

情报语言学的若干研究心得和收获

——张琪玉学术思想自述



张琪玉

南京政治学院上海分院军事信息管理系 上海 200433

张琪玉,男,1930年6月7日生,1954年7月毕业于北京大学图书馆学系。曾任武汉大学图书情报学教授、图书馆学情报学研究所所长、空军政治学院图书档案系(现南京政治学院上海分院军事信息管理系)主任、教授,2001年3月退休。从20世纪70年代末起开拓情报语言学新领域。学术观点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。出版专著20部,主编参编专著16部,发表论文、译文约400篇,《张琪玉文库》光盘1张。

〔摘要〕著者自述情报语言学学术思想,包括情报检索语言的研究方法、检索效率、情报检索语言一些具体问题的研究心得、术语学、自然语言检索等。

〔关键词〕张琪玉 情报语言学 情报检索语言 自然语言检索 术语学

〔分类号〕G254.0

Zhang Qiyu's Academic Thoughts on Information Linguistics

Zhang Qiyu

Military Information Department, Shanghai Branch of Nanjing Political Institute, Shanghai 200433

〔Abstract〕The author expounds his academic thought on information linguistics, including the research methods of information retrieval language, retrieval efficiency, his research experience of some specific issues of information retrieval language, terminology, and natural language retrieval etc.

〔Keywords〕Zhang Qiyu information linguistics information retrieval language natural language retrieval terminology

我研究情报语言学,假如仅仅从写作《情报检索语言》一书算起,也已有30年的时间。有一些来之不易的心得和收获,介绍出来与同行们交流、分享。

我有一个基本观点,认为各种情报检索语言的基本原理是一致的,它们只是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用的方法不同,才形成了不同的类型和语种。因此,对分类检索语言、主题检索语言和其他情报检索语言以及自然语言在情报检索中的应用问题进行统一研究,可探索它们影响检索效率的共同规律和有效的改进途径。提高检索效率是情报语言学研究根本目的和核心问题。

正是从这一基本观点出发,我写成了《情报检索语言》一书以及之后的许多情报语言学文章和专著。写

作《情报检索语言》一书是我对各种情报检索语言进行统一研究,以探索它们影响检索效率的共同规律和有效的改进途径的第一次实践。

我的学术研究,以情报语言学(包括索引学)为范围。对于情报语言学,我可以说是“情有独钟”,自1980年之后写的几百篇著作,几乎全部是关于情报语言学的。

我认为,人生很短暂,一个人不可能做完一切想做的事,所以只能去做最必要的、最有意义的并且能够做到的事情。做学问,要有一个专业范围,在专业范围内力求深而广。

我把情报语言学定义为“是研究情报检索中语言保证问题的一门学科,其主要研究对象是情报检索语言,同时也研究自然语言在情报检索中的应用问题”。

这样表述,是要在概念上弄清情报检索语言与自然语言的区别,情报检索语言(简称检索语言)是人工语言,自然语言不能称为“自然检索语言”。

1 情报检索语言的研究方法

我体会到,研究的成效在很大程度上取决于研究方法。

情报检索语言的检索效率是其功能决定的,而其功能则是由其结构决定的。所以,研究情报检索语言的性能,主要就是要分析解剖其结构。结构功能分析法是研究情报检索语言最为有效的方法。后来,我总结出研究情报检索语言的一整套专用方法,如历史演进研究法、比较研究法、调查整理法、归纳法和演绎法、原理或方法的移植法、理想语言设计法、现用语言改进法、数学方法和统计方法、实验方法等,我发现这些方法都含有结构功能分析的内容,都是从结构功能分析法衍生出来的。

关于比较研究法,我写过一篇题为《情报检索语言原理的一致和方法的差异》(《图书馆建设》1994年第6期)的文章,提到比较研究法可以使我们观察出情报检索语言各种结构与功能的“异中之同,同中之异”,使我们“既见树木,也见森林”。这篇文章可以帮助人们了解各种情报检索语言的性能。正是在利用结构功能分析法及其各种衍生方法的基础上,我对情报语言学的内容进行了丰富、发展。

