●陈万寅　金明华　邹小筑

中文网络学术数据库数据质量分析*

摘　要　对中文期刊全文数据库"重庆维普中文科技期刊"、"清华同方中国期刊网"和"万方期刊"收录核心期刊的种类数量、收录率以及它们之间重复收录的情况进行分析比较。对这 3 个期刊数据库和两个电子图书数据库进行检索测试。3 个期刊数据库在收录核心期刊数量上差别不大,但重复建设突出。中文电子图书分类等有待规范统一。表 5。图 1。参考文献 5。

关键词　中文数据库　网络数据库　电子图书　评价

分类号　G250.74

ABSTRACT　The authors make a comparative study of some online Chinese full-text databases, including VIP Information's Chinese Scientific Journals Database, Tsinghua Tongfang's CNKI, Wanfang Data and other databases, concerning their coverage and duplication. After tests and analysis, they find that the numbers of journals they cover are similar, but there are quite a large number of duplicated titles in all these databases. They also think that the classification methods of Chinese electronic books are to be unified. 5 tabs. 1 fig. 5 refs.

KEY WORDS　Chinese database.　Network database.　Electronic book.　Evaluation.

CLASS NUMBER　G250.74

目前国内学术文献信息服务市场上,利用率最高、影响范围最广、市场份额最大的基于互联网的中文期刊全文数据库主要有"重庆维普中文科技期刊"(以下简称 VIP)、"清华同方中国期刊网"(以下称 CNKI)和"万方数据——中国数字化期刊群"(以下简称万方期刊)。中文电子图书全文数据库则有"超星"、"书生之家"、"方正电子图书"等。

本文着重对上述 3 个中文期刊数据库的基本情况,收录核心刊的种类数量、收录率和质量情况,它们之间数据重复情况以及各自的相关数据等进行统计、分析、比较与研究,并对它们和两个电子图书数据库进行检索和测试(文中所有原始数据截止 2005 年 12 月 31 日)。

1　三大期刊数据库收录总量与类别比较

为保证数据的可比性,我们对 CNKI、万方期刊与 VIP 镜像站点数据进行了更新,而且统一更新至同一时间结点。并采用人工批量复制静态数据即镜像数据的方法,复制导出三大数据库刊名列表。因为 VIP 在 3 个数据库中最早采用中图法对刊名分类,所以我们以它为参照蓝本,将 CNKI、万方期刊进行计算机程序自动比对,并将 CNKI、万方期刊中没有被 VIP 收录的刊名列表由我馆专业人员进行分类,最终形成三大库、且以中图法 21 个分类为统一分类标准的、类与类之间不重复的刊名种类数量比较信息(见表 1)。

表 1　三大数据库种类数量基本情况

中图法分类	VIP		CNKI		万方	
	收录量(种)	占总量(%)	收录量(种)	占总量(%)	收录量(种)	占总量(%)
总量	14415		7260		5016	
A 马克思主义、列宁主义、毛泽东思想、邓小平理论	7	0.0485	5	0.0689	1	0.0199
B 哲学、宗教	57	0.395	38	0.523	16	0.319
C 社会科学总论	719	4.99	464	6.39	316	6.30
D 政治法律事务	813	5.64	385	5.30	174	3.47
E 军事	98	0..680	32	0.441	15	0.299
F 经济	1601	11.1	620	8.54	286	5.70

