分类搜索引擎类目体系研究

[作者] 马张华
[单位] 北京大学信息管理系
[摘要] 论述分类搜索引擎类目结构的编制依据、大类结构、类目体系的特点等；对存在的问题展开讨论。
[关键词] 分类搜索引擎，大类体系，分类法编制，类目体系

目前网上基本上有两类分类体系，一类为传统分类法，主要用于学术性信息资源的组织；另一类为指南型分类体系，主要用于通用性网络资源的组织，是分类搜索引擎的编制依据。其中，指南型分类体系根据网络资源组织的需要，构建了不同于传统文献分类法的类目体系，探索了适合网络环境的技术方法，成为一种不同于传统文献分类法的分类法类型，因此，探讨这类分类体系的特点和规律，已成为探索网络信息资源组织无法回避的问题。

1 分类搜索引擎类目结构的编制依据

分类工具通常是根据分类对象的特点和用户需求，结合一定的技术环境建立的。分类搜索引擎类目结构与传统分类法的根本不同在于，它是网络环境的产物，是直接根据网络环境的特点和要求编制的。与传统分类法相比，其编制依据的不同集中反映在以下几个方面：

●分类对象。与传统分类对象相比，网络资源的特点是：数量大、种类多、动态性强，内容分布特点不同。网络资源中的多种资源类型，如 BBS、聊天室、新闻组、多媒体资源等，是传统文献资源所没有的，而占传统文献集合主体（50%～90%）的单行著作，不足电子资源数量的 30% [1]；同时，在内容分布上，新兴科学技术、商业、娱乐的资源数量较多，传统知识门类的资源相对较少。这就要求有新的、适合处理对象的分类架构。

●用户需求。网络的使用对象涉及所有的终端用户，比文献分类法的用户更广，这就要求类目体系简明、易于使用，使用户不经过预先了解，就可以通过类表进行查找。与传统分类法相比，类目体系应具有较强的通用性、直接性，并且能满足从各种不同的角度查找的要求。

●技术环境。传统分类体系是按照文献组织和手工检索工具编制的需要确定的，基本上为线性形式；网络分类体系则是以电子文本为处理对象，根据网络检索的需要编制的。链接技术的使用，使得它能够按照主题之间的关系和用户的需要，灵活地、多维地进行揭示，在体系构建、类目设置等方面，发展不同于传统分类法的技术特色。

上述因素决定了分类搜索引擎在分类架构和技术特点等方面都不可能等同于传统的分类体系。

2 大类结构

大类体系是分类法的大纲，反映和规定了分类体系的整体展示方式，是分类法能否合理揭示资源、有效服务于用户的重要因素。表 1 为几个国内外比较典型的分类搜索引擎的大类体系。可在一定程度上反映综合性分类搜索引擎大类体系的设置情况。从表中大类体系的情况看，国内外分类搜索引擎的大类体系有以下特点：

●从类目体系设置的角度看，基本上放弃了文献分类法以学科为中心建立类目体系的传统，而采用了以主题为中心或主题与学科相结合的两种设类方式。其中，学科与主题相结合的方式，可以使其在具有直接性的同时增加包容性，使用更加普遍。从表中可以看出，除 Magellan 的大类体系基本上是以主题、对象为中心设置外，其他的搜索引擎均在不同程度上采用了主题对象与学科相结合的设类方式。从目前掌握的资料看，我国分类搜索引擎的大类体系，大多采用这种类目设置方式。

表 1 国内外典型分类搜索引擎大类体系示例（略）

●从大类体系的整体功能看，突出了教育、娱乐、旅行、生活等与日常生活密切相关、普

通用户感兴趣的类目，弱化了科学技术、学术性类目的设置，基本上是一个通用性的大类结构。

●从文献保证的角度看，基本上反映了网络资源的内容分布情况。资源较多的类目，如“ 计算机与互联网 ”、“ 新闻媒体 ”等，在传统分类法中都只是学科类下的子类，分类搜索引擎大多将其提升为一级类目；相反，对传统分类体系中详尽展开的领域，如自然科学、应用技术门类等，分类搜索引擎只设置了概括性类目。

