中国图书馆学报(双月刊)2003 年第 1 期
ZHONGGUO TUSHUGUANXUEBAO

●侯汉清 薛春香

中文搜索引擎分类体系兼容互换工具的设计

摘 要 对新浪、搜狐、网易三大搜索引擎分类体系的分析表明,现有中文搜索引擎分类体系之间虽存在差异但有实现兼容互换的基础。借助情报检索语言兼容互换原理,可以设计出中文搜索引擎分类体系兼容互换工具。表 4。图 1。参考文献 3。

关键词 网络信息检索 搜索引擎 分类体系 检索语言兼容互换 中介词典
分类号 G254

ABSTRACT After analyzing the classification systems of three major search engines, i.e. Sina, Sohu and NetEase, the authors think that although there are differences in classification systems of search engines, there are common grounds for their compatibility and interchangeability. By using the principles of compatibility and interchangeability of information retrieval languages, we can design compatibility and interchangeability tools for Chinese search engines. 4 tab. 1 fig. 3 refs.

KEY WORDS Network information retrieval. Search engine. Classification system. Compatibility and interchange of search engines. Intermediary dictionary.
CLASS NUMBER G254

搜索引擎是对网络资源进行标引和检索的工具,通过一定的机制和方法对网络信息进行搜索,将搜索来的信息进行分类,建立索引,然后把索引的内容存放到数据库中,以供用户检索时用。搜索引擎一般提供分类浏览和关键词查询两种方式的检索服务。本文只对其分类浏览检索进行讨论。

分类浏览,又称为"目录式"查询。它首先建立一个知识分类框架,然后将每一个大类由宽到窄逐层细分出若干等级的类目,形成知识树,把信息分门别类地组织起来。此类搜索引擎以雅虎(Yahoo!)为先,依靠专业人员进行分类标引,具有标引质量高,内容全面,方式直观,检准率高,能够限定检索范围,可以对不明确的检索目的进行引导等优点。

目前网上中文搜索引擎多如雨后春笋,为用户查找网络信息提供了便利,但是由于各搜索引擎的分类体系存在差异,限制了用户对此工具的充分利用。本文拟以新浪(http://www.sina.com.cn)、搜狐(http://www.sohu.com)、网易(http://www.163.com)为例来分析其差异,并借助情报学的相关理论,提出改进方案。

1 搜索引擎分类体系的主要差异

除了极少数的搜索引擎采用传统的图书分类法(如《杜威十进分类法》、《国际十进分类法》、《中图法》)作为分类体系外,绝大多数搜索引擎采用自创的分类体系,存在着下列多方面的差异[1]。

分类体系不同。多数搜索引擎根据自身的特点自创分类体系,类目的设置和划分各不相同。表 1 列出了新浪、搜狐、网易的"教育"类下的二级类目。通过表 1,我们可以看出,对于"教育"这个大类,新浪共分出 44 个下位类,搜狐有 34 个,网易只有 24 个。再比较这些类目,我们得到表 2 的结果。此外,在类目的划分上也存在差异,例如对各级教育的划分,新浪为"胎教—婴幼儿教育—初等教育—中等教育—高等教育—研究生教育—成人教育",搜狐则为"幼儿教育—中小学教育—高等教育—继续教育",而网易则只有"中学教育—高等教育—成人教育",没有单独列出"幼儿教育"。再如搜狐中有"各科教育",按科目内容划分教育,而网易中则设置"各地教育"按地区来划分教育。由此可见,这些分类体系在类目设置和划分上存在一定的差异。

表 1 新浪、搜狐、网易"教育"类下位类类目

新浪(44)	搜狐(34)	网易(24)
综合教育网	高等教育	语言教育与学习
学生生活	留学与移民	教育媒体
校友会与同学会	考试与入学	教育产业
网上教育	科学普及	校园生活
留学	综合网站	高等教育
家庭教育	新闻与媒体	各地教育
胎教	中等专业教育	教育技术/电教
婴幼儿教育	国外大学	中等专业教育

续表

新浪(44)	搜狐(34)	网易(24)
初等教育	组织机构	艺术教育
中等教育	中小学教育	私立学校
高等教育	国内大学	职业技术教育
研究生教育	中国教育科研网	健康教育
成人教育/ 继续教育	幼儿教育	留学与移民
职业教育	聊天与 BBS	成人/ 继续教育
师范教育	教育援助	家庭/ 校外教育
民族教育	职业技术教育	特殊教育
特殊教育	教育理论	中学教育
多媒体教育	公司企业	助学工程
语言教育	继续教育/ 培训	网络化教育
艺术教育	竞赛/ 比赛	组织机构
音乐教育	希望工程	综合性网站
计算机教育	语言教育	招生与考试
科普教育	人才/ 招聘	教育研究
成功教育	教师园地	教育学习类个人主页
性教育	远程教育	
普法教育	图书/ 书店	
军事教育	校园生活	
作文	特殊教育	
考试与招生	电脑学习资源	
题库	学生天地	
竞赛	各科教育	
学校	参考资料	
图书馆	民办教育	
教育机构组织	图书馆	
教育工作者		
教学设备与软件		
教育学		
奖学金助学金贷学金		
教育合作与援助		
论文		
新闻媒体		
论坛聊天		
书店		
公司企业		

