大学数字图书馆国际合作计划
CHINA ACADEMIC DIGITAL ASSOCIATIVE LIBRARY

CADAL
项目标准规范丛书

CADAL数字图书馆知识标准规范及应用研究

潘云鹤◎丛书主编
刘柏嵩◎编　　著

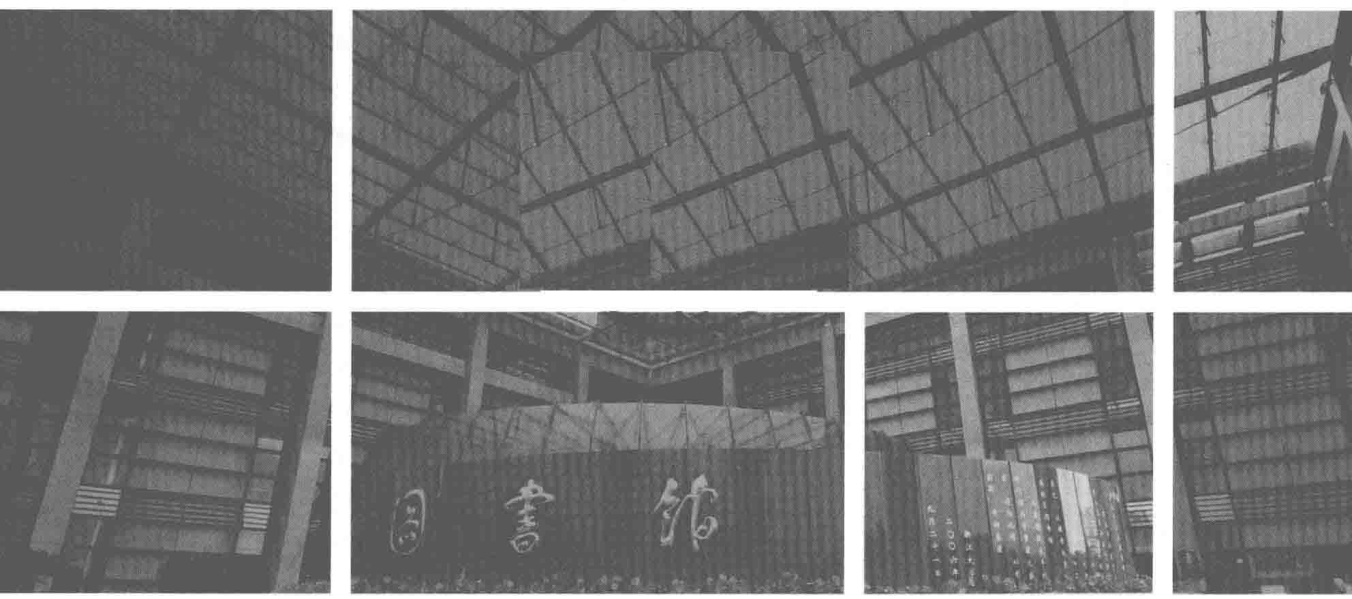
ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

大学数字图书馆国际合作计划
CHINA ACADEMIC DIGITAL ASSOCIATIVE LIBRARY

CADAL
项目标准规范丛书

CADAL数字图书馆知识标准规范及应用研究

潘云鹤◎丛书主编
刘柏嵩◎编　　著



ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目（CIP）数据

CADAL数字图书馆知识标准规范及应用研究 / 刘柏嵩编著. —杭州：浙江大学出版社，2017.9
（CADAL项目标准规范丛书 / 潘云鹤主编）
ISBN 978-7-308-16885-4