情报检索语言的检索效率由其各种检索功能决定,而情报检索语言的各种检索功能则由其结构产生。产生检索功能的是情报检索语言的微观结构。若干微观结构的有机结合才能构成一种情报检索语言。各种类型的情报检索语言都是一种结构模式,即其宏观结构。微观结构的数量、种类及结合模式,形成情报检索语言检索效率的整体水平。

正是这一原理,使我树立起情报检索语言可创新、可改进的信念。我常常异想天开,去寻找理想的情报检索语言结构模式,去发现新的功能,去探索创新之路。例如《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年第2期)和《探索21世纪的情报检索语言》(《北京大学学报:信息管理系建系五十周年专刊》(1997年)以及《对未来分类法的憧憬》(《图书馆理论与实践》2003年第1期)三篇文章。

《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》和《探索21世纪的情报检索语言》是经过十多年的研究才寻找到的模式。该模式的本质属性可归纳为:①学科聚类系统与事物聚类系统的结合;②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;③人工语言与自然语言的结合;④号码标识与语词标识的结合、系统序列与字顺序列的结合;⑤不变概念代码与可变概念体系的结合。其主要实现方法可概括为“分面分析+概念代码+概念对应转换+数据库技术”。这种语言的性能可概括为:①分类法与主题法彻底一体化的;②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的;③用户可十分方便地进行标引和检索的;④概念可不断增补及概念的代表词可进行更换的;⑤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的;⑥修订不受已标引文献所牵制,故分类体系可逐步完善的;⑦可以挂接英文索引、分子式索引等以及可用于机助标引的。

《对未来分类法的憧憬》一文认为理想分类法结构模式应是:①学科分类与事物分类并行又能相互结合的,既可形成学科分类体系又可形成事物分类体系,两种体系可变换的分类法,即学科-事物概念组配型的分类法;②多聚类中心的、线性结构与网状结构相结合的,学科和事物概念全向聚类的分类法,单纯线性结构和单纯网状结构都有局限性;③只依据文献内容的学科属性或文献研究对象之间的相互联系的客观事实进行分类,对其思想观点不加区分和褒贬的分类法(某种意识形态、宗教信仰、政治主张等的优先、重点揭示可作为照顾各国各民族的特殊需要来处理),即分类体系的建立更侧重于揭示各种知识的内在联系,重点不在于构建表示某种信仰的宏观框架;④国际化与民族化妥善结合的分类法。依据“求同存异”的原则,以国际通用为基本,照顾各国各民族的特殊需要,阮岗纳赞的“优惠类”原理用超链接方法实现;⑤体系列举方式与分面组配方式相结合的分类法;⑥在一个整体框架下由众多专业分类法集成的分类法;每个学科领域或事物范畴允许有不只一个分类体系,但不同体系之间大部分类目有对应转换的可能性;⑦分类体系和类目可不断革新的分类法;摆脱了已标引文献和藏书排架牵制的分类法(排架仍可用已分类了大量藏书的原有分类法);⑧能与世界上现有主要分类法通过类目对应转换达到基本兼容的分类法;⑨伸缩性很好的分类法;⑩类名与术语学成果尽可能取得一致的分类法,即其类名可作为主题词使用的分类法,实际上就是分类

法主题法一体化;⑪有详细的、完善的自然语言入口的分类法,并可用于人机结合赋号标引;⑫多语种的分类法;⑬计算机化并可在网络上应用的分类法;⑭有充分文献保证的分类法,但不是仅限于专著的文獻保证。我设想,未来分类法的编制可能是先构筑部分(专业分类法),再形成整体。事实上,未来分类法的构成原理和方法大多已存在于现有情报检索语言中,只是找出它们完善结合(实现)的方案还需要作出很大的努力。

