中国图书馆学报(双月刊) 1999 年第 5 期
ZHONGGUO TUSHUGUANXUEBAO

●曹树金 罗春荣 汪东波

开创情报语言学的新天地

——论张琪玉教授对情报语言学的新贡献

摘 要 张琪玉教授在自然语言检索和情报检索语言研究方面成就卓著; 对情报检索语言的未来发展有独到见解。参考文献 26。

关键词 自然语言检索 情报检索语言 张琪玉
分类号 G354

ABSTRACT Professor Zhang Qiyu has made remarkable achievements in natural language searching and information retrieval language research. He also has original views of the future development of information retrieval language. 26 refs.

KEY WORDS Natural language searching. Information retrieval language. Zhang Qiyu.
CLASS NUMBER G354

1 导言

80 年代初, 张琪玉教授以保证情报检索效率为中心, 开拓性地对各种情报检索语言进行系统、整体的研究, 著就《情报检索语言》一书, 标志着情报语言学的诞生。张教授不仅因此成为情报语言学家的开山鼻祖, 而且代表着中国人对世界图书馆学情报学的一大贡献[1, 2]。

以后, 张教授将全部身心倾注在情报语言学的研究、实践和教育上, 孜孜不倦地探索和耕耘。至今, 他已经发表论文、译文 160 余篇, 论文被各种文集等转载约 60 篇次; 编著、主编、参编的专著有 20 多种。因在编制《中国图书馆图书分类法》中作出较大贡献而获 1985 年国家科学技术进步奖一等奖; 因在编制《中国分类主题词表》中作出的贡献而获 1996 年国家优秀科技信息成果奖二等奖。特别是近几年, 张教授年均发表情报语言论文 10 余篇, 为情报语言学的深化和拓展做出了新的巨大贡献。

最有代表性的是, 1980～ 1997 年期间, 《情报检索语言》及其增补修订本《情报语言学基础》, 以 5 种版本累计印刷 11 次, 共 99900 册, 被 288 位作者的专著和论文共引用约 450 次。作为教材, 在图书馆学、情报学的本科、函授和电大等教学中被广泛采用, 估计使用该教材学习过这门课程的不少于 25000 人。

本文将着重讨论张琪玉教授近些年来在自然语言检索、情报检索语言和标引自动化、未来检索语言模式等方面的研究。

2 对自然语言检索的研究

从 80 年代开始, 自然语言检索成为国外情报检索和自然语言处理领域的共同研究热点。这从 1981～ 1995 年间, *ERIC*, *INSPEC*, *Compendex Plus*, *LISA* 和 *Information Science Abstracts* 五个数据库所收录的"全文检索"的文献数分别是 133, 471, 40, 525 和 107, 就可见一斑[3]。张教授是国内较早关注自然语言检索的学者之一。至今, 他对自然语言检索的研究最为全面、系统, 其主要研究成果已集中反映在《情报语言学基础》增订二版的第十二章中[4]。

2.1　对自然语言检索研究的呼吁

早在1983年张教授就呼吁我们在研究情报检索语言的同时,应注意对自然语言检索的研究[5]。不久之前,他回顾了我国自然语言检索研究进展[6],鉴于"深入考察自然语言性能者不多见",他再次呼吁:"情报语言学研究者应当积极参与自然语言检索的研究,当前亟需从情报语言学角度深入研究自然语言检索方法,把情报语言学的原理和方法引进自然语言检索的研究。"[7]

2.2　自然语言检索与情报检索语言关系的辩证观点

根据情报检索的基本要求以及自然语言检索和情报检索语言的各自优缺点,张教授一方面肯定,"在计算机检索越来越发展的条件下,自然语言具有不可阻挡的发展前途。特别是在互联网络的检索环境中,它将成为一种必然的优先选择"。另一方面,并不赞同自然语言将替代情报检索语言或最终将替代情报检索语言,情报检索语言或情报检索语言研究已经过时的论点。他认为:自然语言检索系统与情报检索语言检索系统并不是绝然对立的,它们各有长处和短处,可以并行发展,可以互相结合,互相补充。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。对于高要求的情报检索来说,控制是绝对必要的。而"对检索过程进行控制",正是情报语言学的精髓。我们应当把情报语言学的理论成果运用到自然语言方面去。为了提高自然语言的检索效率,需要采取后控制措施。语义关联对情报检索是绝对必要的,但既可以在先控制系统中,也可以在后控制系统中[8]。