Ⅰ.①C… Ⅱ.①刘… Ⅲ.①院校图书馆—数字图书馆—著录规则—研究 Ⅳ.①G258.6 ②G254.31

中国版本图书馆CIP数据核字(2017)第097493号

CADAL数字图书馆知识标准规范及应用研究

CADAL Shuzi Tushuguan Zhishi Biaozhun Guifan Ji Yingyong Yanjiu

刘柏嵩　编著

责任编辑　张凌静(zlj@zju.edu.cn)
责任校对　冯其华
封面设计　续设计
出版发行　浙江大学出版社
（杭州市天目山路148号　邮政编码310007）
（网址：http://www.zjupress.com）
排　　版　杭州中大图文设计有限公司
印　　刷　杭州日报报业集团盛元印务有限公司
开　　本　787mm×1092mm　1/16
印　　张　12
字　　数　300千
版 印 次　2017年9月第1版　2017年9月第1次印刷
书　　号　ISBN 978-7-308-16885-4
定　　价　49.00元

总 序

春华秋实，“CADAL项目标准规范丛书”即将正式出版了。该丛书不仅汇编了历年来CADAL项目标准规范建设的成果，而且包含了CADAL人在标准规范建设中的经验总结和理论探索，是CADAL项目建设的基石，也是CADAL项目成果中最璀璨的明珠之一。

CADAL项目的建设始于2000年。那时，全球的数字图书馆建设热潮初起，包括卡内基梅隆大学、中国科学院和浙江大学等机构的中美两国计算机科学家共同发起了“中美百万册书数字图书馆合作计划（China-US Million Book Digital Library Project）”，目标是建设一百万册电子图书，并逐步向用户提供服务。2000年12月，“百万册书计划”项目启动，项目定名为“高等学校中英文图书数字化国际合作计划（China-America Digital Academic Library，CADAL）”。

2002年9月，国家计划委员会、教育部、财政部在《关于“十五”期间加强“211工程”项目建设的若干意见》的文件中，将“高等学校中英文图书数字化国际合作计划（CADAL）”列入“十五”期间“211工程”公共服务体系建设，CADAL与“中国高等教育文献保障系统（China Academic Library & Information System，CALIS）”一起，构成了中国高等教育数字化图书馆的基本框架。2005年11月，作为首个有外资参与的“211工程”项目，CADAL完成了百万册资源的加工工作，拥有了较为齐全和完整的数字化特藏，初步建成了中国高等教育数字化图书馆。作为当时全球最大的公益性数字图书馆，CADAL不仅数据量大，而且学术性强、开放程度高，100万册图书向全世界开放提供数字化信息服务，CADAL一期建设圆满完成。

2008年正值我国数字图书馆建设规模、数字资源的海量管理乃至数字图书馆技术皆迅猛发展的时期。为了更好地实现“构建拥有多学科、多类型、多语种海量数字资源的，由国内外图书馆、学术组织、学科专业人员广泛参与的，具有高技术水平的学术数字图书馆”的项目建设总体目标，教育部决定继续投资CADAL建设，并将二期更名为“大学数字图书馆国际合作计划（China Academic Digital Associative Library，CADAL）”。建设方案几经酝酿、修改，2009年8月，CADAL项目二期可行性研究报告通过了教育部专家的评审和论证。作为“211工程”“高等教育文献保障体系”的两大专题之一，2010年4月，项目二期正式启动，并于2012年5月通过项目二期验收。通过一、二期十年的建设，CADAL项目以250万册的数字资源总量，继续保持国内外公益性数字图书馆规模的领先地位，实现了对科学技术与文化艺术的多种类型媒体资源的数字化整合，可以向读者提供一站式个性化知识服务，成为国家创新体系中重要的学术信息基础设施之一。

CADAL的建设，见证了中国数字图书馆事业的成长历程：从蹒跚学步到健步如飞，而今不仅风华正茂，而且更充满了自信和创新的活力。

标准为人类文明的发展提供了重要的技术保障，CADAL标准规范研究始于建设之初，

前　言

标准规范的建设，尤其是在开放和可互操作基础上的标准与规范建设，是数字图书馆建设高效、经济、可持续的根本保证，是数字图书馆能够长期发挥作用的必要条件。忽略数字图书馆标准规范体系建设，将会导致资源重复开发，影响资源共建共享，限制数字图书馆的作用空间和发展能力。

