自动分类研究进展

肖 明 沈 英

（北京师范大学信息技术与管理学系 北京 100875）（中国科学院文献情报中心 北京 100080）

【摘要】 在对自动分类研究状况进行概述和分析的基础上，指出了自动分类研究的主要发展方向。

【关键词】 自动分类 专家系统

Development of Research on Automatic Classification

Xiao Ming

(*Information Technology and Management Department of Beijing Normal University, Beijing*)

Shen Ying

(*The Documentation and Information Center of CAS, Beijing*)

【Abstract】 In this paper, the authors summary and analyze the existing research works on automatic classification, and point out the primary research orientations of automatic classification.

【Keywords】 Automatic classification Expert system

1 国外研究进展概况

手工分类文献存在着工序复杂、效率低下等缺点，所以国外有不少专家和学者利用计算机来实现自动分类(Automatic Classification)。自动分类的优点是速度较快，易于实现“一文多号”，它特别适合用于自动标引新兴的跨学科文献。从标引方法来划分，自动分类标引可分为抽词分类标引法(Derivative indexing)和赋词分类标引法(Assigned indexing)。抽词分类标引法操作较为简单，易于实现；赋词分类标引法操作相对来说难度更大，目前还缺少这方面的成功报道。

1.1 抽词分类标引法的主要流程是：

①删除无实质意义的语词和泛指词；

②删除无检索意义的高频词；

③对剩余文献进行排序并统计词频；

④统计文献中词的共现频率并排序；

⑤选择词频最高的词或者词组作为文献的标引词；

⑥将计算机选出的词或者词组归入一定的类，并将分类号与文献地址存储在数据库中。

1.2 赋词分类标引法的主要工作流程是：

①为受控词表中的每一个叙词建立一个小词表；

②由计算机利用词频统计标准去识别文献中的关键词。如果某个叙词的小词表中的词能够同从文献中抽取的词相匹配，则该叙词就被计算机指定用来标引该篇文献。

从自动分类的实现途径进行划分，可将自动分类分为自动聚类和自动归类两大类型。其中，自动聚类是指从待分类对象中提出特征，再将提出的全部特征进行比较，并根据一定的原则将具有相同或相近特征的对象定义为一类，设法使各类中包含的对象大体相等。自动归类是指先分析被分类对象中的特征，然后将其与各种类别中对象所具有的共同特征进行比较，再将对象归化为特征最接近的一类并赋予响应的分类号。

自动分类研究始于本世纪50年代。H. P. Luhn在这一领域进行了开创性研究，他提出了词频统计思想，主要用于自动分类。1960年，Maron发表有关自动分类的第一篇论文。1962年，博科(H. Borko)等人提出利用因子分析法进行文献的自动分类。其后，K. Sparck、G. Salton以及R. M. Needham、M. E. Lesk、K. S. Jones等众多学者在这一领域进行了卓有成效的研究工作。概括起来，他们主要从文本的词频统计分析、句法分析和语义分析等三个层次上进行研究。其中，以基于词频统计分析的自动分类试验较为成功。自动聚类多限于理论上的探讨，很少投入实际应用。

收稿日期：2000－05－10

20世纪80年代末，日本庆应义塾大学文学系的图书情报专业和日本IBM东京基础研究所合作开发了一个自动分类专家系统，该专家系统基于日本十进分类法实现了图书资料的自动分类。

自20世纪90年代以来，随着世界范围内出现了一轮又一轮的数字图书馆研究热，国外计算机界和图书情报界陆续展开了对因特网信息资源自动分类的研究，相关研究项目包括：

①北欧WAIS/万维网自动分类项目：该项目由瑞典伦德大学图书馆和丹麦国立技术图书馆合作进行，探讨利用机读版《国际十进分类法》实现因特网资源自动标引来的可能性；

②诺伊斯等人的概念分析试验：主要是利用分面分类法和概念分析等手段来组织因特网资源；

③日本的国际十进分类法数字自动组合系统(UDC－AUTCS)：主要是利用著名的"国际十进分类法(UDC)"进行自动分类试验；

④用于分类体系自动交叉参照的基于知识的系统(KBS－CROSS)：该项目就建筑学领域在UDC(国际十进分类法)和LCC(美国国会图书馆分类法)之间进行交叉参照，目标是解决因特网资源使用中的多语种问题。

2 国内研究进展概况

我国开展自动分类研究起步较晚。1981年，侯汉清对计算机在文献分类工作中的应用作了探讨，并介绍了国外在计算机管理分类表、计算机分类检索、计算机自动分类、计算机编制分类表等方面的概况。此后，我国陆续研制出一批计算机辅助分类系统和自动分类系统。从实现技术上划分，可分为基于词典法的自动分类系统和基于专家系统的自动分类系统两大类；从用户参与的程度上划分，可分为辅助分类系统和自动分类系统两大类。