●从类目检索的入口看，多数大类体系提供了多维检索入口。除信息资源的主题、学科外，一般还同时从地区、资源类型、机构等角度设类。如 Yahoo!、Hotbot、搜狐等都提供了从地区和资源类型角度检索的入口，搜狐还将个人主页、公司企业等直接设为一级类目，便于用户从不同角度检索。

不少论者呼吁建立统一的网络分类法。实际上，在网络条件下，资源的生成和处理是动态的，分类体系也应根据资源和用户需求的变动而随时调整。人为地追求统一是没有必要的。统一的网络分类法只有在特定的系统或联合体内才是可能的。就基本大类而言，主要决定于资源情况和对功能需要的认识。个性化的服务需要个性化的资源选择和组织形式。国内的某些分类搜索引擎在大类设置上出现的趋同现象，实际上只是它们在资源收藏范围和服务方向上趋同的反映。

值得一提的是，OCLC 网站在使用 DDC 组织学术资源的同时，最近在其《杜威分类法》下，推出了一个名为 DDC 浏览器的分类搜索引擎 [2]，共设置 15 个大类，依次为：艺术、文娱、商业、计算机、教育、政府与法律、健康与家庭、人文、新闻、娱乐、休闲、综合参考、宗教、科学与技术、社会科学、旅行。这一设置的思路，基本上与一般分类搜索引擎的大类结构相类似。这一事实说明，即使在老牌文献分类法 DDC 的编者看来，通用网络资源的分类也需要一种与传统文献分类法不同的分类架构。

3 类目体系的展开

与文献分类法相同，分类搜索引擎的类目体系，基本上也是通过层层划分，按照从总到分的方式逐级展开的；但由于资源特点和用户需求的不同，特别是超文本技术提供的多维检索能力，使其在类目体系的展开以及类间关系揭示等方面，形成了不同于传统分类法的特色，具体表现在：

3.1 重视按事物对象设类

这是大类体系重视按主题对象设类的继续。为了使分类体系有较强的直接性，网络分类体系往往直接按照特定的事物对象设类。以搜狐、新浪等分类体系为例，其“ 计算机与网络 ”

中的“软件”、“互联网”、“电脑艺术”、“计算机安全”、“集成系统”;“经济”类中的“公司机构”、“房地产”、“开发区”、“就业信息”等均为较具体的对象类目。这种设置方式不仅减少了常用类目的显示等级，而且可以增强分类体系的直接性与易用性。与此同时，网络分类体系还重视按用户关注的问题设置类目，如“教育”类中的“考试与招生”、“出国留学”；“医疗与卫生”中的“心理健康”、“紧急救护”,“求助 BBS”、“性教育”以及各类中的热点问题等，都是按用户关注的问题设置的，从而形成了根据用户的需求，按问题组织相应资源的类目设置架构。使得整个分类体系成为一个与传统分类法设类方式不同，更加重视直接性、实用性，侧重于从事物、问题为中心设类的系统。

3.2 多元划分

所谓多元划分，是指类目展开时，同时采用多种划分标准。传统分类法类目展开时，通常遵循逻辑分类的原则进行，一次只采用一个标准，只有在必要时才采用两个或两个以上的标准。与传统分类法相比，网络分类法的类目展开，受划分标准的束缚较少，往往在一些类下同时采用两个或多个标准。从国内综合性搜索引擎的编制结构看，各大类下进一步划分的标准通是多元的。以新浪社会文化类为例（见表 2），该类同时采用了主题对象、学科、资源类型等多重标准设置类目；即使是按对象设置的类目，也并非属于一个标准。这就使得分类搜索引擎在类目展开时，各层次下的类目类型多、数量大、范围广，同时也减少了类目展开的层次，增加了类表的直接性，为从不同角度展开类目体系提供了条件。

从目前实际使用的情况看，搜索引擎大类下的编制结构包括以下 3 种情况：①当大类为主题对象类时，以主题对象类为中心设置相关学科类，同时收入地区及各种资源类型类；②当大为学科类时，以学科类为中心设置相关主题对象类，同时收入地区及各种资源类型类；③当大类为地区或资源类型类时，地区下按国内外政治区划区分；各种资源类型进一步按类型的形式区分，再按主题内容区分，从而构成了与传统分类法特点迥异的划分和类目设置方式。表 2 新浪“社会文化”类下的多元划分（略）

