用于海量文献关键词标引的计算机辅助加工系统构建实践*

□ 杨贺 杨奕虹 吴广印 / 中国科学技术信息研究所 北京万方数据股份有限公司 北京 100038
林霄剑 / 北京万方数据股份有限公司 北京 100038

摘要：为缓解海量文献关键词标引的巨大压力，文章构建了用于海量文献关键词标引的计算机辅助加工系统，对标引数据预处理规范、自动标引核心工作区和人工标引校对平台进行了具体阐述。文章采用数据测试方法确定了自动标引软件，在单一软件不能满足标引要求后探索了多种机标结果后处理方式提升机标质量，最终由人工标引校对平台保证海量文献关键词标引质量的同时，将机标出现的问题和改进意见反馈给软件设计和词表维护，保证了计算机辅助加工系统的持续改进。

关键词：文献加工，关键词标引，自动标引，计算机辅助加工系统
DOI：10.3772/j.issn.1673—2286.2013.06.008

1 引言

标引是提高信息检索查全率与查准率的重要方法，尽管现代技术已经可以实现全文检索，但“其检索过程是暗箱操作，局限性很大”[1]，标引预先对文献提取精华、过滤噪音，使检索快速准确，知识挖掘程度深刻[2]，也利于知识发现。然而科学技术的快速发展带动了信息量的剧增，面对爆炸式增长的文献数量，人工标引早已难堪重负。

从上世纪50年代以来，国内外学者研究开发了多种利用计算机进行文献标引的方法，即自动标引，高燕[3]、张静[2]将其归纳为统计分析、语言分析、人工智能三大类型进行优劣势比较，章成志[4]则是从标引自动化程度和揭示知识深度两个维度绘制了自动标引研究路线，李法运[5]撰文总结了近十年来国内外分类主题一体化的自动标引研究进展，余丰民[6]综述了国内2000-2009年十年间的自动标引研究热点。与人工标引相比，自动标引处理能力强、速度快、成本低、稳定性强[7]，但标引质量还不完美，尤其应用海量文献标引的效果还有待提高[3]。

本文研究的自动标引集中在文献关键词标引，考虑到目前叙词表修订滞后与科技文献新词频现的矛盾，本文提到的关键词标引属于不受限于词表的自由标引方式，其用词介于受控词表和自然语言之间[8]，严格选用体现文献主题且满足检索意义的词或词组，并非对关键词毫不加控制，因此其标引质量是有保证的。这也是当前大型文献数据库主题标引的主流模式。

本文利用现有自动标引软件构建用于海量文献标引的计算机辅助加工系统（下文简称机标系统），该系统由三部分组成，分别是标引数据预处理规范、自动标引核心工作区和人工校对平台，其中自动标引核心工作区引入多个不同类型的自动标引软件同时标引，增加多个软件自动标引结果（下文简称机标结果）后处理环节，通过计算机的自动处理进一步提高文献自动标引质量。机标系统工作流程如图1所示，蓝色箭头指引了文献标引过程的数据流向，红色箭头显示了人工智能对自动标引核心模块的反馈作用。下文详细介绍这三部分的具体功能。

2 标引数据预处理规范

本文所指的海量文献不仅仅是指数量上的多，还在于其类型、类别的多样性。在标引数据预处理规范环节中，各种类型和各种学科的

* 本文系国家高科技发展计划（863计划）“云计算一期”重大专项课题“以科技文献为主的搜索引擎研制”子课题（编号：2011AA01A206）成果之一。

文献会经过预先设定好的元数据参数对特征项进行提取后形成待标引文本数据。这是一个著录的过程，而非标引的过程，具体工作包括规范著录数据、识别文献类型和分类特征项、确定待标引数据。表1以学位论文为例说明了这一过程，各环节工作均是为数据进入自动标引前做充分准备。

