中文搜索引擎分类检索工具发展的大趋势

袁文莉

（河南职业技术师范学院图书馆　新乡　453003）

摘　要　中文搜索引擎分类检索工具有五大发展趋势：众多体系统一化、分类主题融合化、体系组配一体化、同位类排列规范化、自然语言受控化，就五大发展趋势分别进行了论述。

关键词　中文搜索引擎　分类检索　检索工具

自从 1996 年开发出第一批中文搜索引擎之后，到目前为止已发展到 282 个，这些中文搜索引擎一般都提供分类检索和关键词检索两种方式，而分类检索都是按照事先组织好的分类检索工具进行的。虽然目前中文搜索引擎分类检索工具各俱特色，良莠不齐，都面临着改进和完善，但其发展趋势是一致的。

1　众多体系统一化

出于个性化特别是知识产权的考虑，往往是有多少种中文搜索引擎，就有多少种分类体系。即使是大类的类目数量相等（如中国白垩纪和秀佳搜索引擎都是 24 个大类，搜狐、网易和新浪网都是 18 个大类），但其类名及其排序是绝对不会雷同的，就是同一种中文搜索引擎的分类体系，由于时间的推移和处所的变更，也有面目全非的时候。亦凡搜索，原有大类 10 个，后又增补 2 个，现共 12 个大类；TOM 搜索引擎，原有 10 个大类，后增补 9 个，删去 3 个，改换 5 个，现有 16 个大类；网天，原有 14 个大类，后增补 6 个，删去 3 个，改换 3 个，合并 1 个，现有 16 个大类。特别是木子网（中文）首页分类体系的类名与点击后的类名迥然不同，如把"娱乐与明星"改为"娱乐与影视"，把"艺术与表演"改为"艺术与建筑"，把"社会与历史"改为"社会与人物"，把"科学与技术"改为"科学与研究"，把"旅游与风景"改为"旅游与交通"等。

我们知道，用户在查询信息时，往往需要使用多个搜索引擎，而每一种搜索引擎的分类体系都不一样，就是同一种搜索引擎的分类体系也变幻莫测，这就给用户熟悉和运用中文搜索引擎的分类体系，造成很大的困难和负担，因此，目前的中文搜索引擎分类体系已经到了非改不可的时候了。那么，怎么改？如何改？换句话说，中文搜索引擎的分类体系究竟向着什么方向发展？其发展的大趋势是什么？历史上有着极其相似的现象：1910 年《杜威十进分类法》传入我国之后，当时的学术界出现了"仿杜"、"补杜"、"改杜"之风。据统计，当时以《杜威法》为蓝本编就的中国文献分类法不下 30 种，解放后又编制了 28 部分类法。从"军阀割剧"到一统天下，虽然经历了一个漫长的历史时期，但其历史发展趋势是很明显的，那就是：从群雄争霸，到"三国鼎立"（《中图法》、《科图法》、《人大法》，全国常用的三大分类法），再到《中图法》的大统一。目前中文搜索引擎的分类体系虽然还处于各自为战的阶段，犹如解放前后 20 年的传统分类法，但其发展趋势也是很明显的，那就是九九归一，编制全国通用的网络信息分类法，达到众多体系统一化。因为这是提高中文搜索引擎分类体系质量的唯一途径，是中文搜索引擎分类体系走上标准化的唯一选择，是实现资源共享和信息化的必由之路。

2　分类主题融合化

20 世纪 80 年代初期，当时学者曾对分类目录与主题目录孰优孰劣展开过激烈的争论，争论的结果导致了分类主题一体化的提出，并公认为这是传统分类法发展的趋势之一。进入 90 年代之后，在众多同仁的努力下，一部由《中图法》和《汉表》（《汉语主题词表》简称）合二为一的产物——《中图分类主题词表》便应运而生。它象征着分类主题一体化的研究终于结成了硕果，付诸了实施。但我们认为，这种分类主题一体化只是初级的、表象的、形式的，而不是内在的、有机的、融合化的。