3.3 多维展开

这里的多维展开有两个含义，其一，从类目划分的个体而言，是指在多元划分的基础上，分别从不同类目的特点出发加以展开；其二，从类表展开的整体而言，则是指同时使用不同的引用次序列类，多维度地展开类目体系。例如，在主题类下按“主题—地区”、“主题—文献类型”的引用次序展开类目体系的同时，在新闻媒体、BBS、个人主页、机构团体以及地区下，按“形式—主题”,“地区—主题”的引用次序展开分类体系。通过超文本链接，在相应类下对相关类的重复反映，可使整个类目体系成为一个分别从主题内容、地区、形式的角度展开的网状结构（见图 1）。这一展开方式是文献分类法在传统环境下不可能做到的。传统分类法为了适应不同用户的需要，虽然有时也采用多个分类标准展开，但由于受线性结构的限制，在实际使用时，一般采用选择的方法加以控制，并将其限制在单维展开的范围以内。在网络分类体系中，这种限制为超文本结构所突破。多维展开的结果，是可以对某一领域的类目同时以不同的引用次序建立不同功能的分类系统，供用户根据需要从不同角度着手进行检索。

多维展示最典型的例子，是国内搜索引擎对文学作品类的设置。不少搜索引擎对文学作品同时采用了引用次序组织类目：①体裁—国别—时代—题材；②国别—体裁—时代—题材；③时代—国别—体裁—题材；④题材—体裁—国别—时代。这样，每个引用次序便组成

了不同功能的资源组织系统，用户可选择最适用的系统进行检索；同时，这一系统还增加了检索入口，便于用户从不同角度出发查找文学作品。

图 1 超文本在多元展开中应用示例 （略）

3.4 横向关系揭示

包括多属类目和相关类目。与文献分类法相比，分类搜索引擎在横向关系揭示上的不同主要表现在显示形式和揭示范围两个方面。

传统分类法一般通过交替类目和类目参照揭示横向关系，以作为类目纵向关系的一种补充。网络分类体系则使用链接方式，通过在相应类下重复反映，使其成为类目关系的有机组成部分。目前搜索引擎对这类关系的揭示，最常见的是将横向关系在相关类下重复反映，同时采用@加以表示；国外有的分类法还将横向关系区分为两种类型，以@表示重复反映，以参照类揭示相关类目；也有的分类法对多重重复不使用符号予以显示。不管采用何种方式，都有效加强了类目之间多重关系的揭示。传统分类法的线性结构要求对横向关系有严格的限制。比较而言，网络分类法对横向关系的揭示更为充分。目前网络分类法对多重关系的揭示包括：①多属性主题；②交叉学科；③边缘学科；④总论与专论；⑤资源形式与主题，如期刊、BBS、聊天屋、个人主页等及相应主题；⑥地区与主题；⑦机构、人物与相应知识门类等。参照关系的范围则更加灵活，有的系统还将相应的信息资源与其相联结。这类横向关系的揭示，有效反映了知识之间的联系，使类目的范围更加完整，相关类的揭示也更为充分，有利于资源的选择和查找。

3.5 类目设置与显示的新形式

结合使用需要和新的技术环境，网络分类法对类目设置进行了改进，形成了一些新的设类形式，比较典型的有：

●通过重复反映，提前设置热点类。即对一些热点类目，在其相应位置上设类的同时，还根

据使用需要，在作为上一级的类目中突出反映，以便于用户使用。如将“体育与运动”下的“足球”、“篮球”、“乒乓球”等类目在“球类运动”下设类的同时，打破原有的逻辑等级层次，以“球类运动”并列类的方式加以突出反映，方便用户对这些类目的使用。