* 本文为江苏省哲学社会科学基金项目"中文网络学术资源利用调查与分析"(04XWB009)研究成果之一。

续表

中图法分类	VIP		CNKI		万方	
	收录量(种)	占总量(%)	收录量(种)	占总量(%)	收录量(种)	占总量(%)
G 文化、科学、教育、体育	1870	13.0	996	13.7	511	10.2
H 语言、文字	78	0.541	63	0.868	24	0.478
I 文学	278	1.93	87	1.20	42	0.837
J 艺术	189	1.31	102	1.40	37	0.738
K 历史、地理	148	1.03	89	1.23	37	0.738
N 自然科学总论	642	4.45	483	6.65	391	7.80
O 数理科学和化学	185	1.28	145	2.00	108	2.15
P 天文学、地球科学	429	2.98	228	3.14	197	3.93
Q 生物科学	131	0.909	88	1.21	77	1.54
R 医药、卫生	1673	11.6	997	13.7	879	17.5
S 农业科学	941	6.53	530	7.30	429	8.55
T 工业技术	3787	26.3	1575	21.7	1232	24.6
U 交通运输	463	3.21	183	2.52	129	2.57
V 航空、航天	151	1.05	55	0.758	48	0.957
X 环境科学安全科学	156	1.08	93	1.28	67	1.34

VIP 拥有期刊种类最多,万方最少。VIP 中工业技术类所占份额最大,文学、政治法律、交通运输类数量也超过其他库的两倍;CNKI 除在语言文字类上份额稍有优势,看不出特别的偏重,属于各学科比较均衡的数据库类型;而万方期刊在医药卫生、农业科学、数理和化学类上份额超过其他两个数据库,比较偏重此方面,其余两数据库相比较而言则没有优势。

2　三大期刊数据库收录核心刊的数量质量比较

2.1　数量比较

我们以国内较权威的北京大学图书馆《中文核心期刊要目总览》(2004 年版)(简称北大)认定的核心期刊(1797 种)为参考蓝本,统计三大库收录的核心期刊数量和收录率。VIP 收录 1716 种,收录率为 95.492%;CNKI 为 1632 种,90.818%;万方期刊为 1314 种,73.122%。

我们仍然以 VIP 分类为参照蓝本,先对它收录北大刊名数据进行计算机程序自动比对,得到收录北大的分类列表,并采用同样方法,分别得到 CNKI 和万方收录北大的分类列表(见表 2)。

表 2　收录核心刊的种类及收录率

北大核心刊分类	VIP		CNKI		万方	
	收录量(种)	收录率(%)	收录量(种)	收录率(%)	收录量(种)	收录率(%)
A 马克思主义、列宁主义、毛泽东思想、邓小平理论 4	4	100	4	100	1	25
B 哲学、宗教 20	19	95	19	95	8	40
C 社会科学总论 108	103	95.370	99	91.667	79	73.148
D 政治法律事务 77	70	90.909	67	87.013	25	32.468
E 军事 11	11	100	6	54.545	5	45.455
F 经济 150	142	94.667	127	84.667	62	41.333
G 文化、科学、教育、体育 176	170	96.591	159	90.341	82	46.591
H 语言、文字 29	26	89.655	26	89.655	13	44.828
I 文学 51	50	98.039	30	58.824	13	25.49
J 艺术 37	30	81.081	26	70.27	10	27.027

续表

北大核心刊分类	VIP		CNKI		万方	
	收录量(种)	收录率(%)	收录量(种)	收录率(%)	收录量(种)	收录率(%)
K 历史、地理 30	30	100	28	93.333	13	43.333
N 自然科学总论 105	84	80	88	83.81	86	81.905
O 数理科学和化学 81	75	92.593	78	96.296	72	88.889
P 天文学、地球科学 82	81	98.78	78	95.122	75	91.463
Q 生物科学 48	47	97.917	47	97.917	46	95.833
R 医药、卫生 211	210	99.526	201	95.261	200	94.787
S 农业科学 115	115	100	111	96.522	106	92.174
T 工业技术 387	373	96.382	364	94.057	347	89.664
U 交通运输 36	36	100	35	97.222	34	94.444
V 航空、航天 18	18	100	18	100	17	94.444
X 环境科学安全科学 21	21	100	21	100	20	95.238