表1所列“依据”中的《文献机标预案》是这一过程的主要工作，如同检索需要制定检索策略一样，文献加工人员依据来源文献的类型、分类特点以及数量规模等特征，平衡标引质量与标引效率，合理有效地制定自动标引的软件组合、词表以及各参数阈值等预案，必要时进行人工调整。其他辅助工具表涉及词表选择，下文详述。

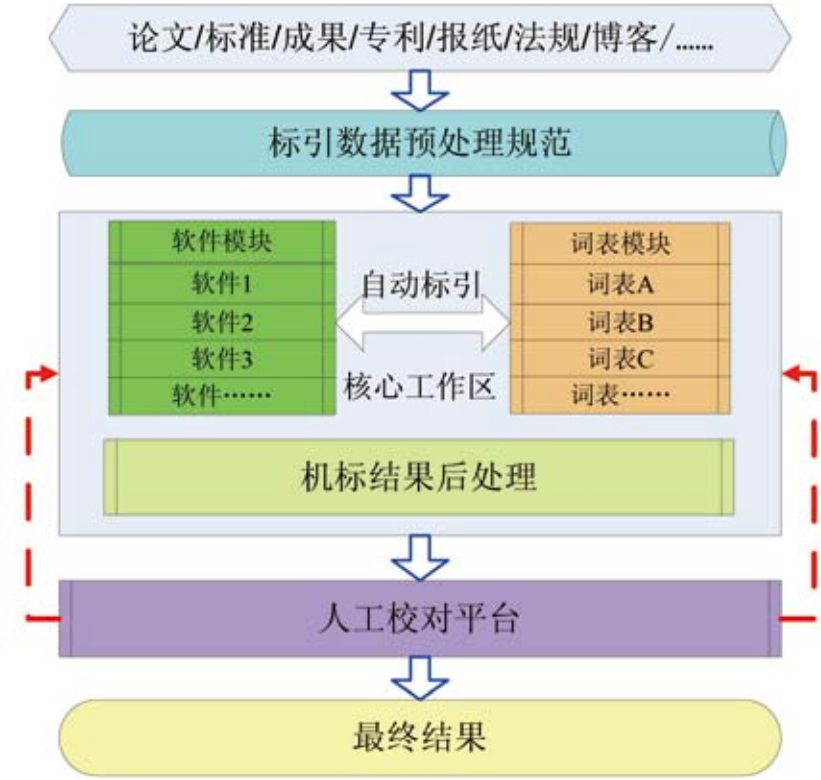

图1 计算机辅助加工系统工作流程图

3 自动标引核心工作区

自动标引核心工作区包括三部分：软件模块、机标结果后处理模块和词表模块。

3.1 软件模块

组建软件模块先期选用了6个相对成熟的自动标引应用软件，因涉及商业秘密，下文均以代号区分，不对具体算法进行赘述。软件1利用分词词典对标引源切词后，基于词频统计计算权重得出最相关的关键词。软件2采用概念语义空间的思想实现了概念语义检索系统。该系统能识别文本类别并从概念层次上理解文本信息，包括文本分类、聚类、文本概念抽取。软件3核心技术为自动分词技术，即针对现代汉语字序列文本，自动分解为词序列文本，进而抽取核心词。软件4为多语言文本挖掘软件，具有自动分类、自动聚类、自动摘要+主题词标引、自动分词等功能。软件5采用知识库算法，软件算法经过训练学习后，可将测试文集中全部有意义的实词进行聚类，形成指定组数的词集合，每组词内部各词彼此关系最近，一般第一组词最突出文本主题。软件6将智能文本处理技术与搜索引擎技术进行融合，能够实现自动抽取文章关键词。如果按类型区分，软件1属于统计分析法，软件2、3、4属于语言分析法，软件5、6属于人工智能法。

3.1.1 测试办法

为了解六个软件的文献标引效果，选用1万篇已由资深标引人员标引过的科技论文作为测试数据，专业范围覆盖理工农医文，偏重于理工。选择论文题名、文摘作为标引源。因各软件对词表要求各异，为公平起见，统一选用这1万篇论文的人工标引关键词集合作为配合词表，要求输出的关键词必须在该词表中。

