张琪玉教授对自然语言检索的研究

司莉，杨君正
（武汉大学信息管理学院，武汉 430072）

摘要：张琪玉教授是我国情报语言学创始人，在情报检索语言领域做出了卓越的贡献。本文阐释了张琪玉教授对自然语言检索的主要研究内容，包括自然语言在情报检索中实际应用，影响自然语言检索效率的因素和提高效率的对策，以及编制自然语言词表对自然语言检索进行控制。张琪玉教授认为自然语言检索必然会得到进一步的发展，同时，自然语言与情报检索语言两者不可相互取代而是相互结合或融合。

关键词：张琪玉；自然语言检索；自然语言词表
中图分类号：G350
DOI：10.3772/j.issn.1673—2286.2015.02.007

1 引言

从20世纪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[1]。张琪玉教授是国内较早关注自然语言检索的学者之一。他对自然语言检索的研究最为全面、系统，其主要研究成果集中反映在《情报检索语言实用教程》[2]的第五章中。

张琪玉教授作为我国情报语言学的宗师，早在1983年就曾提出应该对自然语言检索进行研究，他在《情报检索语言》[3]一书中就专辟第七章介绍“关键词描述语言”，是取自自然语言，而不作规范化处理，或极少量的规范化处理的一种主题法系统。关键词法是国外自然语言应用于情报检索最为成熟、最为普遍的一种方式。张琪玉在上个世纪90年代开始将研究的重点转向自然语言检索，并发表了相关论文，如《自然语言在情报检索中的应用》、《自然语言检索中各种因素对检索效率的影响》、《关于自然语言检索问题》等。张琪玉教授在自然语言检索领域做出了卓越贡献。本文主要采用文献研究的方法，总结张琪玉教授自然语言检索学术思想，为当今自然语言检索的发展提供借鉴。

2 自然语言在情报检索中的应用方式

在情报检索中应用自然语言，其实质就是使用文献作者原来所用的语词，或文摘编写者原来所用的语词，或标引人员自拟的而不是取自词表的语词来作为文献检索标识[4]。

张琪玉教授认为自然语言在情报检索中应用的方式有：①关键词法；②文本检索；③单汉字检索；④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；⑤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；⑥自动赋检索词和自动赋分类号；⑦自动分类（自动聚类法）；⑧自由标引。在之后的研究中，张琪玉教授又将自然语言在情报检索中的应用进一步归纳为6个方面：①关键词索引及以关键词为检索标识的文献数据库；②全文数据库；③搜索引擎及由搜索引擎自动建立的网络资源数据库；④自动甄别（知识本体语言）；⑤自动标引；⑥自动分类。在实际中应用中，关键词索引及数据库、全文检索、搜索引擎已经实现，而且实质都是关键词检索。

3 影响自然语言检索效率的因素与对策分析

为了提高自然语言检索效率，张琪玉教授从情报语言学角度对自然语言检索效率的各种影响因素进行了深入探讨，包括检索依据的文本类型（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等。

针对这些影响因素，张教授提出提高自然语言检索效率的对策。他认为应该：①在自然语言系统中，对文本的题名、小标题和章节名、摘要、正文应分别标注，以便在抽词或检索时有所选择。②如果对文本进行抽词，应尽量抽取专指词。③检索用语优先使用专指词，需要扩检时再使用较泛指的词。④在进行组配检索时，最好在句、段范围内检索。⑤构造检索表达式时，尽量要把同义词、近义词、反义词、否定词等用“逻辑和”连接起来包括进去。⑥配备后控制词表是提高自然语言系统检索效率的重要措施。⑦应采用人工自由标引，以采用自由标引+文本字词匹配检索作补充最为理想。他的这些思想即便是在今天，对于改进目前的一些检索系统性能仍然具有指导意义[5]。

4 对自然语言检索应用的评价

4.1 关于关键词检索

张琪玉教授表示，目前仅在关键词检索的层次上自然语言检索已经实现。关键词检索是以出现在文献题名或正文及文摘中的描述文献主题内容的关键词为标目的字顺索引。起初，关键词索引是“用作检索刊物的临时性索引（期索引）”，后来“数据库的关键词检索”才被重视起来，用来“代替人工标引”。再后来，关键词则应用于“自动主题标引和自动分类研究的前期处理”。张教授在分析关键词索引应用的历程后，提出关键词目前主要用于：①题录数据库；②全文数据库（文本数据库）；③自动抽取关键词，可用于全文数据库索引库的建库，以方便检索[6]。