目前的中文搜索引擎主要向用户提供两种检索途径：一是分类检索，二是关键词检索。分类检索途径的优点在于便于系统地查找某一学科、专业或主题范围之内的知识信息，便于扩检和缩检，但目前大多数分类检索工具都只是靠人工筛选和人工分类方式建立供其使用的数据库，建库成本高，时效性差，缺乏及时维护；关键词检索途径的优点在于：由于关键词属于自然语言，可在网上自动采集、自动抽取，建库容易，成本低，时效性好，维护及时，便于用户直接检索到最新的知识信息；但由于关键词系统缺少规范化处理，用户往往会检索出大量无用信息，使人望而却步。

由此可见，分类检索途径与关键词检索途径各有优缺，那么，如何吸纳各自的优点而避免各自的缺点呢？这就要走分类主题一体化的道路了。当然，我们这里所说的一体化并非历史上的简单重复，而是更高阶段的发展，不是二者机械地一一对等，而是二者有机地融合，即分类检索系统和关键词检索系统的融汇贯通，学科聚类和主题聚类的融汇贯通，系统序列和字顺序列的融汇贯通。犹如目前中文搜索引擎分类体系中所使用的自然语言与人工语言浑然一体，既要保留分类检索的系统性，又要体现主题检索的直观性。虽然达到理想的境界有一定的难度，但它确是中文搜索引擎分类检索工具发展的方向和趋势。

3　体系组配一体化

体系组配一体化正像分类主题一体化一样，并非是新鲜的课题，但它却是难以解决的老问题。《中图法》已含有 5 种组配成份，即主类号直接组配、复分表、仿分、多重列类法、并列法（《资料法》规定可使用联合符合"＋"），但《中图法》组配的

比重并不大，而且由于受到分类实践的抵制，已有的组配条件，也未得到利用。实践证明，《中图法》要实现体系组配一体化的道路还很长，也可以说仅仅是一种难以实现的发展趋势，于是人们便把希望寄托在中文搜索引擎分类检索工具上面。

目前的中文搜索引擎分类检索工具多数是采用等级分类系统，并且在主页上列举全部一组类目和部分二级或少数三级热门类目，然后是第二层再列举有关的二级类及其所属的三级类目，以下类推；少数是采用分面组配分类体系，将网页信息或网站内容，按照不同的标准分析为面，面内由若干个代表特征概念的类目组成，各个分面的类目进行组配，于是便形成了更加专指的组配类目。前者如搜狐、网易、新浪网等，后者如中华网目等。前者的主要问题是：不能反映多维的知识空间；不能反映事物多向成类的性质；不能满足检索专指度高的信息需求；不太适应信息变化和科技发展的需要。虽然目前采用等级分类系统的中文搜索引擎大都采取了许多措施，如大量设置交替类目和镜像类目，经常对其分类体系进行增、删、改，甚至建立第二分类体系等，但都杯水车薪，不能从根本上解决上述问题，反而增加了用户的认知负担，降低查准率，甚至使用户感到紊乱不堪，无所适从。后者的主要问题是：基本分面的划分很难作到全面、科学，而且不同的学科都有不同的分面；由于组配检索会降低检索速度，从而增加运营成本；对标引技术要求高，一般网络信息标引员很难适应。

鉴于上述，中文搜索引擎分类检索工具的发展趋势是体系组配一体化，即先体系后组配。二级以上类目采用等级分类系统，二级或个别的三级类目以下采用分面组配系统，实现先组式检索语言和后组式检索语言的有机结合，界面上再配以菜单和视窗，实现可视化检索。当然，关于体系组配一体化还有许多理论和技术问题需要我们研究解决，但中文搜索引擎分类检索工具的体系组配一体化是大势所趋。

4　同位类排列规范化

同位类排列是衡量中文搜索引擎分类检索工具质量高低的一个重要方面，也是影响用户检索效率的一个重要因素。目前几乎所有的中文搜索引擎的分类体系都存在着同位类排列混乱的通病：新浪网、网易、木子网(中文)、263在线等，都把"文学"与"艺术"、"电脑网络"与"科学技术"、"经济"与"文化"等原本密切的类目分割开来；搜狐对"文学"大类之下27个同位类的排列是19个类是按汉语拼音字母顺序排列的，其余8个类是随意排列；新浪网对"文学"大类之下29个同位类的排列是按访问量排列等。