3.1.2 评价办法

评价方法以人工标引关键词作为对比答案，以理论与人工两种方式作出评价。

理论评价采用著名的Turney[9,10]理论，即计算：召回率（即查全率）R=a/b、准确率P=a/c、F=2*R*P/(R+P)，其中：a——每篇文献中机标词与人工标引相同词的个数；b——每篇文献中人工标引词的个数；c——每篇文献中机标词的个数；F——平衡召回率R和准确率P的参数，F≤（召回率R+准确率P）/2。

表1 数据准备过程样例

提取特征项			目的	依据
步骤	导入内容	导出结果		
规范著录数据	ID 中文标题 英文标题 作者姓名 作者专业名称 导师 授予学位时间 授予学位单位 学位级别 中文关键词 英文关键词 中图分类号 中文摘要 英文摘要 论文页数	ID 中文题名 英文题名 作者 专业名称 导师 授予学位时间 授予学位单位 学位级别 中文关键词 英文关键词 中图分类号 中文摘要 英文摘要 论文页数	实现多种文献跨库统一检索。	《元数据规范标准》
识别文献类型特征项	导师 授予学位时间 授予学位单位 学位级别	学位论文	适配机标软件和词表	《文献机标预案》
识别文献分类特征项	作者专业名称 中图分类号	行业类别 中图大类	确定文献分类特征，适配词表	《学科专业-行业词表对照表》 《中图大类-行业词表对照表》
确定待标引数据项	中文标题 英文标题 中文摘要 英文摘要 中文关键词 英文关键词	中文题名 中文摘要 中文关键词 英文题名 英文摘要 英文关键词	确定进入软件模块的源数据	软件模块参数要求

本文定义的a严格遵循Turney理论，即机标词与人工标引词以完全相同数作为a，未作同义词、近义词、上下位类词、部分匹配词的比较。国内有些研究者将a的定义做了扩展，采用了相似词或部分匹配的概念[10,11]。作者认为拥有用代关系的词判定为相同词可以接受，但两者若是属、分、族、参或部分包含的关系，则不能判定为相同词。这是基于中文科技文献标引的要求，也是汉语词语重心后置的特点决定的，例如应标引“超低碳不锈钢”却只标引了“不锈钢”；应标引“凹板印刷机械”却标引了“凹板印刷”；还有在化学领域分歧更为显著，多一个字的组合化学物质就变化了，如“吖啶”和“吖啶酮”。这样的“部分相同”不符合文献标引要求，是不可取的。

六个软件的机标结果与正确

答案比较，计算每篇文献的R、P、F值，得出全部数据的平均R、P、F值，结果见表2。

在人工评价方面，仅提交理论评价中排在前三位的机标结果供专家评价，鉴于1万条数据人工判断工作量太大，本文随机选取了《中国图书资料分类法》类号为F（经济）、P（天文学、地球科学）、R（医学、卫生）、TP（自动化技术、计算机技术）四类专业，各250篇文献提交给擅长这四个专业的标引专家评价，要求专家对每篇论文的三个机标结果给予好、中、差的评价。总计1000篇论文的人工评价结果如表3所示，总体上软件1得到的好评较多，占到69.1%，其次是软件2；而以最不能接受的程度来看，软件6占28.0%。

经访谈，四位专家普遍认为软件6的结果精度高，但是漏掉词过多，而软件1和软件2虽然相对全面，但冗余偏多，不够精炼。这一结果与理论评价结果一致，软件1、2查全率高，软件6查准率高，在当前情况下可采用为系统自动标引软件。

3.2 机标结果后处理

为进一步提高自动标引的质量，受编辑出版行业中的交叉式校对思想的启发[12]，系统增设了机标结果后处理环节，将多个机标结果进行交叉融合，扬长避短。本文通过数据实际测试来阐述这一环节的工作办法。