“大学数字图书馆国际合作计划(China Academic Digital Associative Library，CADAL)”与“中国高等教育文献保障系统(China Academic Library & Information System，CALIS)”一起，共同构成中国高等教育数字化图书馆的框架，目的在于构建拥有多学科、多类型、多语种海量数字资源高技术水平的学术数字图书馆。CADAL 项目建设的数字图书馆，提供一站式的个性化知识服务，将包含理、工、农、医、人文、社科等多学科的科学技术与文化艺术，包括书画、建筑工程、篆刻、戏剧、工艺品等在内的多类型媒体资源，数字化整合后通过因特网向参与建设的高等院校、学术机构提供教学科研支撑。面对分布、异构、变化和开放的数字信息资源与服务环境，CADAL 数字图书馆需要建立自己的标准与规范描述体系，按照统一原则、框架和基本方式，规定应遵循的各个层次的标准与规范，从而支持在整个数字信息环境中有效使用、广泛获取和长期保存信息。

CADAL 知识标准规范及应用体系主要由数字资源知识标引标准规范、学科资源自动分类标准规范、数字资源学术水平自动切分标准规范、数字图书馆知识服务标准规范等几部分组成。

目　录

图表索引

第 1 章　文献标引概论

1.1　引　言

进入 21 世纪，随着计算机技术、网络技术、通信技术的高速发展，信息的存储和处理能力得到迅速提高，信息量持续增长，纸质文档被不断转变为电子文档，可以说，我们正在被呈几何量级产生的信息所淹没。但大量的信息却因为没有经过挑选、加工、整理、解释而埋没在数据的海洋中，不能成为有效的知识呈现在用户面前，从而使用户陷入“信息泛滥、知识匮乏”的局面。

要转变这种局面，迫切需要各专家学者对信息与知识之间的转化进行研究，把信息组织转变为知识组织，把信息资源管理转变为知识资源管理，把信息服务转变为知识服务，把等级式的组织结构转变为网络式的组织结构，把以文献为单位的传统标引转变为以知识为单位的知识标引。①

虽然数字图书馆可以通过标题、作者、主题词、关键词、文摘，甚至全文进行检索，但现有的知识组织对象还长期停留在文献级别上。然而文献只是知识的载体，而非知识本身。因此，检索出来的文献虽然包含用户所需要的知识，但它并不能揭示出知识之间的联系，只有把知识的组织方式深化到知识元级别上，实现知识元的链接，才能快速地为用户提供全面、准确的知识，这可以说是知识管理上的一场革命，它涉及对知识的理解。知识的机器发现、自动抽取、语义表示，知识的检索和利用，是信息服务向知识服务转变的基础。更为重要的是，在知识组织与管理的知识链中，知识标引与检索也存在相应的知识链，这种知识链是由知识元、知识单元和知识系统组成的知识标引与检索系统。人们把在科学研究中取得的发现、发明、经验、教训等知识创新点组织成知识单元——文献，知识标引工作则是从已发表的文献中提取出作者的知识创新点——知识元。② 知识标引的目的是让用户直接有效地使用知识单元中的知识元，而不是文献。知识标引起着对文献中的知识过滤和使文献中的知识元重现的知识增值的作用。③

传统的以文献为单位的标引制约了用户更有效地利用知识，成为当前知识管理的瓶颈。

首先，不同的文献之间存在着质量差异，质量越高的文献，被利用率越高，但这类文献在数量上往往并不多。传统的标引和检索方式可以在一定程度上让用户从海量的文献中找出高质量的文献，但是需要花费大量的时间与精力，而通过对参考文献的剔出和再加工，建立

---

① 原小玲. 基于知识元的知识标引[J]. 图书馆学研究，2007(6)：45-47.

② 温有奎，徐国华. 知识元链接理论[J]. 情报学报，2003，22(6)：665-670.

③ 温有奎，徐国华，赖伯年，等. 知识元挖掘[M]. 西安：西安电子科技大学出版社，2005：16.

