

论张琪玉先生超前的情报语言学思想^{*}

侯汉清¹, 薛春香²

(1. 南京农业大学信息科技学院, 南京 210095; 2. 南京理工大学经济管理学院, 南京 210094)

摘要: 从张琪玉先生创立情报语言学学科的历史背景入手, 剖析先生在情报检索语言的性质、情报检索语言与自然语言关系、文献标引工作自动化等方面超前的学术思想, 论述先生超前的情报语言学思想和引领性的研究成果对我国图书馆学情报学学科的卓越贡献。

关键词: 张琪玉; 情报检索语言; 情报语言学; 学术思想

中图分类号: G350

DOI: 10.3772/j.issn.1673—2286.2014.08.005

1954年张琪玉先生毕业于北京大学图书馆学系。60年来他先后在文化部社文局图书馆管理处、新疆维吾尔自治区图书馆、武汉大学图书情报学院、空军政治学院上海分院等单位工作过, 将全部身心倾注于图书情报事业的实践、研究和教学, 为我国图书馆学情报学的发展, 尤其是情报语言学的发展作出了开拓性的贡献。他知识渊博、经验丰富、眼光独到、勇于创新, 因而在图书情报领域的不少方面创造了一系列令人难以忘怀的记录, 例如:

他最早(1955年)在国内专业杂志上介绍《苏联图书分类法草案》^[1];

他最早(1956年)设计县图书馆馆舍的建筑标准;

他最早(1950年代)为新疆维吾尔自治区图书馆存储图书馆设计专用的木制可移动书架;

他最早(1963年)系统研究列宁的图书馆学思想, 编纂《列宁与图书馆》大型文献目录^[2];

他最早(1963年)关注图书馆书目工作自动化^[3];

他最早(1964年)向国内同行介绍美国刚刚问世的《科学引文索引》^[4];

他最早(1983、1987年)撰写、出版情报语言学的专著《情报检索语言》和《情报语言学基础》;

……

下面专门讨论张琪玉先生超前的情报语言学思想和相关的具有前瞻性、引领性的研究成果。

1 情报检索语言的提出

检索语言这一术语在20世纪60-70年代中外情报学家的著作(包括译作)中曾经出现过, 但是系统地论述检索语言或情报语言学、并建立一个完整的学科体系, 当属张琪玉先生的独特贡献。

长期以来, 分类法、主题法、索引法研究在中外图书馆学、目录学中, 应当属于“显学”, 研究学者之多及其出版的著作之多, 当属“名列前茅”。20世纪以来这方面的国内著名专家学者就有蒋元卿、姚明达、刘国钧、洪业、杜定友、皮高品、钱亚新、丁柯、陈长源等一批学者, 编纂出版了几十部图书分类法。他们为图书分类法、标题表和各类索引的编制、相关理论方法的建立和完善, 都作出了自己的贡献。

20世纪60年代到70年代, 随着图书情报事业的迅速发展, 英美等国的专家, 诸如英国伦敦分类法研究小组的米尔斯、布立斯、维克利、艾奇逊、奥斯丁, 印度的阮冈纳赞、美国的鲁恩、谢拉、加菲尔德、兰开斯特等学者编纂、出版了一大批新的叙词表、标题表和分类表, 同时他们也撰写、出版了一批相关的理论著作, 大大推进了这一领域的研究和教学。

为什么张琪玉先生会在20世纪80年代创立一个情报检索语言或情报语言学学科呢? 我们以为, 除了学习、借鉴这些中外学者的成果以外, 还有两个重要的历史背景。

^{*} 本研究得到中央高校基本科研业务费专项资金(编号: 30920130132005)资助。

其一,十一届三中全会以后,国内出现了否定文革、批判极左思潮,全面拨乱反正的大好政治形势。文革前及文革中被当作“毒草”批判的中外分类法,诸如《国际十进分类法》、《杜威十进分类法》、《苏联图书分类法》以及国内60年代编制出版的《大型图书馆分类法》等一大批分类法在这一时期都得到了重新评价和使用。这大大促进了国内分类法和主题词表的编纂和理论研究。

