索引研究论丛

·中国索引学会主编·

索引技术和索引标准

北京图书馆出版社

索引研究论丛

索引技术和索引标准

侯汉清　主编

北京图书馆出版社

图书在版编目(CIP)数据

索引技术和索引标准/侯汉清主编．－北京：北京图书馆出版社，1997.10

ISBN 7－5013－1452－7

Ⅰ.索… Ⅱ.侯… Ⅲ.①情报检索－文集②情报检索－标准

Ⅳ.G252.7

中国版本图书馆 CIP 数据核字(97)第 18927 号

书名 索引技术和索引标准

著者 侯汉清 主编

出版 北京图书馆出版社(原书目文献出版社)

发行 (100034 北京西城区文津街 7 号)

经销 新华书店

印刷 北京航空航天大学印刷厂

开本 850×1168 毫米 1/32

印张 10.31

字数 257(千字)

版次 1997 年 10 月第 1 版 1997 年 10 月第 1 次印刷

印数 1－3000

书号 ISBN 7－5013－1452－7/G·385

定价 15.00 元

目　次

情报检索语言的发展趋势

——与吴建中的对话

张 琪 玉

吴： 80年代初，张教授针对当时人们把注意力集中在体系分类法的思想性方面，使图书分类研究陷入所谓"三性（思想性、科学性、实用性）"怪圈问题时指出："要改变研究方向，把研究的重点转移到如何提高情报检索语言的检索效率方面来"，并独辟新路，开创"情报语言学"，提出从更高的层次对分类法、主题法等各种检索方法进行统一研究。张教授这一开拓性的研究，对我国情报检索语言领域的理论与实践起到了积极的导向和推动作用。我想首先请教一下张教授，为什么要提出情报检索语言的概念，它与分类法、主题法是什么关系？

张： 我认为，分类法也好，主题法也好，都是情报检索系统的组成部分，都是在寻求更佳的检索效果中创制出来的，所以，它们的基本原理是一致的。只是它们在表达各种概念及其相互关系时，和在解决对它们提出的那些共同要求时所采取的方法不同，才形成了不同的类型和语种。在情报检索语言的概念下，对它们进行综合研究，可以找出它们之间最本质的东西，以及结构和功能上的相同或相异之处，概括出影响检索效率的共同规律。总之，情报检索语言是表达一系列概括文献情报内容的概念及其相互关系的概念标识系统，其职能是作为情报检索系统的语言保证，它的核心问题

是检索效率。

吴： 张教授，能不能给我们谈谈当今情报检索语言的主流是什么？

张： 这要从两个方面来谈。1.目前在文献数据库的标引方面，世界各国都主要使用叙词语言。《叙词表指南》收录有500多种世界各国的叙词表，我国也主要使用叙词语言。近10多年来，为了适应建立数据库的需要，我国编辑出版了上百种汉语叙词表。叙词语言比较适应计算机检索系统的组配检索；2.在图书馆藏书目录的标引方面，美、加、澳等国主要使用标题词语言，我国和独联体国家以及英国主要使用分类检索语言。这与各国图书馆使用情报检索语言的传统有关。

吴： 你能否告诉我们世界上这几年情报检索语言发展的一些主要情况？

张： 可以概括地说：1.国外，特别是图书情报事业发达的美国，对自然语言在情报检索中应用的研究很活跃，这是他们研究的超前性，这已成为国外情报语言学研究的热点。但至今未见公认的振奋人心的突破性进展。自然语言检索实验系统达到上千个，这类实验目前处于这样的困境：以自然语言研究的主要方面——自动标引来说，“在某种意义上恰似机械鸟的制造，经过20多年的试验，有些外貌开始像鸟，有些能够模仿几声鸟鸣，有些能扑打一番翅膀，但至今还没有一只会飞会鸣，”“绝大多数自动标引系统始终未能走出实验室大门，投入使用”（储荷婷：《自动标引的主要方法》，载《情报学报》1993年12卷3期）。由于自然语言检索的方案很多，需要弄清其孰优孰劣，因此对各种自然语言系统的评价试验，也是自然语言研究的一个重要方面。在自然语言接口与人工语言结合使用等方面，则已有很大成绩；2.自然语言的初级应用（如关键词检索、文本匹配查找）迅速扩大，但一些著名的检索工具和数据库并没有放弃人工语言的迹象。新的情报检索语言语种仍

