我国情报学硕士学位论文的共词聚类分析

李长玲，翟雪梅

（山东理工大学 科技信息研究所，山东 淄博 255049）

摘 要：利用《CNKI中国优秀硕士学位论文全文数据库》中收录的624篇情报学硕士学位论文，对高频关键词进行共词聚类分析，研究各高频关键词之间的内在关系，探索情报学硕士学位论文的研究热点。

关键词：情报学；学位论文；共词分析；聚类分析

中图分类号：G350　　**文献标识码**：A　　**文章编号**：1007－7634(2008)01－0073－04

Co－word Clustered Analysis of Doctor Information Science Dissertations in China

LI Chang－ling，*ZHAI Xue－mei*

（*Science and Technology Information Research Institute*，*Shandong University of Technology*，*Zibo* 255049，*China*）

Abstract：This paper gives a statistical analysis of 624 dissertations of Information Science from CNKI China Excellent Doctor Thesis Full－text Database. A co－word clustered analysis was made for highly key－words and co－word and the inner relations among them to investigate the hot points of Information Science dissertation.

Key words：information science；dissertations；co－word analysis；clustered analysis

我国情报学的研究对象先后经历了从文献到信息再到知识的变革。情报学研究对象的转变，使情报学研究内容不断更新。研究内容的广泛、研究领域的扩展、研究方法的丰富，给情报学发展带来了新的希望，网络时代的到来，更为情报学提供了广阔的发展空间。因此，关于情报学热点的研究具有重要的理论和现实意义。

硕士学位论文一般都具有专深的理论和卓越的见解，具有内容新颖、信息量大、专业性强、学术价值高等特点，其发表状况被认为是衡量学科发展水平和科技产出的一项重要指标[1]。本文应用共词聚类的方法，对近几年的情报学硕士学位论文进行定量分析，找出当前我国情报学硕士研究生的研究热点。

1 数据来源

《CNKI中国优秀硕士学位论文全文数据库》是目前国内相关资源最完备、高质量、连续动态更新的中国硕士学位论文全文数据库，收录了1999年至今全国652家硕士培养单位的优秀硕士学位论文。

本文选择中国知网的《CNKI中国优秀硕士学位论文全文数据库》的免费题录数据库[2]，于2007

收稿日期：2007－05－08

作者简介：李长玲（1969－），女，硕士，副研究馆员，从事知识管理与科学评价研究；翟雪梅（1982－），女，硕士研究生，从事知识管理研究.

年4月9日，以“学科专业名称”作为检索途径，输入“情报学”作为检索词，时间选择2002年到2006年，共检索到624篇学位论文。

对检索结果用Excel进行数据统计，共得到1633个关键词。选择词频数不小于10的关键词作为高频关键词进行分析，同时去除了对反映主题没有积极意义的“对策”（词频为12）和“研究”（词频为10）两个关键词，得到的20个高频关键词，如表1所示。

表1 情报学硕士学位论文高频关键词表

序号	关键词	词频	序号	关键词	词频
1	电子商务	44	11	客户关系管理	12
2	知识管理	30	12	图书馆	12
3	网 络	24	13	信息系统	12
4	企 业	18	14	电子政务	11
5	数据挖掘	17	15	信息技术	11
6	信息化	17	16	数字图书馆	10
7	信息服务	15	17	信息检索	10
8	竞争情报	14	18	信息资源	10
9	人力资源	14	19	信息组织	10
10	信 息	13	20	本 体	10