●动态设类。指根据使用需要组织和显示相关资源，使分类体系能及时反映用户需求和资源的变化。如通过设置镜像类目，以链接的方式，在春节、圣诞节临近时将有关该节日的类目提前设置；在有关事件进行的过程中，对相应类目，如“计算机 2000 年问题”，“千禧年”等类目突出反映；事后，则可以取消该镜像类目，使分类法具有传统分类法所没有的动态性。●多角度设类。即根据用户需要，从不同角度出发设置类目，如在历史类下，同时从各代史、历史事件、人物、专门史等设置类目；在哲学类下，同时从哲学史、流派、著名哲学家著作等多个角度，同时序列哲学家的著作和研究。在可能的情况下，对多属类目还可采用链接方式，并将其在相关类目下重复显示，使用户在不增加工作量的同时，可以从不同角度检出该类著作。

●类目显示方式直观。搜索引擎放弃在检索界面使用标记的做法，直接显示类名，同时通过在检索界面上部揭示的各级上位类，限定该类的等级和含义，这一形式还可供点击，直接回访相应的上位类，整个界面简练、直观，易于理解，便于使用。搜索引擎的类目索引也采用了一种不同于传统分类法的索引形式，即直接以对应的类目及其上位类表示，使用户可

以通过类目的直观显示了解其含义。

上述方法发展了分类法的类目设置方式，使网络分类体系具有更大的动态性、适用性，在类目显示特点上呈现出与传统分类法不同的面貌。

4 问题讨论

分类搜索引擎的类目体系目前还存在许多问题。其中，用户反映比较突出的问题是，类目设置的随意性比较大，类目关系缺乏规律性等。笔者认为，要解决上述之不足，主要应注意以下问题：

4.1 应遵守基本的逻辑规则

实用分类法的类目设置，一般都需要根据情况，在遵守逻辑规则的同时对其进行必要调整。传统文献分类法在类目设置中，对逻辑划分规则的执行也并不严格，存在一度划分同时采用两个或多个划分标准的情况。分类搜索引擎要求以多维方式展开类目体系，并使用多种重复反映方法，对逻辑规则必然需要作相应的调整，但应明确调整的幅度以及必须遵守的基本准则。在目前的搜索引擎中，对逻辑规则的调整缺乏必要的规则系统，有时甚至出现一些违背基本逻辑方法的做法。解决的办法应在类目展开过程中，遵守分类的基本逻辑要求，至少应做到：①应保持从总到分的展开序列，上位类应能涵盖下位类，不能在类目展开中出现上下位类颠倒；②应研究多元划分时划分标准的类型，研究此情况下类目之间的关系和规律，逐步形成常规使用的模式；③一类下包括的类目范围不能过广，不能把不相从属的类目收入其下；④类名应该正确反映类目的内涵和外延，在生动、鲜明的同时准确反映类名的含义。

同时，应对类目动态反映、提前设置等形式的效果和影响进行分析。可以根据使用需要，利用超文本形式动态设类，对类目体系进行必要调整，加强类表的适用性，但这样做会影响分类体系的系统性。应探讨如何在进行灵活揭示的同时，保持一定的度，使其既能够增加类目体系的实用性，又能将其对类目体系层次性的影响限制在一定的范围内，使类目体系在整体上能反映类目展开的规律性。

4.2 应解决好类目的排列问题

同位类的排列直接影响到同位类之间关系的揭示，是分类体系建立的一个基本内容。在

传统文献分类法中，同位类排列通常是按照类目之间的关系进行的，但分类搜索引擎多数没有采用这一方法。

在目前的分类搜索引擎中，英文分类搜索引擎一般都采用主题字顺的方式组织类目体系；这一形式符合外国用户按字顺检索的习惯，易于排列同位类，是国外分类搜索引擎普遍采用的方法。缺点是，这种排列方式在“子类较多的情况下，无法集中相关文献”[3]；国内分类搜索引擎对同位类排列问题一般没有明确说明，从实际使用情况看，有以下 3 种：①采用字顺排列。个别系统部分采用了这一方法。但由于汉字排检不如西文便捷，效果并不理想；②在排序中参考检索频率的因素。即将检索频率高的类目排在前列。但这一方法缺乏稳定性，同时也不能揭示类目的相关性，因此，在实际使用中，一般并没有将这一方法贯彻到底，使得这些系统的同位类基本上成为一种任意的排列，不仅不能给用户带来使用的便利，还会在多维揭示的情况下，加剧无序和混乱的感觉；③对同位类进行系统排列。目前已有个