由此可见目前中文搜索引擎分类检索工具对同位类的排列大致有三种情况：一是按照类目的汉语拼音字母顺序排；二是按照类目的被访问量排；三是随意排放。第一种方法割裂了类目之间的逻辑关系，使得彼此关系密切的类目因割裂而不能触类旁通；第二种方法不但不能体现类目之间的相关性，而且缺乏稳定性，因为访问量是个变数；第三种方法最省事，也最混乱。

由此看来，同位类排列的规范化是中文搜索引擎分类检索工具的当务之急，也是大势所趋。在这里，可以借用传统分类法中按逻辑次序排列同位类的作法，也可借用主题法中按字母顺序排列同位类的作法。其规范化的要求是：首先按照类目之间的逻辑次序排，只有当同位类无法使用逻辑次序排列时才可采用字顺排列法，如国家、省(市)、人物、机构、民族、公司、网站等。

5　自然语言受控化

中文搜索引擎分类检索工具的术语平面是通过类目的命名来实现的，用户也是通过类名来识别和选择检索路径的，因此，对类目的命名应能正确地反映其内涵和外延。

目前中文搜索引擎分类检索工具中类目的名称可划分为三种情况：一是与传统分类法中类目名称相一致，如社会科学、自然科学、文学、艺术、教育、政治、法律、军事、经济、科学、文化、体育、旅游、交通、医药等；二是虽与传统分类法中类目名称不一致，但已约定俗成并被广大用户所理解和接受的，如媒体、影视、环保、婚恋、高校、求职、黑客、文革、电脑、IT业等；三是既与传统分类法中类目名称不一致，又有可能增加用户认知负担的，如名捕的"说天说地"、"美眉写真"，天网搜索的"嗜好"、"残障"、"讨论话题"，搜鼠的"东方热线"、"网络传情"，搜豹的"学者"、"杂类"，百度的"一见钟情"，亦凡搜索的"科研精英"，广州视窗的"CRM"、"VOD"、"MP3"等。

我们知道，传统分类法对类目的命名使用的是人工语言，而中文搜索引擎分类检索工具对类目的命名使用多是自然语言，也有人工语言。由于自然语言使用便捷，不用翻查有关的分类主题词表，而且不受词表的限制，随时增补新词，且专指性强，因此，中文搜索引擎分类标引人员和分类检索人员都喜欢使用自然语言。正如上述，自然语言也存在着不规范、概念的含义显示不出来、亦有大量的同义现象、多义和模糊现象，因而检全率和检准率都很低。因此说，自然语言受控化，将是中文搜索引擎分类检索工具发展的大趋势之一。

为使自然语言达到受控化，为自然语言编制后控制词表是有效措施之一。在该系统中，标引用语和检索用语都是自然语言，而控制用语都是人工语言，二者不但能达到和谐统一，而且还能达到同步增长。

参考文献

1　http://ww.baieji.com
2　http://www.shugachina.com/search
3　http://www.sohu.com
4　http://www.yeah.net
5　http://www.sina.com.cn
6　http://www.gotofind.com/opendir
7　http://www.search.tom.com
8　http://www.net-sky.com
9　http://www.go.muzi.net
10　刘延章.文献信息分类学.北京：中国科学技术出版社，1996
11　刘延章.我看中图文献分类法这条河.图书馆论坛，2000；(1)
12　俞君立.中国文献分类法百年发展与展望.武汉：武汉大学出版社，2002
13　中国图书馆分类法编辑委员会.中国图书馆分类法(第四版).北京：北京图书馆出版社，1999
14　http://www.search.china.com.
15　http://www.search.263.net
16　http://www.mingbu.com
17　http://www.e.pru.edu.cn
18　http://www.sosoo.cnnb.net
19　http://www.sobao.com
20　http://www.ix.baidu.com
21　http://www.search.gznet.com

(责编：王京阳)