至于自然语言或情报检索语言的未来,他认为将是自然语言的情报检索语言化或情报检索语言的自然语言化。

2.3　对自然语言检索中影响检索效率的各种因素的深入分析

张教授对自然语言检索的研究,秉承他一贯以检索效率为核心的传统。因此,他具体论述了影响自然语言检索效率的主要因素:检索所依据的文本的类型(文献题名、文献中的小标题和章节名、文献的摘要和正文);检索用语的专指度;在文本的不同范围(句、段、节、篇)内进行组配检索;文本用词的不规范性;不同标引方法(不标引、自动抽词标引、人机结合抽词标引、自动赋词标引和自由标引);对自然语言进行词表控制的程度。在此基础上,提出改善自然语言检索效率的基本原则[9]。

通过调查国内外的相关文献,我们发现,张教授对自然语言检索中影响检索效率的各种因素的分析,最为深入全面。这对正确认识自然语言检索存在的问题,探寻有效的改进方法有很大的启发。

2.4　自然语言在情报检索中应用方式的全面总结及有待解决难题的认识

为了纠正不少人对自然语言应用于情报检索的片面理解,开阔了人们的视野。张教授全面总结了自然语言在情报检索中应用的方式。他指出:关键词法;文本检索;单汉字检索;以自然语言作为自由词进行补充标引,与情报检索语言结合使用;以自然语言作为入口词(接口),利用计算机的换词功能,辅助情报检索语言;自动赋检索词和自动赋分类号;自动分类(自动聚类法);自由标引等都是自然语言应用于情报检索的方式[10]。

张教授明确指出,自然语言在情报检索中的应用,面临着两个难题,一是如何从自然语言文本中抽取最能准确、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;二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。此外,对中文来说,还有一个汉语分词的问题。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[8]。这实际上是强调了我们研究自然语言检索应该把握的方向。

2.5　对后控制词表的系统研究

张教授对后控制词表非常重视,他在多篇论文中强调后控制词表是弥补自然语言缺陷

的主要措施, 他发表《论后控制词表》一文[11], 系统论述了后控制词表的控制机理、控制程度、后控制词表编制上的特点、后控制词表的各种编制方式及其在控制上的差别, 提出并用实例说明了一种"分类词表+字顺/轮排表"的后控制词表模式, 阐述了利用后控制词表检索文献的各种方法。

2.6 关于自由标引的论述

自由标引是情报检索中利用自然语言的一种简易方式, 我国不少单位在专题数据库建设中采用这种不依据词表进行主题标引的方式。但是, 专门论述自由标引的文献很少。张教授的《论自由标引》[12]一文, 全面阐述了自由标引的概念、自由标引的优点及适用范围(主要适用于报纸文献、期刊文献的大型篇名数据库等的标引)、自由标引种类、自由标引基本方法、自由标引要点、自由标引系统配备后控制词表的必要性, 有益于人们在理论上准确认识自由标引, 在实践中正确采用自由标引。

2.7 对汉语关键词法的探讨和试验

在《情报检索语言》一书中, 张教授就将关键词法作为一种"准情报检索语言"予以重视[13]。后来, 针对关键词法应用于汉语的特殊问题, 不仅进行理论探索, 而且亲自编写程序, 利用计算机进行关键词索引的编制试验, 先后发表了多篇论文。

在《汉语关键词法探讨》一文中[14], 论述了关键词法在我国的实用价值; 分析了汉语文献题名进行自动抽词的进展与问题; 指出人机结合抽词的技术路线目前仍有现实意义。文中讨论了人工抽词—计算机整理方式和计算机抽词—人工干预方式的具体问题, 比较了人工抽词、人工干预和自动抽词三种方式的各自优缺点, 最后阐明抽词词典、规范词典和后控词表的编纂对发展汉语关键词法的重要意义。

在《人—机结合的题内关键词索引可回避汉语分词难题》中[15], 他指出, 汉语题内关键词索引可以"含糊抽词", 实际上是"最长抽词"与"词素轮排"的结合。而且还可以对题内关键词索引加以改进。

《汉语关键词索引的一种编制方法》[16]和《汉语关键词索引的另一种编制方法》[17]介绍了张教授将WPS与dBASE结合编制题内关键词索引的两种简易方法及实现程序。

