中文搜索引擎的分类体系研究

吴红勤

（武汉市青山区图书馆，武汉 430070）

[摘 要]探讨了现有中文搜索引擎分类体系的不足，提出构建统一的中文搜索引擎分类体系的必要性和构建方法。

[关键词]信息资源 搜索引擎 分类
[分类号]G254.11

1 中文搜索引擎分类体系的不足

现有的中文搜索引擎以“事物”为中心设置类目，搜索范围很宽，同时将热点类目单独列出，极大地方便了用户检索。但也存在不少弊端，值得我们重视。

1.1 类目体系不一，类目设置不科学，缺乏规律性

无论是一级类目还是二级类目，甚至在同一个搜索引擎中，类目体系也不尽统一。在一级类目中，类目个数不等，二级类目差距更大，类目设置很随意，同时，不少系统同位类的展开是多元的，往往同时采用多个标准，每一种标准使用时又并不完整设类，有时还同时列出不同等级的类目，使得同位类的设置缺乏必要的规律性。如Yahoo大类的设置，就同时使用了主题对象、学科、地区、文献类型等多个标准，各下级类的设置中也贯穿着相同的设类方式。

1.2 类目之间逻辑性差，类目归属存在不合理现象

中文搜索引擎分类体系的类目之间缺乏逻辑性，同级关系不明，上下关系不清，隶属关系混乱的现象比比皆是。如Yahoo将“新闻媒体”作为一级类目，同时又将它列入卫生与健康二级类目下，“电脑与因特网”与“科学”并列等等。在类目归属时，有两种情况不合理：一类是类表在确定类目的下位类时，相关类收入范围过宽，导致收入一些超出母类外延的类目；另一类情况是未能按照知识之间的关系合理确定类目的归属，如在已经设置“政治军事”的情况下，将法律归入“社会文化”等。

1.3 横向关系揭示上的不合理

传统文献分类法对横向关系的揭示包括设置交替类目、选择类目、类目参照等多种方法。搜索引擎分类法通过链接的方式，将具有多重联系的类目，如交替类、交叉关系类目、总论与专论等在各相关类目下重复反映，以增加类表的多维性。但在处理过程中缺乏一致性及对范围的适度控制。在多重联系类目的处理上，大多重复反映不够，类目缺少提示，交叉参考类目少。用户不能直接找到所需的类名，必须一级一级地向下翻阅，有时为了查找一个类名，需要读取十多个页面，费时费力。此外，单一的重复反映并不能简单代替相关关系的揭示。过度将具有相关关系的类目集中于一个类下，虽然可以改进系统的检全率，同时也会影响其检准率。

1.4 同位类排列没有揭示类间关系

按照类名字顺或检索频率排序虽然可以方便同位类的列举，但这类列类方式不能反映并列类目之间的联系，缺乏对知识门类系统显示的能力和揭示类目相关性的作用。特别是在同位类采用多种划分标准的情况下，更容易造成类目关系的混乱。

1.5 类名不规范、不统一

在中文搜索引擎分类体系中普遍存在类名不统一的现象。比如，对医药卫生类的表达，sohu、sina、yahoo、yoyo分别是“卫生与健康”、“医疗健康”、“健康与医药”和“医疗保健”。sohu有一类是“计算机与互联网”，国家有关部门已下发有关文件，对“Internet”统称为“因特网”，显然sohu在此不够规范。同时，类名有时也不能确切概括类目的内涵，有时只有象征意义。如Yahoo基本大类“政府”下，收入政治、军事、法律、政治学、民族、国际组织等，远远超出了类名的范围。此外，由于网络类目的下位类范围较宽，上位类有时很难有效限定其含义，加之一般不使用注释揭示类目的内涵，容易使一些类目的含义和范围难以确定，影响使用效果。

1.6 类目无注释

绝大多数中文搜索引擎，无论是一级类名还是二级类目都没有注释，这样，用户很难准确判断其外延，如Yahoo的“另类治疗”等。这些类目如果没有注释，用户根本不知道包含了什么内容。

1.7 动态性过强

由于搜索引擎是与网络信息建立的链接，而网上信息的更新又相当迅速，因此，不同时期搜索引擎的分类体系可能还有所变化，所体现的重点也可能不一样。这样，不同的搜索引擎，甚至是相同的，都很难让其用户熟悉它的分类体系，直接影响用户使用。

2 规范搜索引擎分类体系的理论分析

2.1 依据

2.1.1 统一中文搜索引擎分类体系的必要性。现有中文搜索引擎的分类体系，其大类设置与划分、类名的表述与外延、类

目的排列等都各不相同,这对于知识组织与信息交流都是不利的,尤其不便于用户使用。用户上网查询信息一般不只用一个搜索引擎,因此要熟悉多种不同的分类体系,而不同的搜索引擎,即使类名相同,外延也不尽相同,因而造成用户理解和运用的困难。因特网上信息资源最大的特点之一就是共享性。知识组织方法的相对统一,是促进网上信息易检性和共享的重要因素。众多搜索引擎如果使用统一的分类系统组织信息,用户使用起来就方便简洁多了。