文献之间的相互关系(新的知识),这种面向知识发现的标引方法,能很好地解决上述问题。

其次,不同的文献所包含的创造性不同,有的文献所包含的创造性不够,很快会被淘汰;有的文献所包含的创造性很强,在相当长的时间里都会被他人所用。也就是说,文献中知识的创造性是文献知识利用的核心,若文献中的知识用知识元单位来标引,则将大大地加快知识利用的速度。

最后,在传统的标引中,图书馆虽然采集了相关文献,但并没有对文献中所包含的知识进行具体分析,用户因此不能直接获得可以解决问题的知识点。对一项研究而言,通过文献标引技术,学者可以查到几十甚至几百篇相关文献,但是哪些文献是可以解决问题的,学者需要通读所有文献的内容,才能有一个答案,有时甚至读完了所有文献,可能还是无法找到解决问题的知识点,这样既浪费精力又浪费时间。而采用知识标引,就可以帮助用户更快地找到所读文献的知识所在,将用户从海量的文献中解脱出来。

1.2 知识标引

1.2.1 知识标引的定义

标引(indexing),简单说是一种标识和引导,是对文献是什么信息的描述;具体是指在分析文献内容或情报问题的基础上,用某种索引语言或标识符号把文献的主题概念及其他有检索意义的特征标识出来,作为情报存储和检索的依据的处理过程。①

传统标引是以文献为单位的标引,它依据文献的外部特征,如文献的标题、作者、出版时间、出版社或刊物名称等来进行标引,或者依据文献的学科分类进行标引,或者依据文献的主题词进行标引。传统标引虽可以为用户提供检索的依据,但并不能提供准确的知识信息。如何实现传统的信息服务向知识服务的转变,是我们下一步研究的重点。

知识标引是实现知识组织、知识检索的核心,是实现信息服务向知识服务转型的关键技术。实现知识标引,知识才可能被人类有效检索、利用和再创造,起到知识增值的作用,信息服务才可能真正转变成知识服务。② 知识标引是以知识元为单位的标引,它依据文献本身的内容,即知识本身来进行标引,对文本内容进行知识挖掘,最终为用户提供更加准确的知识信息。

知识标引的基础是知识元。知识元是构成知识的最小单位,是文献中的概念、数据、公式、图表、定理、模型、结论等,是构造知识系统的基础。知识元的不同排列组合可以组成不同的知识单元,不同的知识单元按照不同的逻辑关系可组成不同的知识元链接,这是一个知识学习的过程,同时也是一个信息转换为知识的过程。另外,知识标引过程既可以体现出知识结构的背景,又可以体现出知识的创新点,这是一个知识增值的过程。因此,利用知识标引进行知识检索时,用户不仅可以通过知识单元间接获取知识,而且可以通过知识元直接获取知识,这就大大地提高了知识的利用率,从而实现了为用户提供知识服务的目标。

---

① 付蕾. 知识元标引系统的设计与实现[D]. 武汉:华中师范大学,2009.

② 原小玲. 基于知识元的知识标引[J]. 图书馆学研究,2007(6):45-47.

1.2.2　知识标引的分类

标引按照使用的标引语言或标识符号的类型，可分为分类标引和主题标引；按照使用的标引设备，可分为手工标引和自动标引。

1.2.2.1　分类标引

分类标引，又称文献分类或信息分类，是依据特定的分类规则，对文献进行分类标识的过程。分类标引的过程，就是根据已经选定的分类规则，对标引对象的特征进行分析，在确定标引对象所属的类目后，将所要表达的相关信息，用对应分类法中规定的符号代码表示出来的过程。[①] 简单来说，就是按照规则把某些具有共同特征的信息聚类在一起，并依据信息间的关联关系把它们组成一个条理清晰、层次分明的整体的过程。经过分类标引，可以将大量的文献分门别类，纳入特定的分类体系，使得对于原本无序的文献，可按照特定的分类体系对其进行分类标识，使其组成一个有序的学科体系。分类标引还能较好地体现出知识的系统性，把同一领域的知识集中在一起，将不同的区分开来，从而满足了用户按专业领域进行检索的需要。[②]