由此可见,情报检索语言的发展方向远不止一个网络检索问题,我们必须多方面去探索情报检索语言的发展方向。

我认为,情报检索语言的进步主要是结构模式的进步。永远只有更佳而不会有最佳的情报检索语言结构模式。因此,寻找更佳结构模式永远是情报检索语言创新的主流。

我认为,目前寻找更佳结构模式的主要方向是:①学科检索与事物检索的更密切结合;②人工语言与自然语言的融合;③线性显示与网状显示的结合;④族性检索与特性检索的灵活调节;⑤简单易用与功能丰富能兼备;⑥低成本与高效益能兼备;⑦自动化和网络化;⑧既能不断改进,又能回避重新标引;⑨适应性、兼容性、民族化和国际化;⑩与术语学密切结合。这10项中的哪一项,即使是小小的改进,也符合情报检索语言发展的方向。

我在研究中归纳出情报检索语言的4项基本功能,即:①对文献的情报内容(及某些外表特征)加以标引;②对内容相同及相关的情报加以集中或揭示其相关性;③对大量情报加以系统化或组织化;④便于将标引用语和检索用语进行相符性比较。后来,我在《情报检索语言方法综述》(《图书情报知识》1984年第2期)一文中对情报检索语言的各种微观结构如何实现这4项基本功能做了系统说明。

2 检索效率

关于检索效率,我的总结是应当“全、准、快、便、省”(检全率、检准率、检索速度、检索方便性、检索成本与效益),其中最主要的是全和准。

决定检索效率的有4个方面的因素:①情报检索语言的质量;②标引质量;③检索质量;④其他。其中,情报检索语言的质量与检索效率关系最密切。

3 一些具体的心得和收获

至于一些具体的体会,限于篇幅,这里仅列举出相关的文章和出处。在这些文章中,我对情报语言学的某些问题曾做过全面、深入、系统的研究,我认为是比较重要的文章:

- 《论情报检索语言的研究、创制与普及》(《图书情报知识》1983年第4期);
- 《情报检索中的语言保证问题》(《情报理论与实践》1995年第1期);
- 《检索效率及其影响因素》(《情报理论与实践》1995年第2期);
- 《情报检索语言语法体系初探》(《图书馆理论与实践》1986年第3期);
- 《情报检索语言中聚类的原理和方法》(《北京图书馆馆刊》1997年第1期);
- 《组配及其演变》(《情报理论与实践》1996年第1期);
- 《文献主题的构成因素及层次》(《图书情报知识》1985年第1期);
- 《试论隐含主题》(《图书馆理论与实践》1993年第2期);
- 《情报检索语言的国家特点、时代特点和自然语言特点》(《图书馆理论与实践》1989年第4期);
- 《情报检索语言的易用性问题》(《云南图书馆》1990年第4期);
- 《文献标引是需要智慧的近乎艺术创造的处理过程》(《图书馆杂志》2004年第3期);
- 《情报检索全过程中概念与标识的对应转换》(《图书与情报》2002年第2期);
- 《情报检索语言的发展趋势(与吴建中的对话)》(《图书馆杂志》1996年第4期);
- 《事物分类与学科分类》(《图书馆理论与实践》2003年第1期);
- 《体系分类法的准则和惯例》(《晋图学刊》1992年第4期);
- 《体系分类法中“集中与分散”的矛盾》(《图书馆杂志》1982年第1期);
- 《体系分类法中的交替法》(《图书情报知识》1982年第2期);
- 《分类标记原理与方法概述》(《图书馆》1993年第1期);
- 《情报检索语言中语词标识的功能与局限——关于主题法性能的几点分析》(《湖北高校图书馆》1985年第1

期);

- 《主题标引的原理和方法》(《图书馆学刊》1996年第1期和第2期);