2.1.2 统一中文搜索引擎分类体系的可行性。尽管从哲学和科学上对人类知识的体系还有不同的认识,但对知识领域的划分基本是统一的,网络信息也有着共同的体系。很多专家建议集中力量开发为数不多的大型综合性中文搜索引擎,并积极开发研制各类专业搜索引擎,这也为统一中文搜索引擎的分类体系带来了一个契机。尽管香港、台湾及海外也有一些中文搜索引擎,想统一它们的分类体系可能有一定的困难,但网上中文信息源主要是在中国大陆,只要我们的分类体系更适于组织中文信息,其他中文搜索引擎也自然会吸纳我们的精华。

2.2 原则

2.2.1 实用性原则。搜索引擎针对的主要是网络信息,而网络信息分类的类目要少而精,要具有实用性,按学科、专业聚类和展开类目。

2.2.2 直接性原则。分类的层次不宜太多,不应该出现转换十多次还找不到所需信息的情况。一般来说,在6层以下的类目比较符合人们的习惯,如果类目的层次超过6层,用户就会改用其他检索途径,比如主题词等。

2.2.3 自然性原则。网络和搜索引擎是面向大众的,因此,采用的类名基本上应该是自然语言,以满足广大普通用户的信息需求。也就是说,网络信息分类的类名学术性、专业性不应太强,要以便于用户使用为原则。

3 搜索引擎分类体系的构建方法

3.1 聚类的标准

由于网络及搜索引擎的用户范围广且受教育程度不一,网络信息分类不应以学科和专业作为聚类标准,主要标准应该是"主题和专题",其他则为辅助标准。只有这样才能满足广大网络用户希望在一个主题或专题下查全相关信息的检索需求。而按学科、专业聚类和展开系统,将增加浏览查询的难度,不是广大网络信息用户都能掌握的。因此,综合性搜索引擎的分类法应是一个关于主题的分类系统。

3.2 大类的设置

不管是传统分类法还是关于网络信息的分类法,大类的设置都要有较高的稳定性。这是因为人类的知识生产在一定的阶段上总体是稳定的,这对用户把握一个特定的分类系统是十分必要的。通常一二级类目构成网络信息分类法知识组织的核心框架,这是一个较稳定的体系。一级类目的设置就是向用户展现知识范畴的整体框架,大类的数量以15~20个为宜。数量太少,会将相关性不大的信息挤进一个类目中;数量过多,某些需要集中的主题可能被分散,同时也不便于用户在一个屏幕里迅速把握分类体系的脉络,选择浏览入口。

总体来说,网络分类法是主题划分法,以"事物"为主要聚类标准。但是,由于网络信息的多样性,用户及其信息需求的多样性,单一的逻辑划分无法满足网络信息的组织与检索需要,因此,"多重列类"成为网络分类法类目划分的重要方法。划分选取的标准,应是事物主要的特征和用户查询时最可能使用的特征。一级类目的设置,除了人文社会科学、科学技术、政治经济、教育等与学科分类共同的知识领域外,还应把网络信息丰富、用户查询和利用率很高的主题或专题有选择地列为一级类目,如新闻媒体、生活服务等。有些知识在学科分类中的层次较低,但它们是用户普遍需要获取的信息,因此,要根据多数用户查询的需要,把信息量大、点击频率高的知识范畴突出列类而不考虑其在科学分类体系中所处的层次。只有这样,才能使大多数用户在浏览中迅速进入感兴趣的视野,如计算机与因特网等。

3.3 分类体系展开的层次

分类体系展开的层次,决定着分类导航系统的详略程度,层次越多越深,知识被组织得越细密,每一个类目下信息的相关性就越高,但同时处于较低层次的信息就越隐秘,不容易被命中,削弱了分类导航系统的直观浏览功能。从目前各大中文搜索引擎来看,分类体系的层次基本上都控制在3~6级之间,比较符合网络分类法体系层次的要求。

3.4 类目的种类

由于搜索引擎分类导航系统是由分类法、分类目录和网络信息构成的统一体,因此,类目包括"子类"与"网站"两种。每一个子类下都包括可继续延伸的下一级子类和划分到终点的网站。同时,尽管网络分类系统是着眼于对信息的知识内容进行组织,但信息的载体、信息的某种形式也是用户查询信息的重要入口,因此在网络信息分类法中包括内容性类目和形式性类目,其中,形式性类目的比例较传统分类法要高,如免费资源、论坛等。

3.5 类目的名称

类名是类目内涵和外延的概括和缩影。要使用户通过类名就基本了解本类的内容范围,就要做到类名的准确、通用和精炼。准确,就是要通过字面准确揭示该类包含的知识内容,减少用户的猜测和犹豫;通用,就是要通俗易懂,尽量使用网络信息本身和用户查询最常用的术语,常用的缩略语也可以作为类名,必要时加以注释,如BBS、B2B(商家对商家)。为了界面的醒目、紧凑,一级类目的名称尤其要精炼,尽量避免累赘。