1.2.2.2　主题标引

主题标引，是依据特定的主题语言，赋予文献主题标识的过程。主题标引所依据的主题语言可以是标题词语言、叙词语言、关键词语言等。因此，主题标引赋予文献的主题标识可能是标题词、叙词、关键词等。通过主题标引，人们可以把同一主题的相关信息聚类在一起，并按照规定的顺序排列起来。主题标引是对标引对象进行主题分析，在确定标引对象的主题概念后，按照一定的词汇控制方式，对标引对象赋予恰当的语词标识的过程。与分类标引相比，主题标引可以集中有关一个主题的各种信息，有较强的直观性、专指性和适应性。[③] 主题标引一般有两类标引方式，一种是自由标引方式，这种标引方式是标引人员直接从已有的描述标引对象信息特征的语句中选取主题词作为标识，这种方式对标引人员的专业化程度要求较高；另一种是词表标引方式，这种标引方式是从已制定好的各类主题词表中选择相关的语词作为标识，这种方式对主题词表的维护要求较高。

用主题标引文献确实可取得不错的效果，但也存在问题：一是主题词存在不连贯性，使得使用者很难直接从主题词中较准确地获得文献的主题；二是当主题词数量偏少时，标引效果就会受到影响。在这种情况下，情报界提出了主题概念标引，它对文献的主题概括能力较强，可以使标引的效果增强。

目前，获得概念主题词的方法主要有三类：一是在某个主题词在概念层次中没有直接的同义词或准同义词的情况下，直接选取上位词作为主题概念；二是在某个主题词在层次概念词典中有若干直接同义词且这些同义词在文章中也出现的时候，通过聚类产生上位词作为

---

① 冷伏海. 信息组织概论[M]. 北京：科学出版社，2008.

② 付蕾. 知识元标引系统的设计与实现[D]. 武汉：华中师范大学，2009.

③ 冷伏海. 信息组织概论[M]. 北京：科学出版社，2008.

主题概念；三是在若干主题词同时出现在文章的标题或正文的某些字段中的情况下，将两个（或以上）主题词合成生成主题概念。[①]

1.2.2.3 手工标引

手工标引的基本流程为：①阅读文献；②分析文献内容；③提取主题概念；④表达主题概念；⑤使表达规范化；⑥编制索引目录；⑦编辑为索引和文档。

与自动标引相比，手工标引存在很多的弊端，概括起来有以下几方面：

（1）一致性差。人具有主观性，所以不同的标引人员在标引同样的文献时也可能会有不同的结果，这使得手工标引在标引一致性方面存在较大的缺陷。

（2）技术性强。手工标引属于一项技术性较强的工作，对标引人员的专业要求较高，标引人员不仅要具有图书情报理论基础，而且要具备较强的专业素质。

（3）效率较低。手工标引需要标引人员在浏览全文后，才能找出文献的主题信息，并对其进行标引，因此手工标引的速度很难大幅度提高，效率较低。

1.2.2.4 自动标引

与手工标引相比，自动标引具备较好的优势。自动标引是指利用计算机从文献中自动提取相关知识引导的过程。

自动标引的基本流程为：①获得文献文本，以准备标引，此文本须转化为机读式文献；②语句分析；③词语加权；④确定标引词的权值；⑤选出标引词；⑥把标引词转换为受控词；⑦文档生成与索引编辑输出；⑧根据反馈信息，再进行词相关加权计算，以提高标引质量。

按照标引词来源的不同，自动标引可以分为自动抽词标引和自动赋词标引。

自动抽词标引指的是由计算机直接从原文中自动抽取词或者词语作为标引来描述文献的主题内容。它涉及如何从文献中抽取出可以表达其实质意义的词语，并根据这些词汇确定标引词。[②] 后文所讲到的知识抽取就属于此类标引。

自动赋词标引指的是使用预先编制好的受控词表，先取词语对文献进行标引。它涉及如何编制受控词表来反映文献内容中的关键词。后文所讲到的学科文献学术水平等级切分就属于此类标引。

1.3 自动标引方法

近几年来，随着信息技术的快速发展，用户所面向的知识源越来越庞大，对信息的需要也越来越个性化，要充分挖掘文献中所含的知识内容，手工标引技术已远远不能满足用户的需要了。自动标引技术以它的快速性、准确性以及再创造性，愈发得到学界的重视。一种好的自动标引方法的出现，可以大大地提高标引的准确率，为用户提供更优更好的服务。以下着重介绍自动标引方法。

---

① 韩客松，王永成. 中文全文标引的主题词标引和主题概念标引方法[J]. 情报学报，2001，20(2)：212-216.