2.1 基于词典法的自动分类系统主要有以下九种

①莫少强的计算机辅助图书分类系统(C－ABC)

1984年，广东省中山图书馆的莫少强开发出该系统，其设计思想是：以《中图法》军事类中的100个类目为基础，建立模拟的机读分类表；分类员对图书进行主题分析，再键入要分类的主题词，系统自动显示出应该归入的类目，包括分类号、注释和说明等。

②朱兰娟的中文科技文献(计算机类)实验性自动分类系统

1986年，上海交通大学计算中心的朱兰娟在导师指导下开发出该系统，其设计思想是：从随机抽取的一定数量的样本文献标题中抽取全部标题关键词，将其作为候选主题词；接着，建立一个包含类主题词以及类归属度的词表；将全部的类主题词组成一部有限自动机，用该有限自动机去扫描文献标题，抽出其中包含的类主题词，再逐项累加其归属度，当文献类归属度超过某类规定的阀值时，即可考虑将其划归某类。

③张炳恒的半自动图书分类系统

1989年，天津医学情报所的张炳恒开发出该系统，其设计思想是：在《中图法》第二版的基础上，采用元词组配法实现图书的半自动分类，将分类法类名分解为单元词，人工进行主题分析和单元词组配，由系统确定类号。

④苏新宁等人的汉语档案自动分类系统

1995年，南京大学信息管理系的苏新宁等人推出该系统，其设计思想是：以档案的题名与责任者信息作为计算机自动分类的信息源；建立主题词与类号之间的对应关系表；确定主题词与不同类号之间的权重系数；构造主题分类权重词典库、责任者分类词典库、非规范词控制词典库、停用词典库以及跨类类号链接库；分词采用"最长匹配原则"；采用多因素加权分类算法。

⑤叶新明的中文文献自动分类系统

1995年，杭州应用工程技术学院的叶新明推出该系统，其设计思想是：以《中图法》中的类名、《中国分类主题词表》和《汉语主题词表》中的主题词作为词表的主干词汇，并辅以相关工具书以及其它资料作为词汇的补充，建立类名主题词表和组配词表，作为分词元素集合；同时，采用适量的非用词来建立非用词表；以中文文献资料的题名及内容提要作为自动分类标引时获取主题词的信息源；采用"正向扫描，二字先行，半字进或退，最长匹配，异步组词"的自动分词方式，并辅以一定的组词规则。

⑥吴军的自动分类系统

1995年，清华大学电子工程系的吴军推出该系统，其设计思想是：以语料相关系数作为分类依据；以字频、词频以及常用搭配作为补充；采用了停用词表和人工指导分类。

⑦金魏等人的肿瘤学专业文献自动分类系统

1995年，上海交通大学、空军政治学院信息管理系以及上海第二医学院的有关专家合作开发出该系统，其设计思想是：在得到文献的主题词或类主题词后，利用有关专家精心编制的"自动分类用关键词分类归属表"将其转换成对应词的分类归属号，再按分类标引规则形成最终类号。

⑧刘开瑛等人的金融档案自动分类系统

1997年，山西大学计算机系的刘开瑛、郑加恒、刘静等人推出该系统，其设计思想是：以《中国档案分类法金融档案分类表》中的类目词为基础，再加上《金融类公文主题词标引手册》中的部分语词以及从语料库提取的高频关键词来构成类别词库；抽取算法采用正向最长匹配法，并采用模糊匹配法和交叉匹配法以避免漏抽；自动分类算法采用三维加权算法。

⑨张炳恒的全自动图书分类系统

1997年，天津医学情报所的张炳恒提出“全自动图书分类系统”概念，并对研制“全自动图书分类系统”的理论问题作了论述，包括：计算机智能问题、语义结构分析、自动分词问题、全自动图书分类和标引的数学表达式。

2.2 基于专家系统的自动分类系统主要有以下四种

①同济大学计算机系的辅助分类专家系统

1992年，陈大访和陆浩实现该实验性辅助分类专家系统，其设计思想是：将文献的主题词等分类特征作为产生式系统数据库的初始值，通过分类索引知识得到若干相关类号，再根据分类规则知识和文献的其它特征对分类号集合进行裁剪。在此基础上，由用户参与，回答分类号所触发的注释规则和上下位类规则的提问，对分类号集合进一步裁剪。在最小分类号集合中，由用户决定最终类号。

②东北大学图书馆的图书分类专家系统(TSFLZJ)

