◦博士论坛◦

研究热点分析

付鑫金[1,2],方 曙[1],庞弘燊[1,2]

(1.中国科学院 国家科学图书馆成都分馆,四川 成都 610041;
2.中国科学院 研究生院,北京 100190)

摘 要:通过选取我国情报学博硕士学位论文的高频关键词进行两两统计其在同一篇文献中出现的次数,来构造共词矩阵,进而转化为相关矩阵、相异矩阵。对不同的矩阵进行多维尺度分析、聚类分析、因子分析。综合分析结果发现当前的研究热点主要有三方面:信息检索技术的理论与实践,图书馆的建设与服务,电子商务、电子政务的信息化。

关键词:共词分析;情报学;博硕士学位论文

中图分类号:G350 **文献标识码**:A **文章编号**:1007-7634(2011)11-1722-04

Research on Hotspots of Information Science Dissertations of Ph. D. and Master Degree in China Based-on Co-word Analysis

FU Xin-jin[1,2], *FANG Shu*[1], *PANG Hong-shen*[1,2]

(1.*Chengdu Branch of the National Science Library, Chinese Academy of Sciences, Chengdu* 610041, *China;* 2.*Graduate University of Chinese Academy of Sciences*, *Beijing* 100190, *China*)

Abstract: High frequency keywords in Information Science dissertations of Ph. D. and master degree in China are chosen to form the co-word matrix. Then transform the co-word matrix into correlation matrix and dissimilarity matrix. Three different methods analyze the appropriate matrix to find out there are three main streams which reflect current research hotspots. They are information retrieval theory and practice, library construction and service, information-based e-commerce and e-government.

Keywords: co-word analysis; information science; dissertations of Ph. D/master degree

1 引 言

随着时代的变迁及科学技术的不断发展,我国情报学的研究领域也在不断地拓展,情报事业越来越受到重视,新的研究热点不断出现。研究生作为学科研究中的新生力量,其学位论文可以反映该学科发展的最新动态。学位论文专业性强、内容新颖[1],故此本文选取学位论文作为研究对象,进行共词分析,进而探寻近年来我国情报学发展的热点动态。

共词分析方法是从文献数据库中抽取出高频关键词,对一组词两两统计它们在同一篇 文献中出现的次数,以此为基础对这些词进行聚类分析,从而反映出这些词之间的亲疏关系,进而分析这些词所代表的学科和主题的结构变化[2]。共词分析主要研究两个问题:一是探求研究领域间的相互关系,另一个是寻找次要的但是在潜在增长的领域[3]。共词分析的主要方法有多维尺度分析、因子分析和聚类分析。本文将分别应用这三种方法进行共词分析,探寻当前我国情报学博硕士学位论文的研究热点。

2 数据收集

本文选取CNKI的《中国博士学位论文全文数据库》和《中国优秀硕士学位论文全文数据库》作为数据来源,这两个数据库可查询到2000年至今的学位

收稿日期:2011-05-04

作者简介:付鑫金(1984-),女,山西太原人,博士研究生,主要从事情报计量学的理论与实践研究.

论文数据。学科专业名称设定为“情报学”，年限划定为2000年至2009年，于2010年5月6日共检索到1501条记录。利用excel对关键词进行抽取，并进行词频统计，共得到2764个关键词。选取其中词频不小于15的关键词为研究对象，如表1所示。

表1 高频关键词列表

序号	关键词	词频	序号	关键词	词频
1	电子商务	67	11	信息资源	20
2	知识管理	63	12	信息构建	19
3	竞争情报	33	13	信息服务	19
4	数据挖掘	31	14	图书馆	19
5	本体	28	15	搜索引擎	18
6	数字图书馆	27	16	客户关系管理	18
7	企业	27	17	信息系统	16
8	电子政务	25	18	知识服务	15
9	信息化	23	19	信息检索	15
10	高校图书馆	22			