注:类名后的数字为核心刊数量。

VIP在A、E、K、S、U、V、X等类都是100%收录北大核心刊。而VIP和CNKI收录的数量在每个类中都差不多,它们在收录核心刊上没有大的区别。在N、O、P、Q、R、V、X类中,三大库收录的核心刊数量都差不多;在这些类上它们之间也没有太大区别。

2.2 质量分析(缺期刊种类数量和缺期数量)

我们对它们所收录的每个核心刊均采用了三大指标作为衡量标准,即回溯情况、中间缺期情况以及滞后情况,也就是更新速度。用这些指标基本能了解三大库的全貌。

VIP主要起始于1989年和2000年这两个年份,回溯中属于不连续新增的较少,只有15种刊。有近1/4在2001~2005年间缺期或缺年,有1/3在2001年以前缺期缺年。更新速度比较慢,有近90%滞后。

CNKI主要回溯到1994年和1986年以及1986年以后的,它们一共占了将近70%。属于不连续新增的较少,只有10种刊。有接近1/3的刊缺期缺年,且大部分集中在1994到2005年,但1994年以前缺全年的比较多。它的更新速度最好,滞后的刊共有540种占31.47%,其中滞后1~2期占绝大多数。

万方主要回溯到2000年和1999年,一共占了将近70%。属于不连续新增的较少只有10种刊。有接近1/3的刊缺期缺年。更新速度较好,虽然滞后的刊共有699种占53.2%,但是大部分只滞后1期。

三大库缺年缺期的情况仍然比较严重。其中以VIP最为严重,尽管它的总量是CNKI的两倍左右,但是缺期却是CNKI的3~4倍.而其核心刊的收录率只比CNKI多5%。三大数据库的更新速度比以前都有所提高。

3 三大期刊数据库之间重复收录期刊数据分析

我们仍然采用计算机程序自动比对方法,将三大库两两之间进行比较,得到表3和图1。

表3 三大数据库之间重复收录期刊数据

重复数据库	VIP与CNKI 6307	VIP与万方 4599	CNKI与万方 4559	三大数据库共同 4406
占VIP%(重复率)	43.753	31.904	30.565	30.565
占CNKI%	86.873	60.689	62.796	60.689
占万方%	87.839	91.687	90.889	87.839

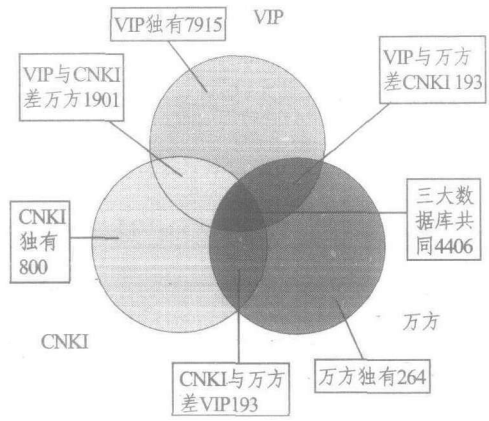

图1 三大数据库期刊重复情况

VIP 独有 7915 种,占自己的 54.91%,CNKI 独有 800 种,占自己 11.02%,万方期刊仅独有 264 种占自己 5.74%,也就是说 CNKI 有 89%,万方有 94% 的期刊基本都被其他两家数据库收录。VIP 也有一半被其他库收录。三大期刊数据库的重复建设非常严重。