张琪玉教授认为关键词有很大的优点：“可以利用计算机抽取，速度极快，索引深度相当大，对标引人员要求最低”。同时关键词也有很大缺点：“不规范，检索效率不高”，“检准率有时很低，往往达到使用户无法容忍的地步”[7]。自由标引的关键词检索效果较好，但不能自动抽取。自由标引的关键词与自动抽取（包括自动匹配）的关键词、题名中的关键词与正文中的关键词有很大差别；关键词与规范词在质量上更有很大差别。

4.2 关于全文检索

全文检索是文本检索的一种，采用任意字词匹配检索技术，是关键词检索的一种应用。全文检索是目前关键词检索技术的主要用途。单纯的全文检索系统，张琪玉教授将其检索性能概括表示为“关键词检索+计算机辅助文本浏览”。

对于一些人认为全文检索可以满足一切的检索需要，可以替代所有其他检索方法的观点，张琪玉教授并不赞同，他认为全文检索系统能够适应的必须有一定的检索要求：①允许使用任意词乃至词的片断，从文本中进行匹配查找；②对于用专有名词表示的检索对象，以及出现的频率很低者检索效果相当好；③非常合适诗词等全文数据库检索。同样全文检索系统不能够适应的检索要求是：①学科或专业的分类检索要求；②一族事物的族性检索要求；③被论述得过多的事物；④有较多同义词、准同义词的检索对象[8]。因此，张教授认为：单纯的全文检索系统既没有分类检索与正规的主题检索功能，所以它并不能取代主要的传统检索方法，它只是增加了一种检索功能——计算机辅助文本浏览功能，可以用关键词从文献原文中直接进行匹配并即时浏览阅读（即检即阅）。只有集成多种检索方法的检索系统（即目录体系、索引体系、具有多种功能的计算机检索系统），才能较好地满足多样性的检索要求。一个好的全文检索系统也必须是一个集成系统，是全文数据库和文献目录数据库的有机结合体。

4.3 关于搜索引擎

搜索引擎的检索是在预先用搜索软件建立的网络信息资源数据库中进行检索。张琪玉教授指出：“搜索引擎建立的数据库属于全文数据库性质，所以，搜索引擎的检索实际上就是全文关键词匹配检索”。但是面对数量庞大的网络信息资源，不可能完全采用人工标引（如分类浏览检索），目前必然的选择是“自动搜索建立网络信息资源数据库和对数据库进行关键词检索”，然而关键词全文检索的现状“并不能令人满意”。

4.4 关于自动标引和自动分类

自动标引是指自动抽取主题概念词标引，自动分

类是在关键词中被确认为表达文献主题概念的词的基础上进一步将其归类。对于自动标引和自动分类的相关研究已经有近半个世纪，但是始终没有突破性的进展。张琪玉教授认为自动标引和自动分类至今未能实现的关键是："计算机还不能识别文献的主题"。计算机要把文献中的关键词抽出来"是可以做到的"。但是要从所抽出的全部关键词中挑选出代表文献主题内容的词来"至今还做不好"。而自动分类是在关键词中被确认为表达文献主题概念的词的基础上进一步将其归类，故与自动标引的困难实质相同。

张琪玉教授认为，目前，自然语言在情报检索中的应用主要面临着"如何从自然语言文本中抽出（或者说确认）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"，并且"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。

5 对自然语言检索进行控制的措施

为了促进自然语言与情报检索语言相结合，张琪玉教授提出应该大量编制自然语言词表，并认为它在促进当代文献标引-检索用语言的进步中将起到关键性的作用，其原因有两点：①自然语言词表是提高情报检索语言易用化的主要手段。要使情报检索语言"自然语言化"，是离不开某种类型的自然语言词表的。用自然语言词表作为辅助的情报检索语言标引和检索将是一个重大的进步。②自然语言词表是对自然语言加以控制的主要形式[9]。在自然语言词表辅助下的自然语言检索，相对于无自然语言词表辅助的自然语言检索来说，检索效率必然会有较大提高。所以，张教授认为有自然语言词表控制的自然语言检索将会是一个重大的进步。张琪玉教授关于自然语言检索控制措施的研究主要体现在《论后控制词表》、《积极为自然语言与情报检索语言的结合创造条件——建议大量编制自然语言词表》、《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》等文章中。