3.6 类目排列

类目的排列和检索结果的排列方法虽然对检全率和检准率不产生影响,但对用户使用的便利性和检索速度会产生影响,同时还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。因此,应采用对用户最有利的排列次序。首先应按照知识的逻辑次序和重要程度排列,共性区分的问题集中排列,采用相同标准区分的类目使用相同的次序排列。只有无明显逻辑联系的,才使用字顺排列,这对用户理解和把握知识体系是有帮助的。同时,搜索引擎的分类体系中,一个类目下展开的除了同位类外,还有若干平行的网站信息,因此,类名排列还要考虑排列这些网站信息。由于网

网络环境下图书馆定题服务刍议

张 月

（抚顺职业技术学院图书馆，抚顺 113006）

[摘 要]论述了网络环境下高职院校图书馆开展定题服务的现状，以及对情报人员提出的更高要求。情报人员如何改变现状，提高自身综合素质和能力，是定题服务工作向更深层次发展的关键。

[关键词]图书馆 定题服务 情报服务 网络环境

[分类号]G252.62

定题服务是图书情报部门根据用户研究课题和高校教学所需，通过对情报信息的收集、筛选、整理并连续不断地为用户提供服务，直到完成课题和教学任务结束。图书馆情报人员根据拥有的大量文献资料，为某一课题的研究者或教学一线的人员提供某专题的文献信息服务。定题服务是一种主动为用户提供所需的文献信息的服务，定题服务人员必须有一定的责任心和奉献精神。定题服务的对象是从事某种专题研究的科研人员和教育工作者，它是一种在一定范围内的针对特定用户的服务。定题服务从科研人员从事科学研究开始为其连续不断地提供专题文献信息服务，直到其完成某项专题研究为止。由此可见，定题服务具有针对性、主动性和连续性的特点。

1 图书馆开展定题服务的现状

1.1 网络环境使用户获取信息的路径加宽

传统图书馆馆藏文献资源处于图书馆独有的状态，用户利用馆藏文献信息必须到图书馆获取或需要图书馆情报人员帮助提供，用户对图书馆情报人员存在一定的依赖关系，所以，传统图书馆这种封闭式垄断性的文献信息资源为图书馆的定题服务工作带来生机。但是，传统图书馆的定题服务受文献检索方式和馆藏文献的数量影响，文献检索以手工检索为主，馆藏文献资源大多数以纸质出版物的形式存在，检索方式落后，检索速度缓慢，查新率较低，这样，难以满足情报用户个性化和多样化的文献信息需求。这是传统图书馆定

站信息数量大、动态性强等特点，通常采用以下方法排列：①按重要程度。通常需要通过人工干预选择，把切题程度高、内容详实丰富的优秀网站排在前面，起推荐的作用。②按点击频率。③按字顺。

3.7 类目注释和说明

搜索引擎分类体系的设计，要达到让用户在浏览过程中很容易判断浏览的方向，减少犹豫和徘徊，也就是通过类名以及与相邻类目的比较，比较确信地选定要进一步浏览的类目。这就要求通过必要的说明和注释，帮助用户了解类目的含义，尽量减少不确定性。通常有两种注释方式：一是通过精炼的文字，指明该类包含的内容范围，不包含的内容范围；二是用列举下位类的方式揭示本类的内容范围，或揭示重点的内容，或揭示热点的内容，或揭示隐藏较深的内容。凡类目下揭示的内容，用户均可以直接点击进入浏览，不必再层层查询。

随着因特网的普及和技术条件的支持，网络信息发展得比以往任何时候都要快，“垃圾信息”也越来越多，如何对网络信息资源进行有效地组织显得越来越迫切。一些专家开始呼吁以“图书馆员的思维”管理网上信息资源，也就是分门别类地把分布式的、无序的信息整理得井井有条，给用户浏览、查询提供最大的便利。强大的检索功能和科学的分类体系相结合，是搜索引擎的发展趋势。由此可以预见，分类法将在网络信息的组织中发挥越来越大的作用。同时，随着越来越多的专家和学者，尤其是图书馆学界的分类专家，开始关注和研究搜索引擎，其分类体系必将日臻完善。

参考文献：

1 俞君立，陈树年.文献分类学.武汉：武汉大学出版社，2001
2 马张华.分类搜索引擎类目体系研究.图书情报工作，2001(2)
3 马张华，张宇萌.指南型网络分类体系初探.大学图书馆学报，2000(3)
4 陈笑辉，范晓虹.Yahoo的分类体系结构及原理探微.图书情报工作，1999(9)
5 谭宁宏等. 中文搜索引擎分类体系研究. 情报学报，2001(6)
6 陈树年.搜索引擎及网络信息资源的分类组织.图书情报工作，2000(4)
7 尚加宁. 图书分类法与因特网分类方法的可比性研究.情报理论与实践，2001(1)
8 徐建华.网络搜索引擎原理、特性分析及未来发展趋势.图书情报工作，2000(8)
9 苏广利.网络信息分类系统的发展趋向研究.图书馆杂志，2002(4)

（收稿日期：2004-08-14；责编：徐向东。）