② 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.

1.3.1　自动标引方法研究状况

自动标引的研究至今大致经历了三个阶段：

第一阶段是 20 世纪 50 年代至 90 年代初。这个阶段主要是关于关键字提取方法的研究。

第二阶段是 20 世纪 90 年代至 90 年代末。这个阶段传统的自动标引方法的效率已达到极限，因此自动标引方法的研究进入低谷。

第三阶段是 20 世纪 90 年代末至今。这个阶段计算机及网络技术迅速发展，用户需求不断提高，因此自动标引方法的研究进入了繁荣期。

1.3.1.1　国外自动标引方法研究状况

国外对自动标引的研究最早始于 20 世纪 50 年代，经过了 60 多年的发展，取得了较多的成果。

1957 年，卢恩(Hans P. Luhn)开始了自动标引研究，他最早将计算机技术应用到了文献标引领域，开创了计算机自动标引的先河。卢恩以 Zipf 定律为其理论基础，采用了以词频为特征的统计标引方法。该方法的优点是简单易行，且具有一定的客观性和合理性，因此在自动标引中占据重要地位。

1958 年，卢恩提出了基于绝对频率加权法的自动标引方法。

1958 年，巴克森代尔(Phyllis B. Baxendale)提出了从论题句和介词短语中自动提取关键词的方法。

1959 年，埃德蒙森(Harold P. Edmundson)等提出了基于相对频率加权法的自动标引方法。

1960 年，马龙(Melvin E. Maron)与库恩斯(Jennafer L. Kuhns)提出了基于相关概率的自动标引方法。

1969 年，埃德蒙森提出了提示词加权法、题名加权法、位置加权法等新的加权方法，并对不同加权法如何形成最优的组合进行了探讨。

1970 年，厄尔(Lois L. Earl)提出了采用词频统计方法和语言学方法相结合提取关键词的方法。

1973 年，索尔顿(Gerard Salton)和杨(Chun S. Yang)提出了基于词区分值的自动标引方法。

1975 年，索尔顿(Gerard Salton)等提出了基于 VSM 模型的自动标引方法。

1983 年，迪伦(Martin Dillon)和格雷(Ann S. Gray)研制出了 FASIT 系统，该系统是一种基于概念的自动标引方法，由概念选择和概念归类两个标引过程组成。

1988 年，西门子公司推出了文本处理项目 TINA(Text INhalts Analyse)，该项目中的一个组成部分是 COPSY(Context Operator SYntax)系统，该系统可对名词短语进行自动识别、选择、规范、匹配等。