从文本(主要是题名和文摘)中自动抽取关键词是较易实现、尚可接受的一种自动标引方案。近20年来, 我国学者对汉语分词技术作了许多研究, 提出了不少分词方案, 但见于实际使用者不多。张教授认为, 原因并不是这些方案都经不起实践考验, 而主要是缺乏抽词词典或缺乏完善的抽词词典。因为只有抽词软件而无抽词词典, 是不能建立自动抽词标引系统的。目前缺乏抽词词典, 已成为限制抽词标引技术推广应用的瓶颈。我们图书情报工作者应当积极地承担编制抽词词典的任务。而且, 可将编制抽词词典与建立实际需要的数据库相结合[18]。

此外, 张教授对汉语自动分词、单汉字检索等也进行了总结研究。

3 对情报检索语言和标引自动化的研究

在1987年出版的《情报语言学基础》最后一章"情报检索计算机化与情报语言学的发展"中, 张教授根据自己的考察, 综述了计算机技术的发展对情报语言学的重大影响[19]。近年, 他敏锐地指出, 在计算机检索正在逐步发展到互联网络阶段的新形势下, 情报语言学研究的基本课程应是如何使情报检索语言适应新的检索环境。为此, 他进行了探索。

3.1 自然语言与人工语言对应转换是情报检索语言走向自动化之路

在实现情报检索语言和文献标引自动化的道路选择上, 张教授明显坚持务实的风格, 也就是注重研究当前切实可行的技术方法。基于他对概念或标识自动转换技术可广泛应用于文献标引和情报检索的认识[20], 论证了自然语言与人工语言对应转换是情报检索语言走

向自动化之路的观点[21]。他认为自动抽词和自动转换是自动标引的主要内容。自动转换必须以自然语言与人工语言的对应为前提,通过对应表将自然语言转成人工语言。所以,把分类表和词表改造成自然语言与人工语言的对应表,是情报检索语言走向自动化的必由之路。

3.2 《中图法》的自动化

张教授认为,《中图法》应该尽快自动化,新一版《中图法》就应是一部多功能的电子版文献分类法[22]。它应是在《中国分类主题词表》的基础上编制的,并要把自然语言对应进去,具有自动抽词、自动赋号、自动赋词功能,可提供分类号、主题词、关键词三种检索途径;优化《中图法》的检索功能,可更多地放在电子版中。如《中图法》的分面化改造,就可以在电子版中实现,但排架分类号可原封不动,即利用对应转换技术,根据类目内容范围划分规则,将电子版中组配分类号标引的概念仍对应到印刷版中原来应入的分类号中去就可以了。

使《中图法》电子版具有自动赋号和自动赋词的功能,还必须编制自动抽词词典。但这种抽词词典可以说是为电子版增加自然语言与人工语言对应转换功能的副产品,只需配备一个汉语自动分词软件就行了。

3.3 易于实现的自动分类方法

为了探索比自动聚类法更为简单、可靠的自动分类方法,张教授仔细研究了基于分类表结构原理和通常的类目内容范围划分规则的自动赋号法[23]。

自动分类必须以自动抽词为前提。中文的自动抽词必须借助于用词词典,或用词词典与非用词词典并用。当然,开始的时候也可以采用半自动方式抽词。

以体系分类号法为基础的文献自动分类需要一种自动分类用分类词表。这种分类词表是词—分类号双向对应表,分为分类号—词对应表和词—分类号对应表两部分。如果使用《中图法》,需将《中图法》分类表改造成分面分类表,把词对应到相应的分面中,但原有类目内容和分类号不需要改变。词—分类号对应表由分类号—词对应表倒转过来编成的,按词的字顺排列。这个表用于对文献自动分类标引。

在自动标引过程中,将从文献题名中自动抽出的词通过与词—分类号对应表核对,赋予《中图法》的分类号,建立分类号索引,提供分类检索途径。同一题名中的词因为分属于不同的分面,其分类号也就有多个,需要确定一个主要分类号。可根据通行的分类规划,将各个方面定出一个优先次序,哪个分类号所属分面在别的分类号所属分面之前,就确定那个分类号为主要分类号。如果遇有两个分类号所属分面并列,则两个分类号均为主要分类号。

如果以分面分类法为基础的文献自动分类,其所用分类词表也分为两部分。但是,分类号—词对应表部分的编制极为简单,只要把从文献题名中自动抽出的词对应到分面分类表相应的类目下即可。词—分类号对应表也只要将分类号—词对应表倒转来就成了。