有创制，原有的情报检索语言语种仍在修订更新，但对情报检索语言新类型的研究已较少；3. 在情报检索语言易用化和兼容方面的研究甚多，而且有些是很新颖的。

吴： 那么，计算机技术的发展是如何影响情报检索语言的呢？

张： 我曾在《情报语言学基础》一书第 11 章中，概括说明过这种影响：50 年代开始的情报检索计算机化，促进了情报检索语言的创新和改造，使词表、分类表向机编化和机读化方向发展，使文献标引和索引编制走向自动化，使自然语言检索得以实现，使多种语言的结合使用成为可能，使检索方法有了很大的进步，并正在使情报检索语言的应用范围扩大（例如开始应用于情报研究和文献计量）。特别是自然语言在情报检索中的应用，使情报检索系统的语言不再局限于情报检索语言。同时，情报检索计算机化对情报检索语言研究提出了许多新课题，并提供了许多新方法和新条件。总之，情报检索计算机化对情报语言学的发展产生了极为深刻的影响。甚至可以说，情报检索计算机化是加速情报语言学形成过程的一个重要因素。可以预见：情报检索计算机化今后将会更快、更广阔、更深入地得到发展，情报语言学也将会有更快的进步。

吴： 一些研究认为，在某些情报检索系统中分类法与词表已无必要使用，张教授的观点如何？

张： 如果说是在某些系统，这是正确的，我赞同这个观点。的确，在某些系统，如报纸数据库，甚至大型综合性期刊论文数据库等，用别的语言比用分类法和词表这类语言工具更合适。但是，如果把某些系统改为在大多数系统或在一切系统（有些自然语言检索研究者曾这样认为），那就不正确了。如藏书实体的组织（这也是一种检索系统）仍然需要分类表；藏书目录的组织，也以使用分类表和词表为好。专利检索系统，不用分类法是不可想象的。一些高要求的检索系统，仍然需要人工语言。为了提高自然语言的检索效率，需要采取后控制措施，特别是采用后控制词表。其实，后

控制词表，按其原理和功能，也可以认为它仍然是一种情报检索语言。将来有可能使用一些人工语言与自然语言结合的、带有自然语言换词、换号功能的分类表和词表。语义关联对情报检索是绝对必要的，既可以在先控制系统中，也可以在后控制系统中。

吴： 张教授，怎样来看待分类法和词表今后的发展趋势呢？

张： 我认为是一体化，是各种类型分类检索语言的一体化，各种类型主题检索语言的一体化，分类检索语言与主题检索语言的一体化。理想的情报检索语言应是：1.既可从学科、专业角度检索，又可从事物角度检索；2.既可按系统入手检索，又可按字顺入手检索；3.既可先组式使用，又可后组式使用；4.既可进行专指性检索，又可进行泛指性检索；5.既可用词进行标引和检索，又可用号码进行标引和检索；6.既可用人工语言进行标引和检索，又可用自然语言进行标引和检索。在计算机检索系统中这是有可能的，因为构成这种理想语言的方法和技术已经存在。我认为，理想的情报检索语言应是“学科——事物”型检索语言。它由学科分类系统面和事物分类系统面构成。两个面可互相组配。当按学科聚类时，藉助于事物及其部分面进行复分；当按事物聚类时，藉助于学科及其问题面进行复分。

吴： 有研究表明，无语义关联、无控制的词汇也能使用，而且在某些情况下要比有语义关联、有控制的词汇使用情况更好，我们还可以看到使用自然语言的数据库数量正在上升。鉴于此，我们能不能说在未来文献数据库中，自然语言的使用将占统治地位呢？

张： 有些研究的结论是令人振奋的，但并没有得到多次普遍的证实，我认为那些结论有点言过其实。还有些文章往往只是引用那些研究结论，而并没有亲自去检验过那些结论的可靠性。对自然语言的研究无疑是一个正确的方向，这里我决没有意思去否定这个方向。相反，我觉得我国对自然语言的研究很不够，与国外有不小差距，应加强这方面的研究。你说的“无语义关联、无控制的词