4　三大期刊数据库检索测试分析

我们在设定相同检索条件的情况下,对这三大期刊数据库进行检索测试,分别测试了主题途径检索、分类途径检索、作者字段检索、全文检索、引文检索功能,结果见表 4。

表 4　三大期刊数据库检索测试

检索字段	检索词	CNKI 返回数量	VIP 返回数量	万方期刊 返回数量
篇名/关键词/摘要	书生之家	37	19	8
	超星	150	96	24
	情报检索	888	2184	185
	方正	4535	3406	480
	数据挖掘	6004	5490	4228
	成形极限	199	187	125
	有限元	26273	33666	21153
篇名	书生之家	4	4	2
	超星	37	39	7
	情报检索	295	295	71
	方正	1367	1773	230
	数据挖掘	2134	2057	1648
	成形极限	63	68	40
	有限元	9095	10757	6453
文摘	书生之家	37	17	7
	超星	150	51	19
	情报检索	758	288	124
	方正	4523	1873	250
	数据挖掘	4165	3508	2879
	成形极限	172	151	104
	有限元	19981	28739	19081
关键词	书生之家	7	10	1
	超星	41	65	5
	情报检索	561	2090	125
	方正	1391	2161	137
	数据挖掘	4552	4831	3438
	成形极限	110	105	60
	有限元	17633	22293	12900
作者	梁战平	55	51	23
	侯惠勤	35	21	11
	高霖	54	53	29
	朱兆达	102	73	42
	朱岱寅	23	9	17

续表

检索字段	检索词	CNKI 返回数量	VIP 返回数量	万方期刊 返回数量
分类	G354(情报检索)	2869	6841	
	D911(国家法、宪法)	59577	665	
	TH122(机械设计)	3274	5216	
	H31(英语)	76478	35564	
	F27(企业经济)	395027	420471	
全文	书生之家	692		8
	超星	2455		35
	情报检索	10596		185
	方正	47185		1398
	数据挖掘	19936		4244
	成形极限	990		125
	有限元	81810		21153
引文检索功能	书生之家	32	篇名、刊名、作者、任意字段4个选项。 篇名=书生之家,命中5篇。 任意字段=书生之家,命中6篇。	"被引论文"=书生之家,命中20篇。
	超星	123	篇名=超星,命中8篇。 任意字段=超星,命中14篇。	"被引论文"=超星,148篇。
	情报检索	1366	篇名=情报检索,命中167,任意字段270。	593
	方正	6347	篇名49,任意字段490。	4699
	数据挖掘	2496	篇名419,任意字段717。	4866
	成形极限	87	篇名22,任意字段27。	65
	有限元	12562	3521, 5143	17254

对表4进行分析可知:

万方期刊收录量最小,CNKI与VIP大致相当。

从主题检索途径看,VIP期刊关键词检索功能较强,CNKI的文摘和篇名检索功能较强。

从分类检索看,CNKI与VIP期刊都按"中图法"进行分类,检索方便。而"万方期刊"比较薄弱。

从全文检索功能看,VIP期刊较弱。

从引文检索功能看,CNKI最强,万方次之,VIP期刊引文检索功能稍弱。但VIP期刊将被引字段分别设置为篇名、刊名、作者、任意字段4个选项,检索比其他两个期刊数据库方便。

CNKI提供的检索字典较多,有6个字段(作者、关键词、机构、基金、中文刊名、主题词)提供了检索词字典,通过使用它可以规范所输入的检索词,有利于更全更准地检索文献信息。VIP期刊只提供了1个字段(作者)的检索词字典。万方期刊没有这项功能。

CNKI具有基金检索、知识关联检索功能。这项功能比较强大,其他两个期刊数据库在这方面比较薄弱。

3个期刊数据库的去重功能均不完善。

5 中文电子图书数据库检索测试分析

我们在设定相同检索条件的情况下,对"超星电子图书"、"书生之家电子图书"也进行了检索测试。按两种电子图书数据库检索字段的并集,分别测试了书名、作者字段检索、索书号、出版时间、出版机构5种字段检索功能和分类导航功能。结果见表5。