其二,“748工程”的启动。748工程是一场科技革命,即国家重点科技攻关项目“汉字信息处理系统工程”,其中包括中文编辑排版系统和中文情报检索系统两个子项目。后者又包括了《汉语主题词表》的研制。从1974年开始,以中国科技情报研究所和北京图书馆为代表的大型情报机构和图书馆聚集了一批国内的一流专家学者,合作攻关,终于在1975年正式出版了《中国图书馆图书分类法》,在1980年出版了卷帙浩繁的《汉语主题词表》。此后又陆续出版了几十种专业主题词表,出现了全国性的“分类法和词表编纂和研究热”。

张琪玉先生这一时期作为专家参加了这两大成果——《中国图书馆图书分类法》和《汉语主题词表》的编纂和出版。在这些科研活动中,张琪玉先生积累了大量的分类表和词表编纂的新经验、新方法,同时也遇到了科研活动中一系列需要回答的问题,诸如,分类法到底是什么性质和功能,为什么还要继续沿用封建时代或资本主义国家编制的大型分类法,分类法和主题法二者到底是什么关系,分类法的编制和使用是不是一定要坚持“三性原则”等等。

这些问题都需要从理论上得到完满的回答和解释。可以说,正是这些社会需求和技术进步、这些挑战和机遇,促成了20世纪80年代张琪玉先生情报语言学学术思想的问世和发展。

2 情报检索语言的性质

1963年国家科委下令全国销毁《国际十进分类法》,《科技情报工作》杂志发表评论员文章指出,“分类法是阶级斗争的晴雨表”。在这种观念的影响下,数以千计的《国际十进分类法》中译本被付之一炬。“城门失火,殃及池鱼”,这场大火同时也毁掉了我国自编的《中小型表》、刚问世的《大型图书馆分类法》,随之而来的是文革中图书馆事业以及图书馆学、情报学研究的瘫痪。

“长期以来,由于左倾思潮的影响,往往用对旧分类法的批判来代替对新分类理论的研究,用对分类法思想体系方面问题的讨论来代替分类技术和方法的研究,诸如‘五大部类’、‘三性原则’、‘大类序列’等问题的讨论和分析几乎充斥了整个分类法理论研究的领域”^[5]。

如何才能走出这样的“怪圈”,跳出分类法的阶级性、思想性的桎梏,针对这种状况,张琪玉先生70年代末解放思想、独辟蹊径,不仅把分类法、主题法综合在一起进行全面系统的研究,而且把它们看作是一种情报检索的语言工具。大家都承认,语言是没有阶级性的,因而阶级性就不再是分类法主题法这些检索语言的基本属性,从而回避了分类法编制、评价的“三性原则”之争,把研究方向转到如何提高检索语言的检索效率上。张琪玉先生这一开拓性的研究对我国情报语言学的实践和研究起到了积极的导向作用。

后来,1989年底张琪玉先生发表《情报检索语言的国家特点、时代特点和自然语言特点》^[6]一文,独具慧眼地对检索语言的性质和特点进行了深入的研究。

首先,他分析了这些特点的表现形式和形成原因。这里所说的国家特点和时代特点,不同于一般所说的阶级性和思想性,通常表现为选词和列类的详略、排列次序的先后、术语的采用等,都要打上编者所处国家和时代的印记,都会突出本国本时代。“这些主要是由于客观的社会因素,而并非某些创制者的主观因素所致。有许多已经成为分类表和词表的惯例”。

其次,他指出,对带有国家痕迹和时代痕迹的中外分类法和词表的苛求,是违反历史唯物主义的。它们虽然很难适应本国当代标引文献的需要,但是它们仍具有一定的参考价值。

第三,他指出,体现国家特点并不是情报检索语言的主要要求,主要要求是它对文献标引和情报检索的适应性,能够保证较高的检索效率。

张琪玉先生这些对检索语言性质特点的精湛而全面的分析,对于广大图书情报工作者解放思想、拨乱反正,澄清一些错误观念,开拓情报检索语言研究的新方向、新领域,都产生了重要的引领作用。