汇也能使用”，我觉得可以这样说，至于“比有语义关联、有控制的词汇使用情况更好”，这就很难说了。目前未见充足的、普遍的证实。我认为，自然语言有优点，也有缺点。单纯使用自然语言是取其长（如时差短，对处理人员要求不高，特别是成本低），而在某些方面则容忍其短，放弃某些质量要求。自然语言要全面胜过人工语言是不可能的，除非它引进许多情报检索语言的原理和方法，而不是单纯的自然语言。自然语言缺少控制，而对于高要求的检索来说，控制是绝对必要的（如果无控制更好，全世界的情报检索早已全面自然语言化了）。检索中需要对表达文献情报内容的语言进行控制，只是控制的程度可随不同要求而异。怎样控制，控制到什么程度，这倒是一个需要研究的问题。使用自然语言的数据库正在增长，这是事实。而且可以说，情报检索系统一直在向增加自然语言检索功能的方向发展，目前国外大多数数据库都提供自然语言检索途径。但是，要区别是单纯使用自然语言的数据库，还是自然语言与人工语言并用的数据库。即使是30%的数据库单纯使用自然语言，60%的数据库自然语言与人工语言并用，10%的数据库单纯使用人工语言，这样，使用自然语言的数据库占了90%，而使用人工语言的只占70%，那也不能说，自然语言占了统治地位，因为自然语言与人工语言在检索中的作用是不能相提并论的（凡并存者，自然语言都只是补充的地位）。现在使用自然语言的系统往往并不放弃人工语言，这是一个充分的证明。只有原来使用人工语言的系统都放弃使用（或大多数放弃使用）人工语言而单纯使用自然语言了，才能说明自然语言全面好于人工语言，占统治地位了。自然语言的根本缺点——词汇无语义关联、无控制，并不因为使用计算机和各种检索方法而不再存在了。

吴：　在自然语言系统中，目前还存在着哪些尚未解决的问题呢？

张：　自然语言在情报检索中的应用，面临着以下两个难题：1.是如何从自然语言文本中自动抽出最能准确、充分地表达文献有价

值内容的词，以及这些词与检索课题有效匹配的问题。这个问题的复杂性在于文献作者的用词无明显的规律性，以及作为人类社会现象的自然语言不可能用纯自然科学的方法去研究解决。这个问题同机器翻译的性质类似。如果去追求百分之百的自动化，至少在短期内是无希望解决的（当然，自然语言自动处理现有的一些中间成果还是有实用价值的）。如果采用人机结合的方法，则可以较为容易一些；2.是克服自然语言由于不规范和缺乏语义关联性而对检索不利的问题。克服这个难题也是不能完全用自动化方法的。使用后控制词表可能解决这个问题，而且后控制词表兼有自然语言与人工语言的性质和优点。但后控制词表的编制需要人的参与，才有可能做到半自动化（兰开斯特曾提出用积累检索提问素材的办法自动编制后控制词表，其可行性十分有限）。除此以外，对中文来说还有一个汉语分词的问题。由于在汉语中，词与词之间没有分隔符号，一个汉字可以同其他许多汉字进行组合构成不同含义的词和词组，因此，计算机很难识别一个句子中，哪个汉字或哪几个汉字的组合是词，而自动把它们分离出来，也难于准确区别有用词与无用词。所以，直接利用汉语进行检索（在文本中进行语词匹配查找和单汉字检索除外），首先必须解决把汉语句子用计算机自动切分成词的汉语分词技术。10多年来，我国进行了大量的汉语分词技术研究，提出了许多分词方案，总的来说，在这方面已取得了很大进展，现在距离解决汉语自动分词的问题已为期不远。但这个问题的解决，只是达到了拼写文字国家的起点水平，拼写文字中未解决的上述两个问题仍有待我们去解决。

从人工语言到自然语言

——与吴建中的对话

侯汉清

吴： 前不久，我随中国电子图书馆考察团参观了美国和加拿大的一些图书馆和情报机构，发现北美联机数据库使用自然语言系统的数量正在增长，其中相当数量的文献数据库，由使用叙词表改为同时使用叙词表和关键词，还有一部分文献数据库则不采用叙词表和分类表，不进行任何标引，直接用文本检索或称全文检索。侯教授，你认为在中文文献数据库建设中，应当如何使用自然语言，使用自然语言有哪些优点？