1990 年，迪尔韦斯特(Scott Deerwester)等提出了潜在语义分析的自动标引方法。

1993 年，席尔瓦(Wagner Teixeira da Silva)与鲁伊(Ruy Luiz Milidiu)提出了基于相信

函数模型的自动标引方法。

1995 年，科恩(Jonathan D. Cohen)提出了基于 N-Gram 分析法的自动标引方法。

1999 年，弗兰克(Eibe Frank)等提出了基于朴素贝叶斯的关键词提取方法。

2001 年，安霍(Anjo Anjewierden)与卡贝尔(Suzanne Kabel)提出了基于本体的自动标引方法。

2003 年，隆友清(Takashi Tomokiyo)与赫斯特(Matthew Hurst)提出了基于语言模型的关键词提取方法。

2003 年，胡尔特(Anette Hulth)利用 Bagging 算法提出了基于集成学习的关键词抽取方法。

2007 年，埃尔詹(Gonenc Ercan)与伊利亚斯(Ilyas Cicekli)提出了基于词汇链的自动标引方法。

2008 年，布罗内(Sarah de Bruyne)等提出了基于 H. 264/AVC 视频标准的视频信息自动标引方法。

2009 年，格里(Gowri Allampalli-Nagaraj)与伊莎贝尔(Isabelle Bichindaritz)提出了基于本体语言的自动标引方法。

2010 年，斯米顿(Alan F. Smeaton)等提出了基于 TRECVid 镜头边界检测的视频信息自动标引方法。

2011 年，帕拉尼韦尔(Sengottayan Palanivel)等提出了基于 LPCC 特征和 K-means 聚类算法的音频信息自动标引方法。

2012 年，埃塞尔(Daniel Esser)利用文档的位置和结构，提出了面向归档类文档的自动标引方法。

2013 年，彼得(Piotr Wrzeciono)与卡沃斯基(Waldemar Karwowski)针对波兰语的农业科学论文，结合波兰语词典，提出了一个基于文本分析的自动标引系统。

2014 年，马尔瓦(Marwa Hendez)与阿舒尔(Hadhemi Achour)基于 TF-IDF，利用领域词典，提出了一个针对教育资源的半自动化标引方法。

2015 年，亚当斯(Joel R. Adams)与贝德里克(Steven Bedrick)针对生物医学领域的文献，利用文献摘要之间的相似性，结合 MeSH 词典，提出了基于潜在语义分析的自动标引方法。