3 情报检索语言与自然语言

检索语言与自然语言的关系一直是情报语言学领域的争论话题之一。张琪玉先生对两者之间的关系一

直持辩证态度,不拘于人云亦云,科学地、发展地论证两者之间的关系,体现了其学术思想的超前性。

(1) 辩证看待检索语言与自然语言的关系

图书情报界对自然语言有两种截然相反的看法,一种传统的看法认为自然语言词量大,缺乏词形、词义和词间关系的控制,因此根本不能代替经过严格词汇控制的受控语言,不能用于信息组织和信息检索。随着搜索引擎和网络检索的出现,又出现另外一种看法,即认为在网络环境中只能唯一地适用自然语言,受控的检索语言必将被淘汰。

实际上,检索语言 and 自然语言各有优缺点,在情报检索发展的不同阶段对于两者的认识和使用也不一样,因此,不能僵化地看待两者之间的关系。张琪玉先生针对检索环境的变化,一直动态地、前瞻性地探索检索语言与自然语言的关系与发展。

在20世纪80年代人工检索盛行阶段,张琪玉先生撰文指出,“任何一种情报检索语言,都要借用某种自然语言作为检索词,或对分类号、代码进行释义”,“因此,自然语言特点成为情报检索语言的一种属性”^[7]。到上世纪90年代,计算机开始在各个领域普及应用,张琪玉先生适时地提出在研究检索语言的同时,要关注自然语言检索的研究,他认为“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”^[8]。

在信息检索计算机化和网络化发展的新环境下,各种自然语言盛行论甚嚣尘上,“自然语言会取代情报检索语言”、“人工语言不适用于网络环境”、“自然语言检索是情报检索语言发展的最高阶段”,张琪玉先生冷静客观地分析道“这种理解是似是而非的。没有任何控制的检索用语言是不可思议的”^[9]。并且,张琪玉先生及时调整研究方向,抱病前往深圳驻点调查研究,重点关注网络信息检索的新形势和新特点,情报检索语言如何适应未来的新需要。在自然语言检索、自然语言与情报语言的结合、未来的情报检索语言模式、网络资源检索工具等方面进行探索。这些研究也推动了网络环境中文献标引工作的自动化和智能化进展。

(2) 面向自然语言检索的“后控制”模式

早在1983年,张琪玉先生就提出在研究检索语言的同时要关注自然语言检索的研究。1993年,在对自然语言检索深入研究的基础上,借鉴国外情报学家兰开斯特^[10]等人的成果,他提出“文献检索过程中,如果要

达到较高的检索效率,控制是永远需要的”,并针对标引阶段和检索阶段的控制提出了四种文献检索系统模式,如表1所示^[9]。其中,“标引不控制+检索控制”模式是一种后控制模式,属于自然语言检索。

表1 文献检索系统控制模式

模式	标引阶段	检索阶段
标引控制+检索控制	情报检索语言	情报检索语言
标引控制+检索不控制	情报检索语言	情报检索语言或自然语言
标引不控制+检索控制	自然语言	自然语言或后控制词表
标引不控制+检索不控制	自然语言	自然语言

这种后控制模式通过配备后控制词表能够有效提高自然语言的检索效率,克服自然语言检索中的漏检和误检。这种后控制的检索思想后来为搜索引擎所接受,目前搜索引擎的相关检索和检索提示实际上就是采纳了这种后控制思想。

对于后控制词表的设计和编制,张琪玉先生提出了一整套的意见和建议。这对于应用于搜索引擎的同义词环这种新型知识组织系统的设计与构建具有重要的指导价值。

(3) 检索语言与自然语言发展路向

计算机和网络环境下,信息检索的大众化使得自然语言检索成为主流趋向,自然语言检索必然还要继续向前发展,但“人工语言在检索中的控制作用是自然语言所无法代替的”,“自然语言检索系统并不排斥人工语言,高级的自然语言系统必然是与人工语言或其原理结合的”^[11]。从现在搜索引擎智能化发展趋势来看,本体、知识库等概念界定严谨、概念关系控制更加严格规范的知识组织系统的构建与应用,体现了在面向普通用户的自然语言检索系统中,底层的、由计算机来实现的控制实际上更加显著。这与张琪玉先生很早就提出的“自然语言的前途还是要走向控制、规范”的观点不谋而合。