别系统采用了这一方式。如下面为蓝帆搜索“教育”类系统排列的例子［4］：

综合网站会议与活动网上教育
政策法规　图书馆　高等教育
新闻媒体教育设施　中等教育
BBS 与聊天屋教育理论　初等教育
个人主页　　考试与招生幼儿教育
机构与团体 就业与招聘职业教育
　人物　留学与出国成人教育
　社会教育
　家庭教育
　特殊教育

上例按照从总到分，从资源类型到方面、对象的次序排列教育类的二级类，可以在一定程度上揭示类目之间的联系。

从分类法的发展历史看，西方分类法中采用过系统排列和字顺排列两种方法。字顺排列主要在 18 世纪分类法编制中流行，后来逐步为系统排列所取代。我国分类法传统上采用的是系统排列方式。系统排列的作用是，①可以揭示类目之间的联系，方便相关类目的查找；②利于结合类目的排列，明确类目的含义；③增加类目排列的一致性和可预见性。从我国分类法的使用传统和实际使用效果看，笔者认为对同位类按照类目之间的关系排列比较好。

将系统排列用于网络分类法应该是可行的。至少有 3 种排列法：①系统排列，即将所有同位类目按相互关系排；②按类型排列，即只将同类信息资源加以集中；③只集中相关类目。上述 3 种排列法无论哪一种都比任意排列好。长期以来，传统文献分类法对同位类的排列总结出了一系列的排列方法，只要将其与网络分类的实践结合，就可以合理地解决分类搜索引擎同位类目的排列问题。国内一些搜索引擎的实践说明，同位类按照类目之间的关系排列是可行的。

4.3 横向类目处理问题

超文本链接技术的使用，可以有效揭示类目之间的多维关系，改进分类体系对相关类的揭示。但在目前的应用中，仍然存在不少问题。其一是，对同一性质类目的处理缺乏一致性。即在一些分类搜索引擎中存在着对同类类目处理上缺乏整体性、一致性的问题。如原 263 搜索引擎中，有关人物的类目是在不同类下分别设置的，没有进行统一的协调，致使不同类下有关人物的类目相互交叉，缺乏类目设置的一致性。这类情况即使是链接技术使用得比较好的搜索引擎如搜狐，也同样存在。如搜狐医药类中“疾病与治疗”外科类，与“临床医学”下外科类的设置，就缺乏必要的协调，一些可以平行设置的类目，没有进行统一设类，使得两者的类目设置都不够完整，影响实际使用效果。

横向关系处理的另一个问题是，目前部分分类体系类下的范围过宽。超文本链接技术的使用，为多重从属和相关联系的揭示提供了便利，但也导致一些类下的类目超出其外延，如一些系统将图书馆作为家庭教育的下位类，将各学科门类作为教育的下位类等。这类情况会造成系统导航的无方向性，使分类检索效能下降。要解决这类问题，一是下位类的收录应有一定的界限；二是应对从属类与相关类进行区分，使得既能揭示类目之间的联系，又能准确、适度地进行揭示。传统检索语言，包括分类法和主题法对主题之间关系的控制，进行了许多研究，对类目之间关系的类型、表达形式、处理机制、适用范围等进行了许多探讨。网络分类中横向关系的建立，应注意在汲取文献分类法、叙词法相应研究成果的基础上，逐步建立起比较完整的理论技术规范体系。

参考文献

1Marcia Lei Zeng. Search for New Ordering Systems for Resources: a Study of the Approaches to Organizing Information in the World Wide Web Virtual Libraries from 1995 to 1997.Knowledge Organization for Information Retrieval: Proceedings of the Sixth International Study on Conference on Classification Research. 16-18 June 1997,London, 28～31 The Hague;FID.

2http:www.oclc.org/vizine/Dewey-Browse/ddc-Top.htm

3Martin.van der Walt. The structure of classification schemes used in Internet search engines. Advances in Knowledge Organization, 1998(6):379～387

4http://www.linefan.com.cn/

5http://www.sina.com.cn/

6http://www.sohu.com.cn/