5.1 自然语言接口用对应表

自然语言接口实际上是在情报检索语言检索系统之前安置一个自然语言语词与情报检索语言语词的对应表，其前端（被转换的源词字段）为自然语言的语词，后端（所转换成的目标词字段）为情报检索语言的语词。检索人员（或标引人员）使用自然语言的语词表达检索课题（或文献主题）进入系统，通过对应表自动转换成情报检索语言的语词在系统中进行实际的检索（或标引）。张琪玉教授认为，对应表只是一个附加部分，并不影响原有的标引工具和标引数据，是有利无弊的，可普遍采用。

张教授提出，为使对应表一目了然以便于管理，并简化转换，可将词表正式词也作为一个自然语言词重复列入对应表，作成"词表正式词→词表正式词"对应款目。词表的双语种对照索引也可编入对应表，它其实就是入口词表的机读版。在对应表中，自然语言与情报检索语言的对应可以有一对多的关系，通过人工辅助转换。

5.2 自动抽词词典

汉语的自动抽词系统，绝大多数都是使用抽词词典的。张琪玉教授指出，"抽词词典不但是抽词标引系统实际投入使用所不可缺少的条件，而且对抽词质量还具有重大影响。抽词词典越丰富和完善，抽词的完全率和正确率越高"[10]，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"，而编制抽词词典比编制抽词软件需要多出花费更多倍的工作量。张教授认为，自动抽词标引技术（此项技术也是自动赋词、自动赋号、自动分类等的基础）难以普及的主要原因就是目前缺乏抽词词典，所以当前迫切需要大量编制汉语自动抽词词典。

5.3 自动赋词赋号用对应表

自动赋词赋号标引系统是对自动抽词标引系统的改进，使自动抽出的自然语言语词转换成情报检索语言语词（检索词或分类号）。其所用的对应表的前端（被转换的源词字段）为自动抽词所抽出的自然语言语词，后端（所转换成的目标词字段）为情报检索语言语词（检索词或分类号）。张琪玉教授认为，"自动赋词标引系统和自动赋号标引系统可以分别建立，也可以合二为一"。自动赋词赋号标引系统可使用"现有的词表或分面分类表"，也可"仿照词表和分类表的编制原理，对自动抽词所抽出的自然语言语词作有限范围的控制（如只对同义词作规范控制，或纳入粗略的分类体系）"。

自动赋词赋号标引系统还可以在赋予文献检索词或分类号的同时，仍保留自动抽词过程所抽出的原词，兼取人工语言与自然语言的优点。这样的系统用于检索时，检索者既可使用检索词或分类号检索，也可使用自然语言检索。张教授指出，“使用现有词表的自动赋词系统和使用现有分面分类表的自动赋号系统”，从检索角度看“也都是一种自然语言接口”。

5.4 自动分类用对应表

张琪玉教授指出，自动分类与自动赋词不同的在于：它使用体系分类法，自动分类得到的分类号可区分出主要分类号和非主要分类号，各个分类号的组配又可表达比原有类目更多和更专指的概念。其所用词表是一种词与分类号的双向对应表，由分类号-词对应表和词-分类号对应表两部分组成。

其中分类号-词对应表的编制法是：“假定使用中图法，需先将中图法的分类表改造成分面分类表，但原有分类号不需要改变。把自然语言语词对应到相应的分面中。由于文献主题一般都是由多个主题因素构成的，各个主题因索在体系分类表中都有其对应的类目”。

词-分类号对应表的编制法是：“将分类号—词对应表的款目倒转过来，按词的字顺排列，供自动分类标引用”。

张教授认为，在自动分类标引过程中，“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。词仍应保留，建立关键词索引，提供主题检索途径”。在对应表中“如果能将等同关系词选定一种词形为正式词，其余为非正式词，提供非正式词转换成正式词的功能，则更好”。