对于网络环境下两者的发展,张琪玉先生提出,“从两者的初步结合到完全融合,即情报检索语言的 自然语言化、自然语言的情报检索语言化”。他还进一步分析,这种融合的实现方式包括:①两者并用;②人工语言增加自然语言成分或自然语言增加人工语言成分;③两者之间的对应转换。其中第三种方式的结合程度最为密切^[12]。在情报检索计算机化发展的今日,情

报检索语言必须走向自动化。张琪玉先生从概念或标识的自动转换技术入手^[13],充分利用入口词原理,建议大量编制自然语言词表来建立自然语言与情报检索语言的对应转换,从而实现情报检索语言的自动化;同此,在自然语言检索系统普及和快速发展的今日,情报检索语言将成为自然语言强有力的后控制手段,自然语言的情报检索语言化是自然语言实现规范控制之途。

4 文献标引自动化

张琪玉先生自上世纪70年代就开始研究检索语言,这期间信息组织与信息检索工作遭遇计算机技术的巨大冲击。在计算机应用尚未普及时,张琪玉先生已经敏锐地意识到计算机技术对检索语言的影响^[14],他认为“自动抽词标引和自动分类标引,以及情报检索语言创制和使用中其他方面的自动化难度极大,进展缓慢,长期未能突破,但这是决不可放弃的方向”^[15]。因此,张琪玉先生身体力行,较早地展开文献自动主题标引和分类标引的理论和实践研究。高手如林的计算机学界在这方面已经取得了一些进展,在技术力量和软件开发能力方面,把我们图书情报界远远甩在后面。如何扬长避短,发挥我们的专业优势,张琪玉先生在这一领域带领我们走出了一条自己的路。

(1) 概念或标识自动转换技术

张琪玉先生提出概念或标识对应转换原理,认为事物概念、自然语言语词和情报检索语言标识三者之间可以对应转换,在情报检索语言创制、文献标引和情报检索的各个过程中都要利用概念、语词、标识对应转换的原理和方法^[13]。这种转换具体包括自然语言与人工语言之间的自动转换、人工语言之间的转换。传统环境下这些转换由人工完成,但在文献标引和情报检索的计算机化条件下,大量的转换过程必须由计算机来完成。这种自动转换技术可以广泛应用于情报检索的自然语言接口、检索后控制,文献标引过程中的自动赋词标引、自动赋号标引等,各种情报检索语言的兼容互换如分类主题对应表、分类法互操作、主题词表互操作等等。张琪玉先生对概念与标识转换的模式、方法、工具以及转换过程中存在的失真、粗化、模糊化等问题都进行了探讨^[16],这些工作对后来的研究起到了启迪和指引作用,尤其在文献标引自动化实践中,基于概念标识对应转换原理和自动转换技术,图书情报界走出了一条

与计算机学界不一样的路。图1^[17]揭示了文献标引自动化的三种类型,其中实现赋词标引依赖于自然语言语词与人工语言标识的对应转换,自动赋号标引则依赖于自然语言语词与分类标识、主题标识与分类标识的对应转换。

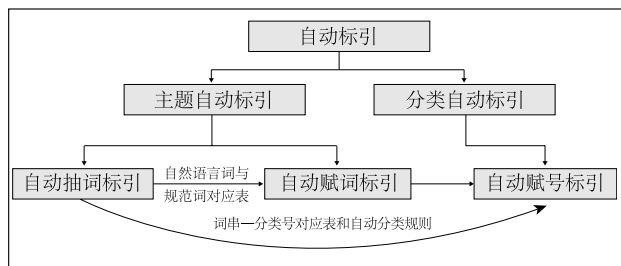


图1 文献自动标引方式及其关系

(2) 自动抽词标引

分词曾经是困扰中文信息处理的一个瓶颈,在计算机、语言学等领域的共同努力下汉语自动分词的精度已经大大提高,达到了相当高的准确率。但简单基于分词结果的文献主题标引往往存在专指度低、准确率不高的问题。因此,在文献标引自动化研究初期,张琪玉先生就提出了自动抽词标引的效果更好,但缺乏抽词词典是制约自动抽词标引普及的主要原因,这也正是适合图书情报工作者的研究课题^[18]。专业词表、文献、标引成果都可以成为抽词词典收词的来源。基于张琪玉先生提出的构建抽词词典提高自动抽词标引效率的思想,南京农业大学信息管理系在为上海图书馆开发《全国报刊索引》数据库自动标引系统时,按照学科从《中图法》类目索引、《中分表》以及海量的标引数据中抽取了大量的专业词汇,建立了大规模的抽词词典用以实现科技文献的自动抽词标引。