为模拟真实海量文献标引过程，作者仍以上述1万篇科技论文作为标引对象，但改由一个数量约为7万的综合词表进行配合标引。该词表综合了已出版的《汉语主题词表》和理、工、农、医等多个行业主题词表，并加入了标引加工人员在一线工作中积累的常用于标引的关键词。各软件机标结果理论评价值见表4。

与表2相比，表4的各项值都下降明显，可见自动标引中所用词表的影响是重大的，这也是系统在数据预处理阶段识别文献类型和分类特征，以期应用行业词表对口标引的初衷。下节将阐述词表模块的设置。

本文对三个软件的机标结果

表2 软件理论值排序结果

软件	F值↓	R值	P值
软件1	0.49	0.63	0.41
软件6	0.49	0.48	0.56
软件2	0.46	0.61	0.39
软件4	0.34	0.47	0.28
软件3	0.30	0.41	0.24
软件5	0.28	0.47	0.23

表4 人工评价统计结果

软件	F值↓	R值	P值
软件1	0.21	0.30	0.17
软件2	0.18	0.26	0.14
软件6	0.17	0.12	0.28

表3 人工评价统计结果

论文类别	软件1结果			软件2结果			软件3结果		
	好	中	坏	好	中	坏	好	中	坏
F（经济）	206	44	0	205	45	0	97	129	24
P（天文学、地球科学）	99	110	41	118	105	27	75	122	53
R（医学、卫生）	234	15	1	178	72	0	28	139	83
TP（自动化技术、计算机技术）	152	86	12	72	158	20	33	97	120
合计	691	255	54	573	380	47	233	487	280
比例	69.10%	25.50%	5.40%	57.30%	38.00%	4.70%	23.30%	48.70%	28.00%

图2.1 两两交集取并集

图2.2 两者并集与第三个结果取交集

图2.3 两者交集与第三个结果取并集

表5 组合结果后处理统计表

组合形式	具体方式	F值↓	R值	P值
图2.3	软件1∩软件2∪软件6	0.56	0.63	0.53
图2.1	两两相交取并集集	0.34	0.30	0.46
图2.2	软件1∪软件2∩软件6	0.33	0.24	0.59
图2.2	软件2∪软件6∩软件1	0.30	0.25	0.44
图2.2	软件1∪软件6∩软件2	0.29	0.24	0.42
图2.3	软件2∩软件6∪软件1	0.24	0.35	0.19
图2.3	软件1∩软件6∪软件2	0.22	0.32	0.17

进行交叉融合处理，方案见示意图2.1、2.2、2.3，紫色区域为最终所需结果，统计结果见表5。

由表5数值可看出，所有的组合方式F值均高于表4所列单一软件的F值，而其中软件1和2取交集后再与软件6取并集的组合方式使F值提高了2倍多，证明采用多机标结果后处理是提高自动标引质量的有效方法。尽管这种方法的结果仍然不能与人工标引质量相媲美，但就效率和成本来看确实是一种良好的改进方式。

3.3 词表模块

实际数据测试表明，词表在关键词自动标引中起到了重要作用，本系统中包括的词表按功能可分为三类：

第一类词表是控制标引关键词质量的词表，包括传统意义上的叙词词表，又称主题词表，以《汉语主题词表》最为著名。常春[13]、鲍秀林[14]撰文梳理了建国后国内编制的叙词表，其中90%为专业型词表，且大部分集中在自然科学和工程技术领域，遗憾的是近九成词表编制后从未进行过修订[14]。另一部分是多年来标引加工人员积累的词单。随着科技的迅速发展，传统的叙词表在词汇数量和新鲜度上都难以满足海量文献的标引加工，张琪玉[15,16]老先生在上世纪90年代就曾呼吁大量编制自然语言词表，作者在这方面也进行了初步研究，另文撰写[17]。

第二类词表实现了主题分类一体化，将词表与分类法融合在一起，同步完成主题标引和分类标引。如《中国分类主题词表》，其第二版[18]已于2006年出版，它实现了《中国图书资料分类法》和《汉语主题词表》两大检索语言的一体化。但遗憾的是，这种一体化的工具还未在其他分类法里实现，这也是作者下一步的研究方向。