5.5 后控制词表

张琪玉教授在《论后控制词表》[11]一文中提出，为自然语言检索系统配备后控制词表，是提高其检索效率的有效措施。后控制词表的性质类似于入口词表，它是一种转换工具，是一种扩检工具，是一种罗列自然语言检索标识供选择的工具。

张教授总结后控制词表的特点在于：“其中的控制词（也可以是分类号）并非直接用于标引，而是对作为文献检索标识的自然语言词进行控制（建立等同、等级、相关关系）。因此，在后控制词表中，标引-检索用词是自然语言，非标引-检索用词却是人工语言，这与在一般词表中的情形正好相反。后控制词表必须在检索系统中使用的自然语言检索标识的基础上进行编制（即必须以作为检索标识的自然语言原词为基础），以达到最大的覆盖率。否则将会大大降低其控制功能”[12]。后控制词表较理想的结构模式是“分类词表+字顺/轮排表”。分类词表可以设置“大类→小类→控制词→自然语言词”的等级。后控制词表的编制过程包括“初始阶段和完善阶段”两个阶段。初始阶段是“后控制词表编制成形，内容还不很丰富，但可以投入使用”。完善阶段是“在使用过程中词汇不断增补积累和体系不断调整细化”。后控制词表的体系可“用现成的分类表或词表作为框架，将关键词添入”；也可“对积累到一定数量的关键词进行归纳整理，形成系统”。

5.6 词素词表

张琪玉教授指出，汉语的词素词表“具有自然语言入口功能”，它在标引文献时，文献主题概念可“全部用自然语言词自由表达”。若“表达文献主题概念的自然语言词”与“词表中的叙词”一致，或与“词表中的入口词（同义词和被组代词）一致”，都可“立即自动转换成叙词，自动将叙词登录入标引结果字段”；若“表达文献主题的自然语言词在词表中没有对应的叙词或入口词，该系统便会对自然语言词进行词素分析，利用词素相似性匹配原理，自动推荐一批含有相同词素的叙词供选择，通过人工判别，选定合适的叙词（包括若干叙词的组配）进行标引；若所推荐的词均不合适，则可将自然语言词作为自由词进行标引并同时作增补记录”。

汉语的词素词表首次见于《军用主题词表》应用管理系统，张琪玉教授总结其在应用词的相似度匹配原理是，“以相同词素的个数为统计单位，并结合叙词词素的位置特征（如词素在词尾、在词首、在词中）及长度特征进行加权，可调整权值来扩充或压缩推荐词的数量以方便选择，并加入同义词素避免遗漏等，从而使所推荐的词更有针对性和全面性”。他认为这种方法，在无形中提高了词表的入口率，无疑使标引工作更为容易。

6 对自然语言检索发展前途的见解

在图书情报界中，很多人对于自然语言检索的观点

是：自然语言检索是发展方向，信息检索要走自然语言道路；人工语言（情报检索语言）不适应网络环境，自然语言不亚于人工语言；目前自然语言虽有缺点，但人工智能可使其达到完善，满足一切检索要求。张琪玉教授则审慎的提出：自然语言的未来与情报检索语言的未来在某种意义上可以说是同一个问题。从一方面看，自然语言不可能全面取代情报检索语言、淘汰情报检索语言，情报检索语言还将继续发展；但从另一方面看，在计算机检索的条件下，自然语言有许多重要的优点，故它也必然会更进一步得到发展[13]。

张教授认为，没有任何控制的自然语言是“不可思议的”，至今也没有找到“在计算机环境下不加控制的利用自然语言”的十分有效的方法。因此他总结道：“网络检索不能唯一地使用自然语言。自然语言的前途仍然要走向控制、规范，当然，控制的方法会与过去人工语言所采用的方法有所不同”[14]。