3 共词分析

3.1 构造矩阵

首先，构造共词矩阵。两两统计这19个关键词在1501篇文献中共同出现的次数，形成19×19的矩阵，如表2所示。

表2 共词矩阵

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	67	2	0	2	0	1	3	1	4	0	2	1	0	0	0	2	1	0	0
2	2	63	6	2	3	1	4	4	2	2	0	0	1	4	0	3	1	3	0
3	0	6	3	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	2	2	3	31	1	0	0	1	0	0	1	0	0	0	3	4	0	0	2
5	0	3	1	1	28	5	0	1	0	0	0	0	1	0	1	0	1	0	3
6	1	1	1	0	5	27	0	0	0	0	1	2	3	0	0	0	0	2	0
7	3	4	0	0	0	0	27	0	1	0	4	0	1	0	0	0	0	0	0
8	1	4	1	1	1	0	0	25	1	0	3	0	0	0	0	2	0	0	1
9	4	2	0	0	0	0	1	1	23	0	1	0	0	0	0	0	1	0	0
10	0	2	0	0	0	0	0	0	0	22	1	0	2	0	0	0	0	1	0
11	2	0	0	1	0	1	4	3	1	1	20	1		0	0	0	0	0	0
12	1	0	0	0	0	2	0	0	0	0	1	19	1	0	0	0	0	0	0
13	0	1	0	0	1	3	1	0	0	2	0	1	19	2	1	0	1	1	0
14	0	4	0	0	0	0	0	0	0	0	0	0	2	19	0	0	0	1	0
15	0	0	0	3	1	0	0	0	0	0	0	0	1	0	18	0	0	0	4
16	2	3	0	4	0	0	0	2	0	0	0	0	0	0	0	18	0	0	0
17	1	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	16	0	0
18	0	3	0	0	0	2	0	0	0	1	0	0	1	1	0	0	0	15	0
19	0	0	0	2	3	0	0	1	0	0	0	0	0	0	4	0	0	0	15

第二，构造相关矩阵。将共词矩阵转化为相关矩阵，可以实现两个目标。其一是相关矩阵既可以直接观察文献的共引频次，也能够分析对象间的相关度；其二是相关矩阵是对原始矩阵的标准化，消除了那些高被引对象与那些与其相似却很少被引的对象在规模上的差别，进而也就消除文献引用特性不同所造成的数据偏差[4]。

本文采用的相似系数为Ochiia系数：

Ochiai相关系数：$S_{ij}=\frac{C_{ij}}{(C_iC_j)^{\frac{1}{2}}}$

其中，C_{ij}代表A、B两次同时出现的频次，Ci代表A词出现总频次，C_j代表B词出现总频次。

于是得到相关矩阵，见表3，对角线上的数据均为1，表示某词自身的相关程度。

表3 相关矩阵（部分）

	1	2	3	4	5	6	7	8	9	10
1	1.0000	0.0308	0.0000	0.0439	0.0000	0.0235	0.0705	0.0244	0.1019	0.0000
2	0.0308	1.0000	0.1316	0.0453	0.0714	0.0242	0.0970	0.1008	0.0525	0.0537
3	0.0000	0.1316	1.0000	0.0313	0.0329	0.0000	0.0335	0.0000	0.0000	0.0000
4	0.0439	0.0453	0.0313	1.0000	0.0339	0.0000	0.0000	0.0359	0.0000	0.0000
5	0.0000	0.0714	0.0329	0.0339	1.0000	0.1818	0.0000	0.0378	0.0000	0.0000
6	0.0235	0.0242	0.0000	0.0000	0.1818	1.0000	0.0000	0.0000	0.0000	0.0000
7	0.0705	0.0970	0.0335	0.0000	0.0000	0.0000	1.0000	0.0000	0.0401	0.0000
8	0.0244	0.1008	0.0000	0.0359	0.0378	0.0000	0.0000	1.0000	0.0417	0.0000
9	0.1019	0.0525	0.0000	0.0000	0.0000	0.0000	0.0401	0.0417	1.0000	0.0000
10	0.0000	0.0537	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000