表 2　新浪、搜狐、网易“教育”类下位类类目比较

比较项目	新浪—搜狐	搜狐—网易	网易—新浪
类目相同	10	9	7
类目相似	17	10	14
类目无关	17.7	15.5	23.3

类名的表述不相同。新浪中称“计算机”,网易中用“电脑”;搜狐中称“互联网”,而网易中则用“因特网”,新浪中用“商业经济”,搜狐中称“工商经济”,网易中用“经济金融”等。

分类深度广度不同。不同的搜索引擎,有的类目设置过细,多达十层,有的则较粗,仅有两三层。譬如我们要查找《路遥作品集》,在搜狐中要经过“文学 > 文学类别 > 小说 > 文艺小说 > 更多作家(按拼音排序) > K—L > 路遥”共七层逐步细分的浏览过程,而在新浪中只要经过“文学 > 小说 > 现当代小说 > 路遥”即可查到。

类目排序不同。虽然类目的排列顺序对于检准检全率不会产生影响,但对于用户的检索效率和使用便利性还是会产生影响。在中文搜索引擎中类目的排序一般采用以下几种方式:(1) 按照字顺排序。(2) 参考检索频率。(3) 对同位类进行系统排列。(4) 无序。各搜索引擎根据自己的特点来安排类目次序。据张琪玉对 43 种综合型网络检索工具进行统计,只有 3 种是按类名字顺排序,其余都没有明显的排序规律[2]。

助检手段不同。有的搜索引擎在大类下列出一些主要的和检索频次高的下位类,用作这个大类的说明,类似于类目注释。有的根据用户的检索频率,系统列出一些热门的检索类目,如新浪有“新浪推荐”用来在每一级类目中列出一些与之相关的最新信息。有的把分类检索和关键词检索结合起来,如在搜狐的分类检索中,还提供了关键词检索,可以通过关键词来检索分类索引数据库中的信息,并且提供“网站、网页、类目、新闻、网址”的范围限定选择,尤其是类目的限定选择。

2　兼容互换的基础

为了充分利用现有搜索引擎的分类目录下的标引质量较高的信息,提高分类检索的检全率,用户就会去多个搜索引擎中查找同一类信息。但是由于各搜索引擎存在着上述差异,用户要在多个搜索引擎之间进行查找,就会出现类目重新选择的困

难。例如,我们要查找"幼儿教育"的信息,在搜狐中通过"教育>幼儿教育"即可检索得到,而在网易中的"教育学习"大类下却检索不到,却要到"少儿乐园"大类下,经过"少儿乐园>各地幼儿园/教育>幼教"才能检索得到。为了提高检索的效率,只能走两条路:统一分类体系或者实施兼容互换。我们不可能设计一个统一的分类体系,要求现有的所有搜索引擎抛弃自己原有的分类体系推倒重来使用这个新体系。这对于个性化强烈的诸网站来说是不可能的,因此统一这条路是走不通的。那么,在不同分类体系之间实现兼容互换这条路能不能走得通呢?

首先,情报检索语言的兼容互换理论给我们提供了理论基础。兼容互换的理论就是从各个系统中间汇集相同或相近的词汇(类目),直接或通过另一通用系统来建立词汇(类目)之间的"等价关系",从而实现系统间的兼容互换。虽然我们不能要求各系统采用统一的分类体系,但是能够通过采用一定的中介系统实现不同体系之间的转换[3]。这对于用户和搜索引擎服务提供商都是能够接受的。

其次,现有的搜索引擎分类体系具备实现兼容互换的条件。(1)现有的搜索引擎分类体系包罗万象,覆盖范围大致相同。(2)分类体系框架相同,从表3的比较可以看出,新浪、搜狐、网易的大类设置是很相似的,这三者都设置了18个大类,其中新浪和搜狐有16个大类基本相同,搜狐和网易有16个大类基本相同,新浪和网易有15个大类基本相同。(3)类目划分详细,一般都在五级以上,而且对于某个类目划分下位类采用多个标准多元划分。(4)类名相同或相近,这一点从表2比较结果中就可以看出。