张琪玉教授强调：自然语言检索系统与情报检索语言检索系统之间的关系并不是绝然对立的。两者各有优点而不可能互相取代，应该使两者相互结合或融合。自然语言或情报检索语言的未来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。未来发展的大趋势是“情报检索语言的自然语言化、自然语言的情报检索语言化”，大方向是“走两者结合之路”。在两者完全融合的新型情报检索语言普及以前的趋势可能是下列三种情况并存：①情报检索语言与自然语言在一个检索系统中并用；②情报检索语言增加自然语言成分；③自然语言适当引进情报检索语言的原理与方法和增加情报检索语言成分。

对于未来的研究方法，张琪玉教授表示，由于人工语言和自然语言都不可取代，因而未来对两者的研究都要重视。要“从情报语言学的角度”来深入研究“自然语言检索中存在的问题”，把“情报语言学的原理和方法”引进“自然语言检索的研究”，并且要“重视利用情报检索语言已往所积累的成果”，也要“积极研究情报检索语言在网络环境下应用中所遇到的新问题，寻找改进方法”，特别是“吸取自然语言的优点来弥补情报检索语言的不足之处”。对于这两方面的研究，张教授相信：“自然语言和情报检索语言应朝着并且必然会朝着从两者的初步结合到完全融合”。

张琪玉教授对自然语言检索的研究为自然语言检索在中国的发展和普及奠定了坚实的基础，他高屋建瓴地提出的情报检索语言的自然语言化、自然语言的情报检索语言化思想，始终引领着情报检索语言研究与实践的发展方向，为我们提出了丰富的研究课题，激励后辈在该领域不断探索与拓新，他的学术思想和学术精神也值得我们继承和发扬。

参考文献

[1] 曹树金,罗春荣,汪东波.开创情报语言学的新天地[J].中国图书馆学报,1999(5):72-77.

[2] 张琪玉.情报检索语言实用教程[M].武汉:武汉大学出版社,2004.

[3] 张琪玉.情报检索语言[M].武汉:武汉大学出版社,1983.

[4] 张琪玉.情报语言学基础(增订二版)[M].武汉:武汉大学出版社,1997:44.

[5] 张琪玉.自然语言检索中各种因素对检索效率的影响[J].情报理论与实践,1997:257-259.

[6] 张琪玉.关于自然语言检索问题[J].图书馆论坛,2004(6):211-213.

[7] 张琪玉.网络信息检索工具增强关键词检索功能的措施[J].图书馆杂志,2001(1):7-10.

[8] 张琪玉.全文数据库、全文检索与全文标引[J].图书馆理论与实践,2002(6):40,55.

[9] 张琪玉.积极为自然语言与情报检索语言的结合创造条件——建议大量编制自然语言词表(上)[J].图书馆杂志,1999(9):7-9.

[10] 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.

[11] 张琪玉.论后控制词表[J].图书情报工作,1994(1):1-4.

[12] 张琪玉.积极为自然语言与情报检索语言的结合创造条件——建议大量编制自然语言词表(下)[J].图书馆杂志,1999(10):8-10.

[13] 走向自然语言与情报检索语言结合之路:与我国著名情报语言学家张琪玉教授的通讯访谈[J].图书馆理论与实践,2001(2):3-5.

[14] 张琪玉.情报语言学的若干研究心得和收获——张琪玉学术思想自述[J].图书情报工作,2009(10):5-9,29.

作者简介

司莉，女，武汉大学信息管理学院教授，博士生导师，研究方向：知识组织与知识管理、图书馆营销与服务。

杨君正，武汉大学信息管理学院硕士研究生。

Professor Zhang Qiyu's Research on Natural Language Retrieval

SI Li, YANG JunZheng
（School of Information Management,Wuhan University, Wuhan 430072, China）

Abstract: Professor Zhang Qiyu is the founder of Chinese information linguistics and he made great contribution to the field of information retrieval language. The article illustrates the main content of professor Zhang's research on natural language retrieval, including practical application of natural language in information retrieval, factors that affect the efficiency of natural language retrieval, strategies of improving retrieval efficiency, as well as the authority control of natural language retrieval by compiling natural language thesaurus. Professor Zhang believes that there must be further development of natural information retrieval. Besides, he considers that natural language may combine with information retrieval language rather than replace it.

Keywords: Zhang Qiyu; Natural language retrieval; Natural language thesaurus

（收稿日期：2015-01-06）
编辑：雷雪