第三，转化相异矩阵。为方便进一步处理，用“1”与全部矩阵数据相减，得到表示两词间相异程度的相异矩阵[5]，如表4所示。相异矩阵中的数据，数值越大表明关键词之间的距离越远，相似度越差；反之，数值越小表明关键词之间的距离越近，相似度越大[6]。因而不同的分析方法，选取不同的矩阵结构。

表4 相异矩阵（部分）

	1	2	3	4	5	6	7	8	9	10
1	0.0000	0.9692	1.0000	0.9765	1.0000	0.9765	0.9295	0.9756	0.8981	1.0000
2	0.9692	0.0000	0.8684	0.9547	0.9286	0.9758	0.9030	0.8992	0.9475	0.9463
3	1.0000	0.8684	0.0000	0.9687	0.9671	1.0000	0.9665	1.0000	1.0000	1.0000
4	0.9765	0.9547	0.9687	0.0000	0.9661	1.0000	1.0000	0.9641	1.0000	1.0000
5	1.0000	0.9286	0.9671	0.9661	0.0000	0.8182	1.0000	0.9622	1.0000	1.0000
6	0.9765	0.9758	1.0000	1.0000	0.8182	0.0000	1.0000	1.0000	1.0000	1.0000
7	0.9295	0.9030	0.9665	1.0000	1.0000	1.0000	0.0000	1.0000	0.9599	1.0000
8	0.9756	0.8992	1.0000	0.9641	0.9622	1.0000	1.0000	0.0000	0.9583	1.0000
9	0.8981	0.9475	1.0000	1.0000	1.0000	1.0000	0.9599	0.9583	0.0000	1.0000
10	1.0000	0.9463	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000

3.2 多维尺度分析

首先利用SPSS对相异矩阵（表4数据）进行多维尺度分析，相关可视化结果如图1所示。该方法是通过测定事物或观测量之间的距离来发现数据结构，指定观测量到概念空间的一个特定位置，使得空间中距离的相似性越近越好[6]。该方法比较直观，并且可作为聚类分析、因子分析的辅助技术[4]。

从图1中可看出，有高度相似性的对象聚集在一起，形成一个类别。越在中间的对象与其有联系

的对象也就越多，也就越核心；反之，则越孤独，越在外围。由此，大致可将关键词分为三组。编号6、10、12、13、14、18的关键词聚为一类，即数字图书馆、高校图书馆、信息构建、信息服务、图书馆、知识服务；编号5、15、19的关键词聚为一类，即本体、搜索引擎、信息检索；其余为一类。

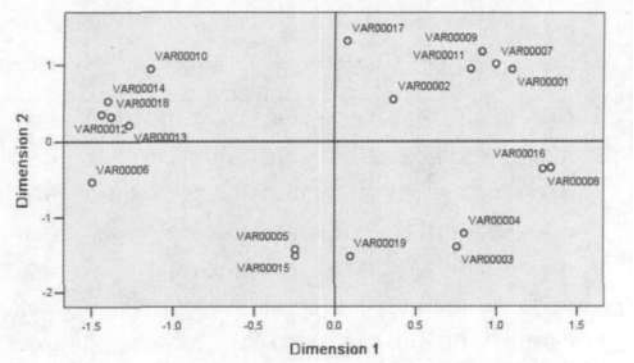

图1 多维尺度分析结果

3.3 聚类分析

利用SPSS对表4数据进行分层聚类分析。方法选择“Between-groups linkage（组间平均链锁距离）”，该方法利用个体与小类的所有距离的信息，克服了极端值造成的影响。将聚类结果绘制为树状图，如图2所示。

从图2中可明显看出，关键词聚为三类。第一类为编号15、19、4、16这四个关键词，即搜索引擎、信息检索、数据挖掘、客户关系管理；第二类的关键词有本体、数字图书馆、信息构建、知识管理、竞争情报、信息服务、图书馆、知识服务、高校图书馆；第三类的关键词有电子商务、信息化、企业、信息资源、电子政务、信息系统。