此外,近年来搜索引擎的发展在技术上出现趋同倾向,这导致它们在分类体系上也日渐趋同。从表3对新浪、搜狐、网易搜索引擎一级类目的比较,可以清楚看出它们大类的设置是很相似的。

表3　新浪、搜狐、网易搜索引擎一级类目对照

新浪(A)	搜狐(B)	网易(C)
娱乐休闲	娱乐休闲	娱乐休闲
计算机与互联网	计算机与互联网	电脑网络
商业经济	工商经济	经济金融
教育就业	教育	教育学习
文学	文学	文学
艺术	艺术	艺术
体育健身	体育与健身	体育竞技
医疗健康	卫生与健康	医疗健康
生活服务	生活服务	生活资讯
新闻媒体	新闻与媒体	新闻出版
科学技术	科学与技术	科学技术
政法军事	政治法律军事	政法军事
社会文化	社会与文化	社会文化
社会科学	社会科学	
参考资料		综合参考
个人主页	个人主页	
国家与地区	国家与地区	
少儿搜索		少儿乐园
	公司企业	公司企业
	旅游与交通	旅游自然
		情感绿洲

所有这些为设计搜索引擎兼容互换工具提供了理论基础和可行条件。

3　兼容互换工具的设计

要实现多个搜索引擎之间的兼容互换,必须通过一个中介系统。我们考虑建立一个中介词典,如图1。本设计以《中图法》类号作为转换中心,X代表转换中心,A、B、C分别代表参与转换的新浪、搜狐、网易的分类体系。

《中图法》是国内最通用,用户最多,维护最好的一部体系分类法,类目详尽,覆盖面广,并且有与《汉表》对应产生的《中国分类主题词表》,有利于今后分类检索与主题检索的结合。但由于《中图法》体系严密,类目划分严谨,而网络分类体系绝大多数是面向用户需要,类目设置不同于传统的分类法,所以要对《中图法》类目进行改造,使其既突出学科体系又面向事物对象。

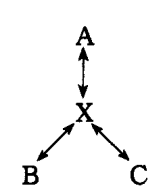

图1　中介词典原理

分类标记是一种很好的助检手段。经常查找某一类信息的专业用户,可能更倾向于使用一个分类标记直接跳转到相应的类目下实现一步到位,而不愿再一层一层浏览查找。但是绝大多数搜索引擎,无论是采用传统的分类法,还是自创的分类体系,都没有采用分类标记。因此,在这个兼容互换工具中引入了分类号用来助检。为了便于转换操作和揭示类目的等级关系,为新浪、搜狐、网易的分类体系分别设置分类标记。具体方案设计如下:

(1)结合目前网络信息资源的特点,对《中图法》类目进行改造,取到五六级类目,热点类目予以细分,检索频率低的类目予以合并,用上位类替代亦可,并保留相应的分类号。

(2)分别从新浪、搜狐、网易中逐级抽取前五级类目词,得到各自的原始类目表。

(3)用层累制分别为原始类目表编配分类标记。

(4)用《中图法》分别标引原始类目表中的各个类目,构建分类号—类目词对应表。

(5)以改造过的《中图法》为主干,以《中图法》类号作为兼容互换基础,建立各个体系之间的对应关系,并按照《中图法》类号来排序。

这样通过《中图法》这个转换中心,就可以在新浪、搜狐、网易分类体系类目之间建立关系。

表 4 是"教育"类的中介词典的片断。如果要检索"幼儿教育、学前教育"的信息,即可通过中介词典查到各系统中的相应类目,知道在新浪中为 A4. 18(婴幼儿教育,二级类目),搜狐中为 B6. 13(幼儿教育,二级类目),网易中为 C12. 5. 1(幼教,三级类目)。尽管类名表述各不相同,类级也不同,但是通过中介词典却可以实现相互之间的转换。

表 4　中介词典片断

类目名称	分　类　号	新浪	搜狐	网易
幼儿教育	G61	A4. 18	B6. 13	C12. 5. 1
幼儿园(中国)	G619. 28	A4. 18. 1	B6. 13. 3	C12. 5. 2
初等教育	G62	A4. 19	B6. 10	C12. 6
中等教育	G63	A4. 20	B6. 10	C9. 17
高等教育	G64	A4. 21	B6. 1	C9. 5
研究生教育	G643	A4. 22	B6. 1. 9	C9. 5
高等院校	G648	A4. 21. 1	B6. 1. 1	C9. 5
综合院校	G648. 1	A4. 15. 1. 1	B6. 1. 1. 1	C9. 5. 21
留学教育	G648. 9	A4. 15	B6. 2	C9. 13
师范教育	G65	A4. 25	—	C9. 5. 5
职业技术教育	G71	A4. 24	B6. 16	C9. 11
成人教育	G72	A4. 23	B6. 19	C9. 14
特殊教育	G76	A4. 27	B6. 28	C9. 16
…	…	…	…	…