该系统由李欣和陈星等人在1994年研制成功，其设计思想是：以书名为基础，以联想字库所提供的联想词作为切分书名的依据，以字典作为补充的切分词方法；系统中的知识系统以《中国科学院图书馆图书分类法》为分类专家知识。

③长春地质学院图书馆的图书自动分类专家系统

该系统由邓要武和王连俊于1997年推出，其设计思想是：通过对图书资料的全面系统的分析，利用专家系统的理论和方法，模拟《中图法》的分类原则和有关分类专家的思维方式来构造图书自动分类专家系统。

④基于神经网络优化算法的中文自动分类系统

1999年，上海交通大学的刁倩、王永成、张慧慧等人提出基于神经网络优化算法的中文自动分类系统，其设计思想是：人工给定分类用词权值初值，系统运用神经网络理论进行样本训练；对于多层分类用词表排布时，可进行逐层训练，由次层向主层，自下而上，直到获得令人满意的分类结果时为止。

2.3 现有中文自动分类系统类分文献的准确率大都在80%左右，离实用化和商品化还有一段距离，都存在着若干不足之处

①信息源不充分

现有中文自动分类系统的信息源主要来自文献的题名或者文摘，其依据是：社会科学文献的题名与内容的平均符合率为84%，自然科学的符合率为89.3%。显然，仅依据文献的题名或者文摘来进行自动分类，必然会导致一定程度的误分率。要想进一步提高自动分类的准确度，必须以待分文献的全文内容作为自动分类系统的信息源。

②分词算法不完善

现有中文自动分类系统中多半采用基于词典的分词方法，这就需要考虑词库建设问题。为了使词库能够反映学科的发展状况，必须经常对其进行更新和维护，所以说词库质量会影响自动抽词和自动分类的实际效果。此外，目前的自动分词系统都没有完全解决好书面汉语的歧义问题。以上不足之处都在一定程度上制约着中文文献自动分类系统的分类精度。

③分类法不完善

现有各种分类法都是为适应人工分类而编制的先组式分类法，本身都存在着跟不上科技发展的痼疾，有许多地方都不适合用计算机来实现自动分类。比如，利用基于词典的自动分词算法来切分文献时总会有一些新词切分不出来，从而导致部分文献无法实现自动分类。

④分类程序不完善

现有中文自动分类系统多数是个人自行研制，由于研制者本人水平有限，再加上做不到集思广益，所以开发出来的自动分类程序只能够适用于某一狭小的学科领域，局限性很大，低水平重复现象较严重。

⑤知识库规模太小

基于人工智能的中文自动分类专家系统大多存在着知识库规模过小等问题。此外，人工智能目前还不能能够从根本上解决知识学习等难题，知识库更新较慢，这也是现有自动分类专家系统不实用的主要原因。

3 未来发展方向

总结国内外有关自动分类研究的现状，我们不难看出国内外自动分类研究正朝以下几个方面向前发展。

(1)基于语料库构建自动分类系统

语料库是能够代表某一学科领域的语言现象的大量真实语言材料的集合，它至少在两个方面可以支持自动分类研究：①词典信息。基于词典法的自动分类系统的核心是词典中主题词的质量。现有自动分类系统的共同缺陷是，其中使用的词典落后于科技发展，利用语料库技术则有助于解决该问题；②分词信息。困扰自动分类系统的主题词获取难题也可通过语料库技术来获得解决。

(2)构建实用性更强的自动分类专家系统

现有自动分类专家系统的共同缺陷是：知识库较小，知识表示方法单一，没有解决好分类主题信息的获取等问题。我们可针对上述缺陷进行改进，以构建实用的自动分类专家系统，具体措施包括：对任务进行分解，将系统分解为由多个小系统有机耦合在一起的协作分布式系统，每个子系统求解策略单一，规模较小，易于构造和维护。此外，还应根据分类领域的具体情况选用合适的知识表达方法。

(3)构建适合自动分类的机读分类法

目前，国内有关人员在研制文献分类法数据库方面做了很多努力。比如，中国科学院成都文献情报中心研制了《科图法》数据库系统；北京图书馆建立了《中国分类主题词表》数据库，并对外发行《中国分类主题词表》机读版；山西省图书馆、北京图书馆、中国科学院文献情报中心和中国人民大学图书馆联合开发了《计算机文献标引对照系统》。但是，现有中文文献分类法多属于先组式分类法，并不适合用来实现文

献自动分类。要想大幅度提高中文文献的自动分类水平，则应该尽快研制出用于自动分类的机读分类法。

(4)要想进一步提高文献自动标引水平

还需要编制将情报检索语言和自然语言紧密结合的若干对应表，至少应该包括自然语言接口对应表和自动抽词词典。利用"自然语言接口对应表"，可在自然语言和情报检索语言(主题词或者分类号)之间进行自动转换，并为实现书面汉语的生动分词和自动分类提供更大方便。如果能够编制出高质量的各种专科性自动抽词词典，则能进一步提高自动分词和自动分类水平。

