 8 千字为宜(按篇幅字数计算,包括图、表)。

2008 年本刊《特邀专栏》的组稿工作已开始启动,欢迎广大专家、学者给予支持、帮助!