4 对未来理想情报检索语言的探索

张教授非常重视把握情报检索语言和情报语言学的发展方向,挖掘研究课题,解决新问题。他一直遐想着运用理想语言设计法和结构功能分析法,创制一种完全新颖的、能够满足人们对情报检索语言的绝大部分要求的情报检索语言结构模式。

4.1 对未来理想情报检索语言的要求

张教授给这种理想的情报检索语言设想的结构要求是: 学科聚类系统与事物聚类系统的结合(事物聚类也应当有系统性,字顺序列可作为进入事物聚类系统的手段),先组式语言与后组式语言的结合,体系分类法与组配分类法的结合,人工语言与自然语言的结合,号码标识与语词标识的结合,系统序列与字顺序列的结合,不变概念代码与可变概念体系的结合。

他给这种情报检索语言提出的功能(性能)要求是: 分类法与主题法彻底一体化; 充分

发挥情报检索语言对知识进行系统组织和对自然语言进行规范控制的功能; 用户可十分方便地进行标引和检索; 概念可不断增补及概念的代表词可进行更换; 用户区别不出是自然语言还是人工语言, 而其实是由严密的人工语言控制; 修订不受已标引文献所牵制, 故分类体系可逐步完善; 并可以挂接英文索引、分子式索引等以及可用于机助标引。

4 2 未来理想情报检索语言的多样性结合

张教授认为, 构成这种理想情报检索语言性能的原理和方法都已存在, 难题在于找到它们的结合方案。经过长期的思索, 他找到了达到上述要求的方案。张教授发表两篇论文[24, 25], 具体描述了他所设计的理想情报检索语言的原理和方法, 他将这种模式的情报检索语言命名为"学科—事物概念组配型检索语言"。

他设计的这种检索语言在本质属性上实现了多样化的"结合":

(1) 学科聚类系统与事物聚类系统的结合。它能够同时提供学科分类系统和事物分类系统, 两者以"学科—面—点: 事物—面—点"或"事物—面—点: 学科—面—点"的形式构成先组式标引句, 既可按学科完全集中文献, 又可按事物完全集中文献。

(2) 先组式语言与后组式语言的组合, 体系分类法与组配分类法的结合。这种检索语言按其本质是一种后组式语言, 但当它采用上述形式的标引句并将其系统排序时, 就具有了先组式语言的性能。同时, 在计算机检索系统中, 采用上述形式的标引句, 并不妨碍进行自由组配检索, 因而它并未丧失后组式语言的性能; 它既是体系分类法, 又是组配分类法。

(3) 人工语言与自然语言的结合。在这种检索语言中, 检索系统内部用于标引文献的, 既不是分类号, 也不是概念词或自然语言词, 而是用概念代码。分类号、概念词和自然语言词都与概念代码对应, 作为概念代码的外部形式, 三者在标引和检索中可任意使用, 通过计算机与概念代码自动转换。自然语言可大量使用, 任意增补, 但在系统内部是受到控制的。

(4) 号码标识与语词标识的结合, 系列序列与字顺序列的结合。在这种检索语言中, 既使用号码标识(分类号), 也使用语词标识(概念词和自然语言词), 两者完全对应, 具有等价关系。可以同时提供系统序列和字顺序列, 也即具有分类检索语言和主题检索语言双重特征。

(5) 不变概念代码与可变概念体系的结合。设置不变的概念代码是这种检索语言的特异之处。概念代码是学科和事物概念的登记号, 它固定不变, 始终与某一概念相对应, 文献实际上是用这种代码来标引的。这样, 不变的概念代码是这种语言的主体, 而分类号、概念词(代表某个概念的正式词)、自然语言词都与概念代码相对应, 仅仅是概念代码的索引而已。概念代码仅在系统内部使用, 标引人员和检索人员使用的仍是分类号和语词。采取这种措施, 就使概念分类体系具有了极大的灵活性, 分类体系的改变与对文献已作的标引无关, 概念词和自然语言词的字面上的改变也与对文献已作的标引无关。