图2 聚类树状图

3.4 因子分析

因子分析是根据相关性大小把变量分组，同组的变量相关性较高，而不同组的变量相关性较低。因此利用SPSS对相关矩阵，即表3数据，进行因子分析，方法选择“主成分分析”，生成的碎石图如图3所示。图3中横坐标为因子数目，纵坐标为特征根。选择特征根大于1的因子有10个，这10个因子大部分地解释了原有变量。从第11个因子开始，对原有变量的解释贡献较小，如“山脚下的碎石”可忽略不计。因子分析结果，见表5。

从表5中可看出，编号15、19的关键词在第1个因子上有较高载荷；第2个因子主要解释了编号4、16的关键词；第3个因子主要解释了编号5、6这两个关键词；第4个因子主要解释了编号7、11的关键词；第5个因子主要解释了编号3、10、12、13的关键

表5 因子分析结果

	Component									
	1	2	3	4	5	6	7	8	9	10
VAR00001	−0.168	0.149	−0.078	0.099	−0.735	0.176	−0.271	0.028	0.046	−0.034
VAR00002	−0.265	−0.003	−0.012	−0.035	0.084	−0.475	0.132	0.295	0.145	0.223
VAR00003	−0.105	−0.100	−0.054	−0.011	0.156	−0.968	−0.217	−0.153	−0.053	−0.138
VAR00004	0.201	0.769	−0.109	−0.062	0.082	0.035	−0.115	0.087	0.028	−0.077
VAR00005	0.131	−0.116	0.821	−0.072	0.164	−0.048	0.077	0.242	−0.118	−0.086
VAR00006	−0.242	−0.071	0.761	−0.040	0.087	0.175	−0.126	−0.130	0.207	−0.069
VAR00007	−0.111	−0.113	−0.043	0.917	−0.019	−0.088	−0.235	0.276	0.004	0.059
VAR00008	−0.087	0.021	−0.004	−0.037	0.119	0.175	0.937	0.024	−0.019	0.036
VAR00009	−0.034	−0.335	−0.171	−0.131	−0.774	0.102	0.068	0.062	−0.073	−0.016
VAR00010	−0.094	−0.092	−0.290	−0.074	0.260	0.010	−0.040	0.151	0.094	−0.624
VAR00011	−0.104	−0.145	−0.080	0.688	0.069	0.196	0.355	−0.058	0.044	−0.135
VAR00012	−0.134	−0.135	−0.115	−0.206	0.064	−0.115	−0.018	−1.008	0.025	0.054
VAR00013	−0.131	−0.063	0.052	0.041	0.422	0.308	−0.280	−0.005	−0.167	0.157
VAR00014	0.003	−0.136	−0.265	−0.055	0.190	0.149	0.009	0.010	0.074	0.803
VAR00015	0.855	0.018	−0.134	−0.069	0.084	0.106	−0.177	0.041	0.063	0.041
VAR00016	−0.301	0.875	−0.072	−0.173	0.047	0.100	0.150	0.098	−0.031	−0.019
VAR00017	−0.200	−0.104	0.018	−0.204	−0.080	0.060	−0.065	0.146	−0.797	−0.046
VAR00018	−0.080	−0.142	0.064	−0.234	−0.127	0.108	−0.105	0.165	0.638	−0.014
VAR00019	0.861	−0.104	0.091	−0.111	0.068	0.075	0.067	0.108	0.056	0.004

词;第6个因子主要解释了编号1、9的关键词;第7个因子主要解释了编号为8的关键词;第8个因子主要解释了编号2、17的关键词;第9个因子主要解释了编号为18的关键词;第10个因子主要解释了编号14的关键词。