侯： 在文献检索中使用自然语言是指使用文献作者或文摘、提要的作者原来使用的语言，其中包括关键词、自由词和出现在文献题名、摘要或正文中的语词。相对于使用情报检索语言（即人工语言）来说，使用自然语言的优点是：1. 可以取消复杂费时的标引工作，或者降低标引工作的难度和成本，提高标引工作的速度；2. 直接使用文献用语和作者用语，可以改善标引的专指性和一致性，从而提高检索的效率；3. 用户熟悉自然语言，使用起来方便得多，尤其是在联机网络环境之中。使用自然语言建立中文文献数据库有多种方式。第1种方式是使用关键词抽词标引。像上海图书馆建立自然科学报刊文献数据库，可以采用这种方式。用手工或计算机直接从文献的题名、摘要、大小标题、正文（尤其是文章开头、结

尾等部分)以致参考文献中抽取表达文献主题内容的、具有检索意义的语词,用作检索的标识。目前,汉语自动分词技术已经接近实用水平,只要把这些抽词用的语料扫描到计算机中去,并使用完善的禁用词词典和关键词词典与上述语料匹配,即可实现自动标引,而且机器抽词标引与人工抽词标引的吻合率极高。中国农科院农林文献数据库在1994年用这种方式进行了成功的试验,标引的速度可以提高十几倍。

吴: 自然科学、技术科学文献题名与文献内容的相符率很高,可以采用关键词抽词标引和增补标引。但是,人文科学、社会科学文献情况就不一样,题不达意的甚多,怎么来解决这些问题呢?

侯: 社会科学文献名词术语稳定性差,更新快,新词多,因而建立社会科学文献数据库宜于采用第2种自然语言标引方式,即自由标引。标引员无需查对词表,自主拟定标引词或选用文献题名、摘要或正文中合适的词来表达文献主题。所谓自由标引,也不是完全自由,需要标引人员有一定的受控标引的经验,参考一定的标引规则和标引模式,这样可以大大加快标引的速度,而且提高标引的专指性。目前,北京学习出版社主持的新闻信息数据库采用这种标引方式。我想,如果上海图书馆的社会科学篇名数据库改用自由标引,一定可以克服标引难度高、速度慢、专指性差等问题。建立中文文献数据库还可以采用其他方式,例如情报检索语言与自然语言并用的方式,即既用叙词和分类号标引和检索,又使用文献题名、摘要或正文中的关键词、专用名词或自由词(作为受控标引的补充)进行检索。当然,还可以采用无标引的全文检索方式或其他方式。

吴: 国外有些学者断言,情报检索语言已经过时,而且很快就会被自然语言所淘汰,能不能谈谈你的看法?

侯: 1981年我曾经撰写过一篇《分类法的发展趋势简论》的文章,认为情报检索语言在朝着分面组配化、分类主题一体化、自动

化和标准化的方向发展。10多年过去了,情报检索语言的发展趋向还应当增加自然语言化这一条,也就是说受控的情报检索朝着与自然语言结合的方向,或者大量使用自然语言标引和检索的方向发展。自然语言化可以说是当代检索语言发展最重要的特征和趋势,应当引起我们足够的重视。

吴: 一些研究表明,在一些大型情报检索系统中,分类表和叙词表已不再是必要的。那么,分类表和叙词表的未来前景如何呢?

侯: 自然语言标引和检索虽然也可以用于手工检索(如关键词目录或索引),但只有在联机检索或联网的环境条件下,才能高效率、高水平地实现。在我国,计算机技术在图书馆和情报机构的应用尚未普及,加之汉语自然语言检索不少技术问题有待于解决,因此,目前自然语言系统不可能取代受控检索语言,当代情报检索语言的主流仍是叙词语言。正如美国著名情报学家兰开斯特在《情报检索词汇控制》(1980年第2版)中所指出的,随着用智力加工所需成本的不断上涨,计算机存储费用的不断下跌,以机读形式存取的文本(包括电子邮件、电子报刊)数量的逐渐增多以及在联机检索中对熟悉的"中间人"依赖的逐渐减少,"自然语言将变成情报检索的规范,普通受控叙词表的使用将会衰退,这似乎是肯定无疑的了"。至于自然语言是否会完全取代情报检索语言,分类表和叙词表是否会被淘汰,仍是一个值得商榷的问题。我认为,假如一个机读数据库要提供一个印刷型主题索引,或者要发挥其组织知识以及相关性检索的功能,仍然离不开分类表和叙词表。如果用关键词来编制主题索引或分类索引不是不可能的,至少说是低水平的。

吴: 说到关键词检索,不知同义词问题将如何解决?假如我们从文献题名及文摘中抽取关键词,关于"降价销售",在不同文献中就会出现"削价销售"、"特价销售"、"优惠销售"、"减价销售"、"折扣销售"等多个不同的关键词,如用这样的自然语言建立文献数据