(5)加强对因特网信息资源的自动分类研究

如果不对海量的因特网信息资源进行有效的分类或者主题标引，则人们将难以利用这些取之不尽、用之不竭的宝贵财富。

目前，国外已经启动有关自动类分因特网信息资源的若干研究项目。比如，瑞典伦德大学图书馆和丹麦国立技术图书馆合作进行了WAIS/万维网自动分类项目，日本正在研究国际十进分类法数字自动组合系统(UDC－AUTCS)，瑞典伦德大学正在开展用于分类体系自动交叉参照的基于知识的系统(KBS－CROSS)项目。国内目前还没有类似的相关报道。尽管国外已经开展对因特网资源进行自动分类研究，但目前取得的成绩仍然非常有限，距离实用化还有许多技术难题尚待解决，并成为今后一段时间内计算机界和图书情报界共同研究的热点、难点和重点。

参考文献

1 成 颖 史九林. 自动分类研究现状与展望. 情报学报，1999，18(1)：20－26

2 细野公男著 董光荣译. 文献资料的自动分类. 国外图书馆情报工作，1990，(2)：37－42

3 刁 倩 王永成 张惠惠. 中文信息自动分类系统及其神经网络优化算法. 信息与控制，1999，28(3)

4 叶新明. 中文文献自动分类研究概述. 情报理论与实践，1992，(5)：39－41

5 白国应. 中国文献分类学研究50年(1949－1999). 中国图书馆学报，1999，(5)：63－67

6 张琪玉. 缺少抽词词典是自动抽词标引难以普及的主要原因. 图书与情报，1988，(2)：27

7 孟广均 徐引篪. 国外图书馆学情报学研究进展. 北京图书馆出版社，1999，9

超星快讯

超星数字图书馆近期将免费赠送500张读书卡

网络的出现，使数字图书拥有了新的载体，同时也引发了一些新的问题与思考。因为上网费用较高，读者迫切要求把书籍下载到本地阅读，但是数字图书馆在提供下载服务之前，首先要突破版权的瓶颈。为了解决书籍下载时所涉及的版权问题，超星数字图书馆(www.ssreader.com)积极与中国版权保护中心洽谈全面的版权代理，创建了全新的超星读书卡数字图书下载收费模式，并将读书卡的销售收入按照传统出版业中的版税制原则，合理地分配给作者，从而使作者、读者都可以从这种良性的商业模式中获益。

世纪超星公司首期发行的读书卡面值有三种：10元、30元、100元，可分别在超星数字图书馆下载图书1个月(31天)、3个月(93天)、1年(365天)。促销期间，购买30元、100元的读书卡可分别获赠2张和4张精品图书光盘，光盘具体内容分为《中国古代文学经典》、《中国古代学术经典》、《世界文学经典》、《世界学术经典》。

读者购买超星读书卡后，在超星数字图书馆登录注册卡号、密码，注册完毕后即可在超星数字图书馆许可的范围内下载图书。世纪超星公司近期将免费赠送面值为10元的读书卡500张，持有此卡的用户可以在超星数字图书馆免费下载图书1个月。具体申请方法敬请关注超星数字图书馆发布的最新消息。

超星图书阅览器免费捆绑及下载量平均每天增长1.3万套

由于超星数字图书馆藏书量达10万余册(3500万页)，可下载图书近2万册，所以从一开始就吸引了很多的读者浏览。目前，超星数字图书馆的日访问量已达25万人次。受其影响，超星图书阅览器的需求急剧增长，免费捆绑及下载量在突破120万的基础上，继续以每天1.3万套的速度递增。

为了更好地服务用户，世纪超星公司进一步加大了研发力度，2000年5月份，最新版本的超量图书阅览器(SSReader3.5)将提前问世。SSReader3.5最显著的特点是嵌入了汉王OCR识别软件，用户可以将扫描的图书资料转换成文本文件，使用更加方便。同时，SSReader3.5还支持多远程浏览，用户可以在超星数字图书馆看到国家图书馆、广东省图书馆、清华大学图书馆等众多采用超星PDG技术制作的所有数字图书馆的全部内容。除此之外，超星数字图书馆还提供图书光盘定制服务，用户只需将自己所选定的图书目录和索书号(SS号)告知超星数字图书馆，超星数字图书馆就会将图书内容刻录在光盘上，以邮寄、快递或其它方式送给用户。

(本刊讯)