4 3 未来理想情报检索语言的实现方法

这种检索语言的主要实现方法是"分面分析+ 概念代码+ 概念对应转换+ 数据库技术"。

它对概念进行分面化处理, 将分面分为学科面部分和事物面部分, 每一部分再分为第一层的分面(学科或事物)和第二层的分面(学科的问题或事物的部分)。

该检索语言由多种文档组成。其中, 主文档设置下列字段: 概念代码(分 2 个子字段); 分类号(分 4 个子字段); 概念词或自然语言词; 子文档区分号; 临时分类; 轮排标志; 参照和注释。由主文档生成概念代码索引文档、概念分类索引文档、概念词索引文档、语词轮排索引文档。概念分类索引文档、概念词索引文档、语词轮排索引文档都可用于联机标引, 也可作为检索途径转入供检索用文档。对词典的各种修改都在主文档中进行, 修改后重新生成

各种索引文档和供检索用的文档。此外,他还设计了文献标引数据文档和检索用文档的结构。

这种情报检索语言模式,在全国继承现有情报检索语言成果的基础上,为下个世纪的情报检索语言设计提供了一种新的思路,是一个创举。即使是美国国家医学图书馆主持开发10余年的UMLS(一体化医学语言系统)的超级叙词表(Metathesaurus),虽然是以概念为中心,但作为一个整体,它主要是为后控制使用,没有建立分类体系来实现分类主题一体化,概念代码不用于标引[26]。因此,没有达到张教授所设计的理想情报检索语言模式的水平。

参考文献

1 吴建中. 21世纪图书馆展望——访谈录. 上海: 上海科学技术文献出版社, 1996
2 范并思. 论图书馆学学科前沿的转移. 图书馆, 1993(4)
3 Sievert, Mary Ellen C. Full-text Information Retrieval: Introduction. Journal of the American Society for Information Science, 1996, 47(4)
4 张琪玉. 情报语言学基础(增订二版). 武汉: 武汉大学出版社, 1997
5 张琪玉. 论情报检索语言的研究、创制与普及. 图书情报知识, 1983(4)
6 张琪玉. 自然语言检索研究进展. 见: 马费成主编. 知识信息管理研究进展. 武汉: 武汉大学出版社, 1998
7 张琪玉. 世纪之交中国情报语言学发展之路. 图书馆杂志, 1997, 增刊
8 张琪玉. 情报检索语言的发展趋势(与吴建中的对话). 图书馆杂志, 1996(4)
9 张琪玉. 自然语言检索中各种因素对检索效率的影响. 情报理论与实践, 1997(5)
10 张琪玉. 自然语言在情报检索中的应用. 情报理论与实践, 1996(3)
11 张琪玉. 论后控制词表. 图书情报工作, 1994(1)
12 张琪玉. 论自由标引. 图书馆学刊, 1995(5)
13 张琪玉. 情报检索语言. 武汉: 武汉大学出版社, 1983
14 张琪玉. 汉语关键词法探讨. 图书馆杂志, 1993(4)
15 张琪玉. 人—机结合的题内关键词索引可回避汉语分词难题. 图书馆杂志, 1993(4)
16 张琪玉. 汉语关键词索引的一种编制方法. 图书馆理论与实践, 1998(1)
17 张琪玉. 汉语关键词索引的另一种编制方法. 图书馆理论与实践, 1998(4)
18 张琪玉. 缺乏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, 1998(2)
19 张琪玉. 情报语言学基础. 武汉: 武汉大学出版社, 1987
20 张琪玉. 概念或标识自动转换技术的应用. 图书馆杂志, 1998(6)
21 张琪玉. 自然语言与人工语言对应转换——情报检索语言走向自动化之路. 中国图书馆学报, 1996(1)
22 张琪玉. 情报检索语言走向自动化之路与《中图法》发展新目标. 北京图书馆馆刊, 1996(4)
23 张琪玉. 分类法主题法一体化自动标引系统的基本原理和方法. 图书馆论坛, 1995(6)
24 张琪玉. 学科—事物概念组配型检索语言——关于情报检索语言的遐想与求索. 图书馆杂志, 1997(2)
25 张琪玉. 探索21世纪的情报检索语言. 北京大学学报: 信息管理系建系50周年专刊, 1997
26 U. S. Department of Health and Human Services et al. UMLS Knowledge Sources. National Library of Medicine, 1999

曹树金　中山大学信息管理系副教授。通讯地址: 广州市。邮编 510275。
罗春荣　中山大学图书馆副馆长。通讯地址同上。
汪东波　国家图书馆业务处处长。通讯地址: 北京白石桥路39号。邮编 100081。

(来稿时间: 1999-05-14。编发者: 李万健)