此外,针对当时研究中更注重各种自动标引算法的寻找,而忽视了对自动标引软件设计之外其他问题的深入、系统研究,张琪玉先生提出在自动抽词标引研究中还应该关注:①学科或主题领域特征词的研究和领域特征词表的编制;②文献结构、叙述模式和表述主题用词规律的研究;③利用分类表词表以及标引成果和术语整理成果来编制特征词表;④引入后控制词表成为特征词表的组织形式;⑤人机结合以提高标引准确率^[19]。这些提醒推动了图书情报界学者从构建词典资源、探究不同位置标引源对主题概念的贡献,以及从人机结合工作模式的视角来研究自动标引抽词效率和准确率的提高,走出了一条有别于计算机领域侧重算法的

自动标引之路。

(3) 自动分类(自动赋号标引)

张琪玉先生对于自动分类的研究融合了概念分面组配和分类主题一体化思想,提出了一种较易实现的自动分类模式:以文献题名作为主要抽词对象;适当采用人工辅助抽词措施以保证抽词质量;采用分面分类法模式以减少概念归类的复杂性;控制自动归类诸环节;利用关键词—概念词—分类号对应表以提供自然语言入口;保留原始关键词以提供自然语言检索功能。形成一个由“软件+分面分类法+关键词-概念词-分类号对应表(抽词与归类一体化对应词表)+数据库”构成的自动分类系统^[20]。

其中,采用分面分类法模式能够有效减少各种复杂、仿分导致的主题概念归类的复杂性。无论是在构建用于自动分类的分类—主题对应表过程中,还是在自动归类过程中,将时间、地点、文献类型等分面从学科主题分面中剥离出来,都能有效提高分类—主题转换的准确性,减少其他因素对于学科主题的干扰。

分类主题一体化是建立概念语词与分类标识之间的对应关系,是基于抽词结果自动赋予分类号的自动分类模式的实现机理。张琪玉先生提出的分类主题一体化自动标引系统的原理、方法和实践^[21],从图书情报工作实际的文献标引需要出发,探索出一套有别于计算机学界的文献自动分类方法。受此启发,我们在为上海图书馆研制《全国报刊索引》数据库自动标引系统时,从《中图法》类目索引、《中分表》以及海量的人工标引成果中获取概念语词与分类标识之间的对应关系和同现情况,通过利用数据挖掘中关联规则发现算法建立关键词串—类目对应表(即分类知识库),取得了较好的分类效果和实际应用。这种分类主题一体化自动标引系统,具有分类层次深、算法复杂度低的优点。我们总结项目时都深有感触地说,张琪玉先生的情报语言学思想是我们项目取得成功的保障。

5 结语

张琪玉先生是我国情报语言学的开拓者和引领者。他在上世纪70年代末,突破图书馆分类法研究“三性原则”的怪圈,另辟蹊径将分类法和词表定位为语言工具,开创了情报检索语言的新视角。面对情报检索系统的计算机化,辨证发展地探究自然语言与检索语言的关系和发展路向,探索网络环境中情报检索语言的新

模式,身体力行进行文献标引工作自动化理论与实践研究。这些研究成果都指引着图书情报界的学人在计算机化、网络化信息环境中,走出一条不同于计算机学界的信息加工、组织和检索之路。