库，检索时检全率岂不是很低吗？

侯： 自然语言确实表达概念较为自由，存在着对大量的同义词、近义词、同义词组、近义词组缺乏控制的弊端。除此以外，自然语言还存在着对词量、词形、词间关系不进行控制的缺点，从而影响检索的效率。从这个意义上说，不管今后计算机技术和自然语言系统如何发展，情报检索的基本原理——对词汇的控制，是永远不会取消的，变化的只是词汇控制的方式、方法和手段。就自由标引的方式来看，在标引阶段不查词表，不实行严格的词汇控制，可以自主地选择标引用词，但是，为了提高检全率，减轻用户检索时拟定检索策略的智力负担，仍需要在检索时提供一种后控词表。这种后控词表只用于检索而不用于标引，采用字顺或分类的方式显示各种关键词或自由词之间的等同关系（即同义关系）和等级关系。有了后控词表，就可以把各种“自由散漫”的自然语言标识组织起来，形成一个语义网络，以便于检索。用户假如从“降价销售”入手查找，计算机可以自动地把标引“降价销售”以及另外5～6种关键词的记录检索出来，甚至还可以把“降价销售”的下位词，如“服装降价销售”、“花卉降价销售”、“季节降价销售”等有关记录统统检索出来，供用户选择，用户就不必挖空心思地思考有关“降价销售”的文献，可能会用哪些关键词或自由词标引，使检索过程更为简单、高效。从上述实例可以看出这种后控词表综合了自然语言和常规受控语言的优点。兰开斯特指出：“后控词表的发展为改进联机网络内的检索效果以及成本——效益提供了良好的前景。事实上这种方法值得引起比以往任何时候更多的重视”。目前有一些全文数据库，不标引，不搞任何词汇控制，建库速度快，成本低。但是普遍存在着误检率高、检全资料困难等缺点，每次检索都可能带出大量不相关的或相关性较小的文献，大大加重用户筛选所需文献的负担。张琪玉教授近两年来进行的关于自然语言标引和检索的研究表明，报刊文献数据库应当走自然语言标引＋后控

制的道路。

吴： 近年来，分类主题一体化词表成了国内情报检索语言研究的一个热点，无论在理论研究方面还是在词表编制方面都取得了不少成果。你能否给我们介绍一下这方面的动态和发展？

侯： 所谓分类主题一体化是指分类法和主题法的有机结合，即对分类表和叙词表的术语、参照、标识和索引实施统一的控制，使两者有机地融为一体。这种词表称为分类主题一体化词表。由于我国具有使用分类法的传统，因此，这种一体化词表在我国受到特别的欢迎。据统计，近 10 年来我国探讨分类主题一体化的论文多达百余篇，编辑出版的一体化词表多达 20 多部，其中，一种类型是分面叙词表，如《教育主题词表》、《社会科学检索词表》等；另一种类型是分类法叙词表双向对照索引，如《中国分类主题词表》、《中图法与 MeSH、中医药学主题词表对照表》等。这种一体化词表是一种分类号和主题词之间兼容转换的工具，其最大优点是可以通过标引数据的转换同时完成文献的分类标引和主题标引，提高标引的数量和质量。上海图书馆的中文社会科学篇名数据库就是利用《中国分类主题词表》的这种功能，每人每天标引报刊文献多达 100 篇。近年来的研究还发现，以《中国分类主题词表》为基础，还可以建成后控词表，用于自然语言检索系统。上海空军政治学院的硕士研究生在张琪玉教授指导下，建成了基于分类主题一体化词表的自动分类和自动标引系统。现在看来，把分类主题一体化词表和自然语言结合起来，增加检索语言与自然语言的对应转换功能，是一条走向自动化的捷径。

吴： 最后，请侯教授简要地谈一谈情报检索语言未来的发展。

侯： 应当指出，计算机技术的应用是情报检索语言发生深刻变化的主要动力。原先很多想做而不能做的事情，如分类主题一体化、自然语言检索、后控制词表、超文本检索等都已经成为现实。随着计算机网络和电子出版物的发展，情报检索语言将会有更多的创

新，预计在世纪之交，诸如自动标引、自动分类、自动摘要、自然语言理解、智能情报检索等都会在我们这一代实现。