第三类词表是配合自动标引软件工作的词表，如分词（或称切词）词表（或称词典）、停用词（或称禁用词）词表、敏感词词表等。它们进一步保障了标引质量，并对标引结果进行有意识的调整。

除此之外，机标系统中还设置了与词表配合使用的工具表，如前文提到的学科专业-行业词表对照表、中图大类-行业词表对照表等。这类工具表通过识别待标引文献分类特征，即便仅仅是一级大类，也

能够实现缩小词表适配范围、提高行业词表适用性的目的。比如医学类期刊配合医学行业词表，再如学位论文，可通过其学科名称识别行业特征，配合相关行业词表标引。

在实际海量文献标引工作中，作者也发现了一个趋势，即学科之间的交叉愈来愈明显，所谓“他山之石可以攻玉”，许多几年前仅应用于某一行业文献的“专有”标引词很快就横跨多个领域文献，变为通用标引词[19]，比如数值模拟、可视化等等。这也是科技发展的必然趋势，因此需要重视行业词表的及时更新问题。

综上所述，词表模块在整个机标系统中不可或缺，尤其是在当今网络环境下词表的应用对于知识组织、知识检索都起到了良好的规范、引导作用。

4 人工标引校对平台

设置人工校对平台有两个目的，一是校对机标结果，保证文献标引最终质量；二是将校对审核过程中发现的问题和改进意见反馈给软件模块和词表模块。平台功能具体表现在三方面：

（1）审核修改功能：标引员参考文献的标引源数据（如题名、摘要等）信息对机标结果加以修改。与之前人工标引相比，这种转变好比自己写答案变为批改答案。实践证明，当机标结果质量较好时，标引员工作量确实有所减轻，但是当机标结果质量较低时，例如机标系统设置初期使用单一软件机标结果测试时，标引员普遍反映修改量太大，还不如直接“写答案”快捷。多机标结果后处理环节解决了这一问题。

（2）查词功能：将词表模块中的词表集中，供标引员随时检索查询。该功能可围绕检索词提供上、下属及前后方一致相关词的同时，还可以提出新词和禁用词的警示。该功能可帮助业务尚不熟练的初级标引员迅速进入“状态”，也减轻了中高级标引员的工作压力。

（3）统计对比功能：在经过一定数量的人工与机标结果对比统计后，机标软件设计人员可从中总结规律发现问题，调节阈值、改进算法；词表维护专家根据文献类型和专业分类特点，补充新词、淘汰坏词、调整词间关系，不断修订相关词表。

人工校对平台在当前海量文献自动标引水平的情况下恐怕还要坚持很长的一段时间，虽然一定程度上增加了标引员的工作量，但从长远来看却是“磨刀不误砍柴工”。

5 结论及下一步工作

计算机辅助文献标引加工系统的应用大大缓解了海量文献标引压力，在单一软件机标质量不高的情况下，系统通过增加标引数据预处理规范、多机标结果后处理达到了提升机标质量的目标，并通过人工校对平台的反馈实现了系统持续改进的目的。在今后的实际应用中，为实现海量文献完全自动标引的理想目标，还需要通过软件改进、词表修订和人工标引智能三方面的共同协作完成。

随着人类对知识需要的多样化和差异化，届时计算机文献加工标引系统的功能将不再仅仅作为减轻人工标引压力的替代品，而是结合信息检索系统实现多层次、多角度、多用途的知识组织系统工具，为用户提供更好的知识服务，这也是作者下一步努力的研究方向。