2017 年，帕伊(Tayfun Pay)等利用名词性短语和修饰名词性短语的若干个形容词，结合基于位置的启发式过滤方法，提出了一种无监督的全自动关键字抽取方法。

1.3.1.2 国内自动标引方法研究状况

国内自 1980 年以后开始涉足自动标引领域，也取得了很多成果。

1980 年前后，陈培久提出了基于词典切分词标引法的汉语科技文献标题自动标引方法，并用该方法展开了“汉语科技文献标题自动标引试验”。

1984 年，王永成与肖玮瑛提出了基于部件词典的自动标引方法。

1985 年，朱纳克博士等利用语义结构分析法进行全文自动标引试验，实验结果证明其可以媲美手工标引。

1985 年，毛玉姣等展开了对关键词标引的试验，并最终开发了汉语文献自动标引检索系统。

1987 年，北京大学图书馆学情报学系开发了汉语科技文献自动标引系统。

1987 年，邓钦和与龙泽云开发了基于词典分词、词频统计、位置加权三者结合起来的自动标引方法——微机中文情报检索系统。

1991 年，赵宗仁开发了语词结构类比自动标引系统。

1997 年，简立峰提出了基于 PAT 树的关键词提取方法。

2004 年，李素建提出了基于最大熵模型的关键词提取方法。

2006 年，张阔提出了基于 SVM 的自动标引方法。

2006 年，田苗苗等提出了基于遗传算法的 Web 信息自动标引方法。

2007 年，原小玲提出了基于知识元的知识标引。

2007 年，沈静、周金治等提出了基于 UCL 的文化网格标引方案。

2008 年，沈静、周金治等提出了基于 ADO 技术的网页信息自动标引方法。

2008 年，张美娜等提出了基于篇章结构的自动标引算法。

2009 年，章成志提出了基于集成学习的自动标引方法。

2012 年，高影繁、徐红姣等提出了基于多重过滤策略的自动标引方法。

2012 年，杜冉冉提出了基于 DOM 的 Web 信息自动抽取技术。

2014 年，高影繁等提出了一种基于过滤和权重平滑策略的标引词自动抽取方法。

2014 年，王星等利用文献之间的引用关系，提出了基于引文的中文学术文献自动标引方法。

2015 年，许德山等基于本体管理平台，实现了科技文献领域词和未登录词的自动标引。

2016 年，李千驹等提出了一种基于知识组织的关键词自动标引方法。

2017 年，李军莲等通过多维特征概念通用度计算算法，结合 STKOS 超级科技词表和专家审核，构建了面向文献主题自动标引的英文通用概念表。

1.3.2　自动标引方法介绍

自动标引方法主要包括统计标引法、语言分析标引法、人工智能标引法、网页标引法、概率标引法、词典标引法等。下面对这些方法做一详细介绍。

1.3.2.1　统计标引法

在各类自动标引的方法中，出现最早且被广泛持续使用的是统计标引法。统计标引法的基本原理在于术语具有一些显著的统计特征，如共现、逆文档词频、熵、互信息等。[①] 统计标引法包括词频统计法、加权统计法、N-Gram 标引法、统计学习法和分类判别统计法。

1. 词频统计法

词频统计法是指通过对文献中词的出现频率、共现频率等统计指标进行统计排序，找出

① BUITELAAR P, CIMIANO P, GROBELNIK M, et al. Ontology learning from text tutorial[C]. In conjunction with the ECML/PKDD 2005 Workshop on: Knowledge Discovery and Ontologies(KDO-2005). Porto, Protugal, 2005: 31-44.

处于临界域内、能真正表达文献主题内容的词，再根据情况选取适当数量的词作为标引词。①

2. 加权统计法

加权统计法是在词频统计法的基础上引入了加权的概念，因为词频统计法虽然原理简单且使用方便，但标引词的选择范围较大，难以获得较理想的标引结果。由此，人们在词频统计标引的过程中，加入了不同的加权概念，由此形成了位置加权法、相对加权法等加权统计方法。

位置加权法是根据词在文献中所在的位置来对词取不同的权值后，再进行统计。例如，出现在文献标题中的词比出现在文献正文中的词更能代表文献的主题，所以出现在文献标题中的词的加权系数就比出现在正文中的大。

相对加权法主要建立在相对频率这一概念的基础之上。相对频率主要包括文内相对频率和文外相对频率两种类型，其中，文内相对频率是指某词的绝对频数与文献中所有词的绝对总频数之比，文外相对频率是指某词在一篇文献中的绝对频数与其在所有文献中的绝对总频数之比。文内相对频率和文外相对频率都可看作是权值，以此加权即可获得自动标引的抽词依据。②③

3. N-Gram 标引法

N-Gram 标引法是指 $n(n>=1)$ 个相邻字符序列，对文本进行 N-Gram 处理即可得到该文本所包括的长度为 n 的字符串的集合。因为一种语言的 N-Gram 是有限的且较稳定，所以这种标引方法几乎不受学科术语发展变化的影响。但是这种方法仅从形式上对 N-Gram 进行统计，会出现一定程度的标引词不准、标引短语中缺词等问题。④

4. 统计学习法

统计学习法由学习和标引两个过程组成，通过一个学习过程建立标引与促进词和削弱词的关系，并在此基础上确定标引词的标引值。

5. 分类判别统计法

分类判别统计法的主要特点是以词的频数或权值为基点，然后利用统计学中的数值分类法（如聚类分析、因子分析、多维排列或判别分析法）确定词在含义上的相近和疏远关系，同时也从统计的角度解决近义词、同形异义词、异形同义词等问题。这类方法在自动赋词标引中用得较多，在对标引文献进行语义分析时也有所应用。⑤

统计标引法不依赖标引词的领域特征，能够比较方便地在不同领域使用，有一定的使用效果，因而使用较普遍。但该方法只是对词频进行统计，忽略了词语的语义信息，所以要取得更高的标引质量，还需同其他方法结合起来使用。

---

① 张静. 自动标引技术的回顾与展望[J]. 现代情报，2009，29(4)：221-225.

② 肖明. WWW 科技信息资源自动标引的理论与实践研究[J]. 图书情报工作动态，2001(4)：25-26.

③ 张敏. 生物学文献的自动标引系统的研究与开发[D]. 上海：东华大学，2006.

④ 李培. 现代标引方法研究[J]. 图书馆建设，1999(6)：4-7.

⑤ 张静. 自动标引技术的回顾与展望[J]. 现代情报，2009，29(4)：221-225.