为了便于检索类目,还可以考虑建立一个类目字顺索引。具体设计:(1)将中介词典中的类目词及各搜索引擎分类体系中对应的类目词抽取出来,保留类号。(2)将《中国分类主题词表》中与改造过的《中图法》类目对应的词也抽取出来,保留相应的《中图法》类号。(3)将这两组词进行去重。(4)建立字顺索引并附相应的《中图法》类号。例如,"幼儿教育"这个类目与它相关的非正式词还有"学龄前教育"、"学前教育"、"婴幼儿教育"、"幼教"等,将这些词附上相应的分类号(索引地址),按照字顺排列所有的款目。这样从这些词出发直接到中介词典中可以查找到相应的正式类目。有了这样的类目索引,检索类目就更方便,而且可以使用分类号来检索。

目前有一种元搜索引擎,能够利用多个独立的搜索引擎进行查询。当用户将检索要求提交给元搜索引擎后,它将这个检索要求同时交给多个独立的搜索引擎进行查找,收到检索结果后,对这些结果进行加权等处理后返回给用户。目前因特网上的元搜索引擎都是基于关键词进行检索的,利用前面设计出的分类体系兼容互换系统和类目索引,我们是不

●林 曦 周 磊

在图书情报服务中导入 CS 战略研究

摘 要 CS 战略即"使用户满意"战略。CS 战略属性表明,它对图书情报服务具有潜在价值、现实价值和未来价值。中国图书馆事业的基本矛盾和实际状况已为 CS 导入提供了可能。参考文献 6。

关键词 图书情报 CS 战略 用户服务
分类号 G251

ABSTRACT The attributes of CS strategy, or customer satisfaction strategy, indicate that it has potential, realistic and future values for library and information services. The basic contradictions and present conditions of Chinese librarianship provide possibility for the introduction of CS strategy. 6 refs.

KEY WORDS Library and information services. CS strategy. Customer service.
CLASS NUMBER G251

在全球日益信息化的新形势下,图书情报服务的实践迫切需要新的理论来加以指导。CS 战略,作为全球企业界在 20 世纪 90 年代末期流行的一种新型的管理理念,作为企业参与 21 世纪市场竞争经营的"通行证",正日益引起人们的关注和重视[1]。因此,将 CS 战略引入运用到现代图书情报服务活动中,不失为一种有效的战略管理形式。

1 运用 CS 战略的现实意义

CS (Customer Satisfaction) 战略即"使用户满意"战略。将 CS 战略运用于现代图书情报服务中,旨在使用户能得到完全满意的图书情报服务和产品。CS 战略是一种服务战略,对其运用可以综合而客观地测定用户的满意程度,并在对读者用户调查分析和研究结果的基础上,使整个图书情报机构作为一个整体来系统改善服务,改善产品及改善图书情报文化。CS 战略要建立的是一种革命性系统,即用户至上的服务体系,使用户满意度达到百分之百[2],从而使图书情报工作效益倍增。

首先,CS 战略的服务对象已超越了"读者"范围,代之以"用户"概念;再者,CS 战略促使图书情报服务所用的信息源突破了"馆藏"局限,其服务方式、服务

是可以考虑建立一个基于分类浏览的元搜索引擎呢?这样,用户就不用再去各个搜索引擎中分别查找,可以在一个统一的界面上从系统提供的、经过改造的《中图法》分类体系中直接查找,就可以获得在多个搜索引擎中检索的结果。

搜索引擎是面向大众用户的。随着它日益成为一种主要的信息源,专业用户越来越多。对于这些专业用户,可以提供分类号检索服务,而且通过类目索引可以从多个入口来查找到相应的类目,为用户提供多途径检索的便利,充分利用各大搜索引擎人工标引的高质量数据库。

总而言之,这种利用情报检索语言兼容互换理论设计的搜索引擎分类体系兼容互换工具,对于提高分类浏览检索的效率是很有帮助的,值得进一步研究和开发。

参考文献

1 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)
2 张琪玉. 网络信息检索工具发展的方向与提高竞争力的途径. 深圳:巨灵信息技术研究所, 2000. 3(巨灵研究报告,编号为 GTJ/ TR002)
3 侯汉清. 当代分类法主题法索引法研究. 北京: 书目文献出版社, 1993

侯汉清 南京农业大学信息管理系教授。通讯地址:江苏南京。邮编 210095。
薛春香 南京农业大学信息科技学院 01 研。通讯地址同上。
(来稿时间:2002-04-29)