参考文献

[1] 张琪玉.文献标引是需要智慧的近乎艺术创造的处理过程[J].图书馆杂志,2004,23(3):24-25.
[2] 张静.自动标引技术的回顾与展望[J].现代情报,2009,29(4):221-225.
[3] 高燕.关键词自动标引方法综述[J].电子世界,2012(6):118-120.
[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).
[5] 李法运.国内外分类主题一体化标引和检索系统研究进展[J].图书情报知识,2004(2):19-22.
[6] 余丰民.2000～2009年国内自动标引研究综述[J].情报探索,2011(5):28-31.
[7] 马张华.论自动标引的实际应用[J].图书情报工作,2003(2):48-51.
[8] 董丽,侯汉清.中文期刊文献关键词标引的分析和改进[J].情报科学,2004,22(11):1355-1358.
[9] TURNEY P D. Extraction of Keyphrase from Test: Evaluation of Four Algorithms [M]. Technical Report ERB-1051, National Research Council, Institute for Information Technology, 1997.
[10] 张庆国,薛德军,张振海,等.海量数据集上基于特征组合的关键词自动抽取[J].情报学报,2006,25(5):587-593.
[11] 章成志,周冬敏.自动标引通用评价模型研究[J].情报学报,2009,28(1):40-47.
[12] 许丽玉.“交叉式编校合一”:以“交叉校对”为条件的“编校合一”[J].中国科技期刊研究,2009,20(5):954-956.
[13] 常春,卢文林.叙词表编制历史、现状与发展[J].农业图书情报学刊,2002,(5):25-28.
[14] 鲍秀林,吴雯娜.中文叙词表发展概况和性能测评(1980-2009)[J].图书馆论坛,2012(4):101-106.
[15] 张琪玉.积极为自然语言与情报检索语言的结合创造条件——建议大量编制自然语言词表(上)[J].图书馆杂志,1999(9).
[16] 张琪玉.积极为自然语言与情报检索语言的结合创造条件——建议大量编制自然语言词表(下)[J].图书馆杂志,1999(10).
[17] 杨贺,杨奕虹,乔晓东,等.用于计算机辅助文献标引加工系统的自然语言词表构建[J].现代图书情报技术，2010(6).
[18] 国家图书馆《中国图书馆分类法编辑委员会》.《中国分类主题词表》（第二版）及其电子版手册[M].北京:北京图书馆出版社,2006.
[19] 赵妍,侯汉清.中文期刊文献通用词标引分析[J].图书与情报,2007(1):63-65.

作者简介

杨贺（1978-），女，毕业于中国科学技术信息研究所情报学专业，硕士，馆员，现在中信所万方数据公司工作，研究方向：文献加工、数据库加工、词表建设。E-mail: yanghe@wanfangdata.com.cn
杨奕虹（1965-），女，中国科学技术信息研究所情报学硕士，研究馆员。研究方向：信息管理、知识组织、数据库建设。
E-mail: yangyh@wanfangdata.com.cn
吴广印（1965-），男，中国科学技术信息研究所研究员。研究方向：非结构数据库管理系统与中文信息检索，RMS系统的总体设计师和主要开发人员。本文“863”专项的技术负责人。E-mail: gywu@wanfangdata.com.cn
林宵剑（1983-），男，北京万方数据股份有限公司数据库工程师。研究方向：数据库深度挖掘、海量数据精细处理。E-mail: linxj@wanfangdata.com.cn

Building Practice of Computer Assistant Processing System for Massive Literature Keywords Indexing

Yang He, Yang Yihong, Wu Guangyin / Institute of Scientific and Technical Information of China; Beijing Wanfang Data Co., Ltd., Beijing, 100038
Lin Xiaojian / Beijing Wanfang Data Co., Ltd., Beijing, 100038

Abstract: In order to alleviate the enormous pressure of the massive literature keywords indexing, the paper builds a computer assistant processing system for massive literature keywords indexing, and three modules are exposited, including Indexing data preprocessing specification, Automatic indexing core workspace and manual indexing proofreading system. The paper selects automatic indexing softwares by data testing, and discovers several combination modes to improve the quality of automatic indexing. The function of manual indexing proofreading system is designed to enhance the literature indexing quality and the whole system as well.

Keywords: Literature processing, Keywords indexing, Automatic indexing, Computer assistant processing system

（收稿日期：2013-04-17）