- 《汉语检索词词素轮排索引编制法探索》(《图书与情报》1992年第4期)。

这些文章都有一定的新颖性,大部分可构成情报语言学的基础理论。

4 术语学

我认为,术语学对情报语言学极其重要。术语是指称概念的规范化符号,情报检索语言的语词则是指称文献主题概念的规范化符号(标识),术语与情报检索语言的语词两者存在着本质的一致性。从这一角度看,术语学与情报语言学的研究对象可以说是同一事物,情报语言学所研究的是如何编制用于情报检索领域的“术语词典”——分类表、词表、代码表。

情报检索语言的编制可以说是术语整理(规范化、统一化、标准化)工作成果的诸多应用领域之一。术语整理工作的原则,要求术语具有:单义性、标准化术语的字面意义同它所表达的概念的一致性、系统性(术语应尽可能反映概念之间实际存在的、在系统化过程中确定下来的关系)、稳定性和普遍性、简洁性、语言的正确性、借用外来语的不可取性等,这些要求与情报检索语言的选词、规范化处理和显示概念关系等的要求都是一致的。所以,术语整理工作成果(术语标准、推荐术语集、术语词典等,特别是有术语定义和分类体系的术语文献)是情报检索语言选词的主要来源和进行规范化处理及建立概念间关系的重要依据。情报检索语言编制中的失误,有很多是忽视对术语整理工作成果的利用所造成的。

所以,术语学与情报语言学有着极为密切的关系,情报语言学研究对术语学的研究必定会获益匪浅。

5 自然语言检索

近10多年,随着自然语言检索的流行,图书情报界议论自然语言检索的文章很多,主流的论点是:①自然语言检索是发展方向,信息检索要走自然语言道路;②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;③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。

但是,如果仔细去看,可发现:发表这些乐观论点

者几乎都不是自然语言检索的专门研究者,而专门研究自然语言检索的学者中多数虽然也认同以上某些观点,但他们所发表的文章比较平和、审慎,不下如此断言。

自然语言在文献(或曰信息、情报、知识)检索中的应用大体可归纳为下列几个方面:①关键词索引及以关键词为检索标识的文献数据库(数据库中的关键词检索标识来自人工自由标引,或略加人工辅助的计算机抽词,或借助于词典的自动抽词);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(自动抽取主题概念词标引);⑥自动分类。

以上6个主要方面,只有关键词索引及数据库、全文检索、搜索引擎已进入实用。其实,这三个方面的实质都是关键词检索。所以可以说,自然语言检索目前仅在关键词检索的层次上已经实现(但还不是非常成熟,不是无可指摘)。至于自动甄别、自动标引和自动分类,严格地说都还没有走出实验室进入广泛应用。自然语言检索的研究已有半个多世纪的历史,进展如此缓慢,可见难度极大。

自然语言在情报检索中的应用,面临着以下两个难题:一是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。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。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。克服这个难题也是不能完全用自动化方法的。除此以外,对中文来说还有一个汉语分词的问题。汉语分词的研究已取得很大进展。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。

有不是一位作者说,自然语言检索是情报检索用语言发展的最高阶段。他们说:从自然语言,到人工语言,再回到自然语言,或者说,从不控制,到控制,再到不控制,这是“否定之否定”,是情报检索用语言的“发展规律”。

我认为,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。至今还没有找到在计算机环境下不加控制地利用自然语言的十分有效的方法。如果有,一定会被检索网站立即用高价收买而付诸应用。但至今在网络上没有发现那样的方法,说明那样的方法至少现在还不存在,将来也未必会出现。

自然语言到底会走向何方?我认为,自然语言的将来与情报检索语言的将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。总之,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。

其实,自然语言检索系统与情报检索语言检索系统并不是绝对对立的。既然两者各有优点而不可能互相取代,为什么不可以使两者结合或融合呢?自然语言或情报检索语言的将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。

情报检索语言的自然语言化、自然语言的情报检索语言化,这是两者发展的大趋势,走两者结合之路是大方向。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。