2 数据分析

2.1 构造共词矩阵

对20个关键词两两配对，统计它们在624条文献中共同出现的频次，形成20×20的矩阵，如表2所示。

2.2 构造相异矩阵

为了消除频次悬殊造成的影响，用Ochiia系数将共词矩阵转换成相关矩阵，即将共词矩阵中的每个数字都除以与之相关的两个词总频次开方的乘积，其计算公式是：

Ochiia系数 = A、B两词同时出现频次/（$\sqrt{A}$词总出现频次 * $\sqrt{B}$词总出现频次）

对角线上的数据表示某词自身的相关程度，经上式计算均为1。为方便进一步处理，用“1”与全部矩阵相减，得到表示两词间相异程度的相异矩阵[3]，如表3所示。

表2 情报学硕士学位论文高频关键词的共词矩阵

序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	44	1	8	2	1	4	0	0	0	1	3	2	2	1	3	1	0	3	0	0
2	1	30	0	8	1	3	0	1	2	0	1	5	1	2	2	0	0	1	0	0
3	8	0	24	10	0	1	6	1	0	0	3	1	0	1	2	0	4	8	5	0
4	2	8	10	18	0	12	1	11	3	0	1	1	7	0	1	0	0	2	0	0
5	1	1	0	0	17	0	0	1	1	0	3	0	1	0	0	0	2	0	0	0
6	4	3	1	12	0	17	0	0	0	2	0	0	2	2	3	0	0	1	0	0
7	0	0	6	1	0	0	15	0	0	0	0	6	3	0	0	1	2	3	1	1
8	0	1	1	1	1	1	0	0	14	0	0	0	0	2	0	1	0	0	0	0
9	0	2	0	3	1	0	0	0	14	2	1	2	3	0	1	2	0	1	0	0
10	3	1	8	2	5	2	3	0	2	13	1	5	5	3	1	3	3	1	5	0
11	3	1	3	1	3	0	0	0	1	0	12	1	1	2	0	0	0	1	0	0
12	2	5	1	1	0	0	6	0	2	0	1	12	0	0	1	0	1	5	1	1
13	2	1	0	7	1	2	3	2	3	5	1	0	12	0	0	0	0	1	0	3
14	1	2	1	0	0	2	0	0	0	0	2	0	0	11	0	0	0	3	0	0
15	3	2	2	1	0	3	0	1	1	1	0	1	0	0	11	1	0	1	0	0
16	1	0	0	0	0	0	1	0	2	0	0	0	0	0	1	10	0	3	0	1
17	0	0	4	0	2	0	2	0	0	0	0	1	0	0	0	0	10	3	2	2
18	3	1	8	2	0	1	3	0	1	1	1	5	1	3	1	3	3	10	5	0
19	0	0	5	0	0	0	1	0	0	1	0	1	0	0	0	0	2	5	10	1
20	0	0	0	0	0	0	1	0	0	0	0	1	3	0	0	1	2	0	1	10

表3 情报学硕士学位论文高频关键词的相异矩阵(部分)

序号	1	2	3	4	5	6	7	8
1	0		0.972476	0.753817	0.928933	0.963436	0.853746	1
2	0.972476	0	1	0.655735	0.955719	0.867158	1	0.951205
3	0.753817	1	0	0.518875	1	0.950493	0.683772	0.945446
4	0.928933	0.655735	0.518875	0	1	0.314006	0.939142	0.307065
5	0.963436	0.955719	1	1	0	1	1	0.93518
6	0.853746	0.867158	0.950493	0.314006	1	0	1	1
7	1	1	0.683772	0.939142	1	1	0	1
8	1	0.951205	0.945446	0.307065	0.93518	1	1	0

2.3 利用 SPSS 进行聚类分析

将表 3 所示相异矩阵导入 SPSS 进行层次聚类分析，选择“组间平均链锁（Between – group linkage）距离”，即个体与小类中每个个体距离的平均值。此种方法利用了个体与小类的所有距离的信息，克服了极端值造成的影响[4]。得到的凝聚状态表如表 4 所示。

表 4　层次聚类分析的凝聚状态表

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	7	12	.900	0	0	10
2	3	18	.964	0	0	4
3	4	8	1.112	0	0	9
4	3	19	1.158	2	0	6
5	5	11	1.366	0	0	11
6	3	17	1.525	4	0	10
7	9	13	1.564	0	0	14
8	1	15	1.566	0	0	13
9	4	6	1.620	3	0	19
10	3	7	1.779	6	1	16
11	5	14	1.833	5	0	15
12	16	20	1.883	0	0	14
13	1	2	1.928	8	0	15
14	9	16	1.993	7	12	17
15	1	5	2.005	13	11	17
16	3	10	2.122	10	0	18
17	1	9	2.181	15	14	18
18	1	3	2.367	17	16	19
19	1	4	2.699	18	9	0

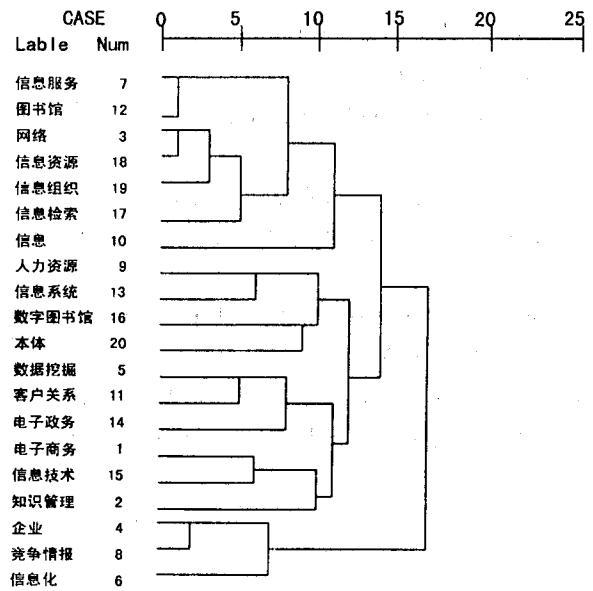

图 1　层次聚类分析的树状图

表 4 中，第一列表示聚类分析的第几步；第二、三列表示本步聚类中哪两个样本或小类聚成一类；第四列是个体距离或小类距离；第五、六列表示本步聚类中参与聚类的是个体还是小类，0 表示样本，非 0 表示由第几步聚类生成的小类参与本步聚类；第七列表示本步聚类的结果将在以下第几步中用到。例如，第一步中，7 号关键词（信息服务）与 12 号关键词（图书馆）聚成一类，它们的个体距离是 0.900，这个小类将在第 10 步中用到。同理可得其它聚类。这个聚类过程也可以从图 1 所示树状图中展现出来。