参考文献

- [1] 张琪玉.《图书馆工作》编者按(关于杰斯林科《论苏联图书分类法草案》一文)[J].图书馆工作,1955(1).
- [2] 张琪玉.列宁与图书馆(文献目录)[M].油印本,1973.
- [3] 瓦西里也娃-切波塔辽娃.书目工作自动化[M].张琪玉,译.1963.
- [4] 张琪玉.美国出版的《科学引文索引》[J].综合科技动态 第二分册 情报工作,1964(5).
- [5] 侯汉清.分类法的发展趋势简论[J].情报科学,1981(1):58-63.
- [6] 张琪玉.情报检索语言的国家特点、时代特点和自然语言特点[J].图书馆理论与实践,1989(4):9-10.
- [7] 张琪玉.关于自然语言检索问题[J].图书馆论坛,2004,24(6):211-213,145.
- [8] 张琪玉.张琪玉与情报语言学学科建设[A]//张琪玉情报语言学文集.北京:北京图书馆出版社,1999.
- [9] 张琪玉.论后控制词表[J].图书情报工作,1994(1):1-4.
- [10] Lancaster, F. W. 情报检索词汇控制[M].侯汉清,戴维民,陆宝树,译.上海:同济大学出版社,1992:87-93.
- [11] 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[J].中国图书馆学报,1996(1):37-40.
- [12] 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[J].北京图书馆馆刊,1996(4):25-27.
- [13] 张琪玉.概念或标识自动转换技术的应用[J].图书馆杂志,1998(6):9-13.
- [14] 张琪玉.情报语言学基础[M].武汉:武汉大学出版社,1987.
- [15] 张琪玉.寻找更佳结合模式是情报检索语言创新的主流[J].图书馆杂志,2005(2):22-25.
- [16] 张琪玉.情报检索全过程中概念与标识的对应转换[J].图书与情报,2002(2):33-34.
- [17] 戴维民.信息组织[M].北京:高等教育出版社,2004:250.
- [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[J].图书与情报,1998(2):27,80.
- [19] 张琪玉.自动抽词与自动分词[J].图书馆杂志,2002(3):13-14.
- [20] 张琪玉.概念分面组配型自动分类系统[J].图书馆学刊,2002(6):9-10.
- [21] 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[J].图书馆论坛,1995(6):3-4,57.

作者简介

侯汉清, 男, 1943年生, 南京农业大学信息科技学院教授, 研究方向: 信息检索、智能信息处理, E-mail: hqhou@njau.edu.cn。
薛春香, 女, 1979年生, 博士, 南京理工大学经济管理学院副教授, 研究方向: 网络知识组织系统、文本分类、自动标引。

Discussion on Mr. Zhang Qiyu's Advanced Thoughts of Information Retrieval Language

HOU HanQing¹, XUE ChunXiang²

- (1. College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China;
2. School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: Beginning with the historical background in that Mr. Zhang set up the discipline of information linguistics, this article discussed Mr. Zhang's advanced academic thinking of information language from the following aspects: the nature of information retrieval language, the relationship between natural language and information retrieval language, automatic indexing and categorization. Nowadays, His creative theories about information retrieval language and leading research results are still making great contribution to the development of library and information science in China.

Keywords: Zhang Qiyu; Information retrieval language; Information linguistics; Academic thinking

(收稿日期: 2014-06-30)

■ 书 讯 ■

《中国高被引分析报告2012》

在您所关注的研究领域中, 哪些论文最受同行关注? 哪些研究主题最为热门? 哪些学者最具学术影响力? 哪些期刊最获同行认可? 又有哪些机构占据领域研究的制高点? 上述问题请您参考近期出版的《中国高被引分析报告2012》。

该书将理、工、农、医、人文、社科等领域划分为51个学科, 综合分析了各个学科的研究热点与前沿、高影响力论文、高影响力作者、高影响力期刊和高影响力科研机构, 并以关联图谱的方式展现各种学术关系, 有助于科研人员及时发现并跟踪研究热点, 可为科研管理机构评估科研能力提供依据, 还有利于期刊编辑部监测本刊学术影响力, 是高等院校、科研院所及期刊编辑部等相关单位和人员参考工具书。

该书以“中国知识链接数据库”为依托, 数据样本覆盖我国6000余种期刊的论文及引文, 分学科揭示高影响力的学者、研究机构(大学、研究所、医院等)、地区(省/自治区/直辖市)、学术期刊、图书、外文期刊和会议录, 并采用共词分析、共被引分析和合著分析等方法绘制出各学科的前沿主题分布以及作者、机构和期刊间关联的知识图谱。

《中国高被引分析报告2012》由中国科学技术信息研究所编著, 科学技术文献出版社出版, 全书86万字, 定价298.00元。