既然人工语言和自然语言都起着不可取代的作用,因而对两者的研究不可偏废。目前,亟待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(这是自然语言检索研究中的薄弱环节),把情报语言学的原理和方法引进自然语言检索的研究,并要重视利用情报检索语言已往所积累的成果(例如分类表和词表对概念和术语的整理成果)。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。这两方面的研究,应朝着并且必然会朝着从两者的初步结合到完全融合。

总之,我对仅仅借助于计算机技术的纯粹自然语言检索(或者说不利用情报语言学原理控制的自然语言检索)在短期内(例如10年、20年)能完满实现并不抱乐观的态度。我越来越觉得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。正是这样,我近年的研究

重点,转移到了自然语言在情报检索中的应用方面。

在自然语言在情报检索中的应用(包括网络资源检索工具)的研究方面,我曾写过不少文章,如:

- 《关于自然语言检索问题》(《图书馆论坛》2004年第6期);
- 《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年第5期);
- 《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年第4期);
- 《题名关键词与正文关键词检索性能的差别》(《中国索引》2004年第4期);
- 《自动抽词与自动分词》(《图书馆杂志》2002年第3期);
- 《文献题名自动抽词-分类标引系统》(《图书馆杂志》1998年第4期);
- 《自然语言与人工语言的对应转换——情报检索语言走向自动化之路》(《中国图书馆学报》1996年第1期);
- 《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》(《图书馆杂志》1999年第9期);
- 《走向自然语言与情报检索语言结合之路》(《图书馆理论与实践》2001年第2期);
- 《人工语言与自然语言、先控制与后控制的界限在计算机系统中可淡化或取消》(《图书馆杂志》1997年第5期);
- 《分类语言、主题语言、自然语言一体化检索系统》(《现代图书情报技术》2002年第1期);
- 《论后控制词表》(《图书情报工作》1994年第1期);
- 《充分利用入口词原理》(《图书馆论坛》1992年试刊号);
- 《论自由标引》(《图书馆学刊》1995年第5期);
- 《概念分面组配型自动分类系统》(《图书馆学刊》2002年第6期);
- 《全文数据库、全文检索与全文标引》(《图书馆理论与实践》2002年第6期);
- 《全文检索系统较好的模式》(《图书馆理论与实践》2002年第5期);
- 《全文检索系统的检索性能》(《江西图书馆学刊》2004年第3期);
- 《字面相似聚类法辅助构造词族表、分面类表和自动标引》(《图书馆论坛》2002年第5期);
- 《文献题名初步研究》(《江西图书馆学刊》2006年第3期);
- 《关于我国网络信息检索工具开发与改进的思考》(《2000年理论学术年刊》);

(下转第29页)

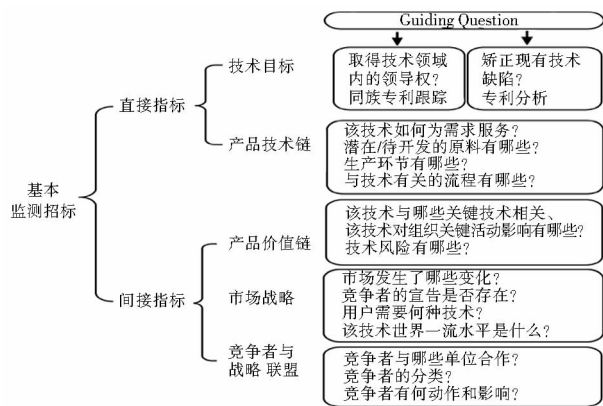


图3 技术预见基本监测指标

与开发的技术指标相关,包括技术的目标以及与技术相关的现有参数指标。通过直接信号,情报人员可以明确需求,对于分析效率的提高有较大帮助。间接信号包括产品价值链活动信号、市场战略信号以及竞争对手相关的信号。间接信号是多重的,分析时需明确问题、各个击破。在得到问题答案的基础上,将各种间接信号相关联,找出其中的技术预见信息。在实践过程中,由于专业的局限性,情报人员需要短时间充分了解该行业的专业信息,因此,跨专业的学习非常重要,在学习过程中要注意不断发现和修正偏差^[7]。