树状图以躺倒树的形式展现了聚类分析中的每一次类合并的情况。SPSS 自动将各类间的距离映射到 0 ~ 25 之间，并将凝聚过程近似地表现在图上。7 号关键词（信息服务）与 12 号关键词（图书馆）距离最近，首先合并成一类，其次是 3 号关键词（网络）和 18 号关键词（信息资源），以此类推。可见，该聚类过程与表 4 所示的凝聚状态图是一致的。

3　结　　论

结合高频关键词的共词矩阵和上述聚类过程，

我国情报学硕士学位论文的研究热点可以概括为以下几类：

（1）图书馆的信息服务，包括关键词7、12、10。网络环境下图书馆的信息服务更加体现出其个性化特点。主要围绕信息服务实现的技术、系统、对策及信息服务能力的评价展开。同时也有涉及专业图书馆的信息服务研究，像医学图书馆和军事图书馆等。

（2）网络信息的组织和检索，包括关键词3、18、19、17、10。以信息构建、Web、本体、XML、统计语言学模型等各种理论为基础的信息组织和检索是研究的热点。还涉及一些企业、医学等专业方面的信息组织和检索，以及检索系统的设计。

（3）数据挖掘在客户关系管理及电子政务中的应用，包括关键词5、11、14。数据挖掘技术是实施客户关系管理的关键技术之一。企业在收集大量的客户基本资料和详细交易数据的基础上，利用数据挖掘发现客户特征、购买模式等有价值的知识，从而有效指导客户关系管理的实践。将这种思维运用到电子政务中，同样具有重要的意义，因此得到了硕士研究生的关注。

（4）企业竞争情报和企业信息化，包括关键词4、8、6。关于企业竞争情报的研究主要集中在竞争情报系统的构建及应用、企业竞争情报需求分析、人才培养及反竞争情报等方面。企业信息化的研究主要是关于企业信息化的经济效益评价、项目风险评估和控制、信息化对策等的研究。

（5）本体与数字图书馆，包括关键词16、20。本体是近几年的研究热点，主要用于知识检索、信息系统建模、领域本体建模和信息服务系统等方面。而这些问题又是数字图书馆建设和运作过程中所不可忽视的。

（6）人力资源与信息系统，包括关键词9、13。企业、高校、银行等领域的人力资源管理系统的设计及实施是其研究重点。

（7）电子商务、知识管理及信息技术的关系，包括关键词1、15、2。电子商务和知识管理是出现最多的两个关键词，它们涉及的范围比较广泛。知识管理主要是针对企业和图书馆，大多研究知识管理系统的设计与实现；电子商务的模式选择、税收问题、消费者信任度以及信息流、信息安全等问题是研究热点。当今网络环境下，知识管理和电子商务的实现都与信息技术有着不可分割的联系。

共词聚类分析不同于普通的文献计量方法，它能定量反映出词与词之间的亲疏关系，进而反映这些词所代表的主题内容的结构。

本文不仅利用高频关键词反映情报学硕士学位论文的研究热点，还通过共词聚类分析反映这些热点内容的结构关系。但是由于所选关键词的多少造成的聚类结果有所不同，因此不排除有些出现频次较低的关键词可能成为未来的研究热点[5]。同时，尽管《CNKI中国优秀硕士学位论文全文数据库》具有很高的权威性，但其录入数据有一定的滞后性和片面性，可能存在数据的漏检或误检。因此，我们的统计分析难免可能会出现差错和缺漏。但我们的目的是通过对我国情报学硕士学位论文的热点分析，进一步了解我国情报学研究生的研究方向。

参考文献

1 栗 莉.21世纪情报学的学科定位[J].情报理论与实践，2001,(3):169-171.

2 http://lsg.cnki.net/grid20/Navigator.aspx? id=9,2007-04-09.

3 郑华川，于晓欧，辛 颜.利用共词聚类分析探讨抗原CD44研究现状[J].中华医学图书情报杂志，2002,(2):1-3.

4 薛 薇.SPSS统计分析方法及应用[M].北京：电子工业出版社，2005:310-313.

5 侯跃芳，崔 雷.医学信息存储与检索研究热点的共词聚类分析[J].中华医学图书情报杂志，2004,(1):1-4.

（责任编辑：徐 波）