表5　中文电子图书数据库检索测试

检索字段	检索词	超星	书生
书名	情报检索	37	1
	数据挖掘	32	2
	信息管理	164	14
	政治	3843	487
	机械	1774	833
作者	梁战平	5	1
	苏新宁	6	0
	马费成	9	0
	侯惠勤	3	0
	刘思峰	4	0
索书号	D911(国家法、宪法)	88	无此字段
	G354(情报检索)	109	无此字段
	TH122(机械设计)	492	无此字段
	F27(企业经济)	15483	无此字段
出版时间	2004	31941	无此字段
	2005	808	无此字段
	2003	43562	无此字段
	2002	53077	无此字段
	2001	45960	无此字段
	1990	17354	无此字段
	1980	1879	无此字段
	1970	7	无此字段
	1960	23	无此字段
出版机构	机械工业	无此字段	1356
	电子工业	无此字段	846
	中国社会科学	无此字段	824
	高等教育	无此字段	7951
分类导航功能	图书馆学	158	124
	信息与传播理论	无法查	107
	信息处理技术	53	29

从总体上看，超星收录书的数量比书生多，综合性强，且检索方便。

从高级检索功能看，超星有“包含”、“等于”2种运算，检索字段与简单检索相同；书生比超星多“丛书名称、ISBN、主题、摘要”4个字段，但无“索书号字段”，书生的类似功能由“图书分类”导航检索完成。

从导航检索功能看，书生可以查到更多级分类，如“信息与传播理论”为“信息处理技术”的上一级，超星无法查到。

超星分类基本按照“中图法”，如：按《中国图书馆分类法(第四版)》类目设置为A马克思主义、列宁主义、毛泽东思想、邓小平理论，而超星的“A经典理论图书馆”下设类目与“中图法”相比增加了“总论”类目。书生设有2种分类标准，一是“中图法”分类，另一是“书生分类”。但目前实际分类是未采用“书生分类”。用“书生”查“企业经济”使用导航找到3条相关类目，企业经济理论和方法110，各种企业经济206，而世界各国企业经济图书暂未上传。超星中“企业经济”无法直接查。

两个数据库相似内容的类目设置名称不统一。比如超星“信息与知识传播”类目下设：总论，信息与传播理论，新闻学、新闻事业，广播、电视事业，出版事业，群众文化事业，图书馆学、图书馆事业，博物馆学、博物馆事业，档案学、档案事业。而书生“知识信息传媒”下设：信息与传播理论，新闻学、新闻事业，广播电视，出版事业，群众文化，图书馆学，情报学、情报工作，博物馆学、博物馆事业，档案学、档案事业。

6　结论

三大期刊数据库涵盖21个大类并且分布都比较平均，都以收录核心刊作为核心竞争力，在收录核心刊数量上差别不大。重复建设特别突出，三大库都有的期刊达4000多种。

标引深度，VIP期刊优于CNKI。VIP期刊适合选择关键词字段检索，CNKI则用篇名和文摘字段检索较好。CNKI与VIP期刊都采用“中图法”进行分类，检索方便而准确，而万方期刊相对比较薄弱。全文检索功能，VIP期刊较弱。引文检索功能，CNKI最强，万方次之，VIP期刊引文检索功能稍弱，但VIP期刊将被引字段分别设置为篇名、刊名、作者、任意字段4个选项，检索比其他两个期刊数据库方便。

电子图书中，从总体上看，超星收录书的数量比书生多，综合性强，且检索方便。两个数据库相似内容的类目设置名称不统一。中文电子图书的分类有待规范统一，两种电子图书采用两种不同的全文格式，也有待统一。

参考文献

1　[2005-12-31]. 中国学术期刊网 http://www.cnki.net
2　[2005-12-31]. 维普中文科技期刊 http://202.195.136.17/
3　[2005-12-31]. 万方数字化期刊 http://202.119.70.19:85/szhqk/index.html
4　[2005-12-31]. 超星数字图书馆 http://202.119.47.40:8080/
5　[2005-12-31]. 书生之家 http://202.119.70.18:85/

陈万寅　南京航空航天大学图书馆研究馆员。通信地址：南京。邮编210016。

金明华　邹小筑　南京航空航天大学图书馆副研究馆员。通信地址同上。

(来稿时间：2006-10-19)