5.4 技术预见中信号分析的研究模式

虽然技术预见中信号分析的方法众多,但并没有统一的、放之四海皆准的研究模式。各类型模式均各有利弊。科学导向型研究可在科研任务为主、科研氛围浓厚、科技人员较多的研究机构内进行,其分析工具包括计量分析、出版物分析、常规的研发会议和网络;网络导向型研究,主要发生在以信息交流为主的机构

〔作者简介〕董悦,女,1983年生,硕士研究生,发表论文5篇,参编著作1部;赵小康,男,1982年生,博士研究生,发表论文8篇,参编著作1部;刘彦祖,女,1986年生,硕士研究生,发表论文1篇。

中,如市场部门、信息部门和管理部门,其工具主要是利用网络信息资源,通过观察、交流和搜索发现有价值的信息;市场导向型研究主要是通过持续的监测和系统的扫描来搜集、分析客户数据和竞争性信息,分析人员多来自销售部门和管理部门,分析方法多采用产品分析法、客户需求评估、市场分析法、SWOT分析、愿景分析法等;战略规划导向研究适合管理人员,运用的工具主要来自档案、政府出版物和市场资料。计算机技术也越来越多地介入了技术预见领域,数据挖掘技术、信息抽取技术、知识发现技术、聚类技术、图像技术都是进行信号搜集和分析不可缺少的技术工具。

参考文献:

- [1] B. Martin. Technology foresight 6: A review of recent overseas programmes. London: HMSO, 1995.
- [2] 联合国工业发展组织. [2009-01-08]. <http://www.unido.org/>.
- [3] 孙成权,曹霞,黄彦敏,等. 战略情报研究与技术预见. 上海:上海科学技术文献出版社,2008.
- [4] 陈劲. 技术创新管理方法. 北京:清华大学出版社,2006:32.
- [5] [2009-01-12]. <http://www.foresight.org/>.
- [6] 上海科学技术委员会. 技术预见在德国. [2009-01-12]. <http://www.stesm.gov.cn/>.
- [7] 费伊. 竞争者:以才智、谋略与绩效取胜. 朱舟,译. 北京:中国人民大学出版社,2005.
- [8] Day G S, Schoemaker P J H. Peripheral vision: Detecting the weak signals that will make or break your company. Boston: Harvard Business School, 2006: 5.
- [9] 中华人民共和国国家统计局. [2009-02-20]. <http://www.stats.gov.cn/tjbz/hyflbz/>.
- [10] 沈国朝,赵小康,董悦,等. 渣处理与利用技术报告. 南京大学信息管理系 CI 工作室, 2008.

(上接第9页)

- 《网络信息检索工具增强关键词检索功能的措施》(《图书馆杂志》2001年第1期);
- 《网络信息检索工具的分类体系》(《江苏图书馆学报》2002年第4期);
- 《因特网大众分类法若干问题的探讨》(《图书馆论坛》2005年第6期);
- 《因特网大众分类法的本质属性》(《图书馆杂志》2002年第11期);
- 《因特网大众分类法是独立创造而不是对传统分类法的改进和发展》(《江西图书馆学刊》2005年第1期);
- 《因特网大众分类法的标准化问题》(《中国索引》2004

年第1期);

- 《分类浏览型网络信息检索工具的主要缺陷》(《中国索引》2005年第1期);
- 《专业型检索工具与导航库在发掘网络信息资源中的重要作用》(《图书馆理论与实践》2002年第6期);
- 《网络信息检索用语言的发展趋势》(《图书馆杂志》2001年第3期)。

在以上文章中,带“•”号的文章,我认为是比较重要的。

至于我在研究索引学方面的心得和收获,因限于篇幅,不再作介绍了。

以上心得中的不正确、不完善之处,请不吝指正。