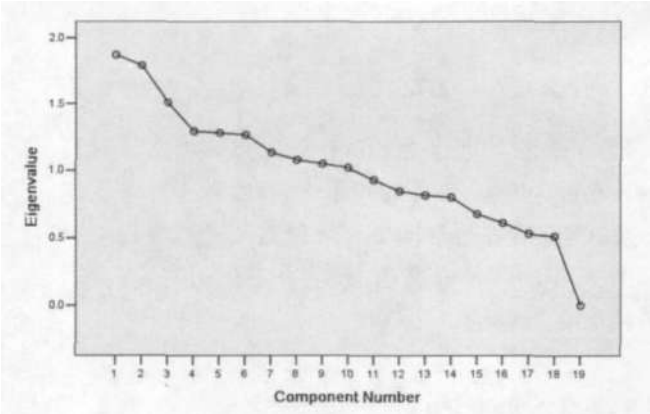

图3 因子碎石图

4 结　语

利用高频关键词构造共词矩阵,并进行多维尺度分析将关键词大类分出,再对矩阵进行聚类分析和因子分析,将关键词大类逐渐细化为小类。以此综合来看其中反映的当前情报学研究现状。

⑴信息检索技术的理论与实践,代表关键词有信息检索、搜索引擎、本体、数据挖掘等。面对"信息爆炸"的热浪,如何才能更快、更准地获取用户想要的信息。为此,情报研究人员更加注重信息检索技术方面的理论与实践,这离不开对搜索引擎、数据挖掘的研究与应用。本体理论的加入,使得搜索更加准确与人性化。数据挖掘、搜索引擎是实施客户关系管理的重要技术,可对客户群体进行划分,帮助网站找到其所关心的客户,如潜在客户、有价值客户、保持客户等。文献[7]也指出,信息检索是情报学最重要的核心领域之一,要不断提高检索的效率和可用性。可以看出,信息的智能与个性化检索、深度挖掘等问题仍是情报学领域的研究热点。

⑵图书馆的建设与服务,代表关键词有图书馆、高校图书馆、数字图书馆、信息构建、信息服务、知识服务等。随着网络时代的到来,如今的图书馆不仅有物理的模式,也出现了数字图书馆。数字化信息资源整合与长期保存是数字图书馆研究的主要方面。不管图书馆的模式是什么,都是要为用户服务。然而用户不再仅仅需要被提供文献、信息,更需要的是解决问题的方法,即知识服务。研究人员应用本体、信息构建的思想来完善数字图书馆的建设,并利用各种技术手段加强资源的共享与服务质量。情报学离不开图书馆这片土壤,如何加强图书馆的建设与服务仍将是情报学领域的热点问题之一。

⑶电子商务、电子政务的信息化,代表关键词有电子商务、电子政务、信息系统、竞争情报、知识管理等。在高频关键词排名中,电子商务名列第一,说明在此方面有相当多的研究。共词分析后,发现电子商务与电子政务、竞争情报、知识管理等词归为一类。这是由于现如今,不管是政府还是企业都相继建设自己的网站,电子政务、电子商务使用户更易获取网站上的信息。尤其对于企业来说,竞争情报与知识管理,一是外炼,一是内修,二者缺一不可。而面对庞大的数据、信息,企业、政府不能只是对信息进行盲目堆砌,而要使信息资源得到更有效地组织与利用,故对其信息系统建设的要求非常高。

本文利用近年来我国情报学博硕士学位论文的高频关键词构造共词矩阵,并进行多维尺度分析、聚类分析、因子分析。通过共词分析挖掘当前情报学的研究现状,并尝试梳理热点内容的结构关系。然而,所选关键词的多少会造成分析结果的不同。研究越深入,某类关键词词频才越高,这需要一定的时间保证。共词分析存在这样的时滞问题,会影响其反映某一研究未来的发展趋势[5]。因此,不排除低频词会成为将来的研究热点。另外,由于一些学位论文涉及保密或其他原因,未被CNKI的数据库所收录,因此统计上难免会有偏差。

参考文献

1 李长玲,翟雪梅. 我国情报学硕士学位论文的共词聚类分析[J]. 情报科学, 2008,(1): 73-76.
2 冯　璐,冷伏海. 共词分析方法理论进展[J]. 中国图书馆学报, 2006,(2): 88-92.
3 崔　雷,郑华川. 关于从MEDLINE数据库中进行知识抽取和挖掘的研究进展[J]. 情报学报, 2003,(4): 425-433.
4 吴　霞, 冷伏海. 基于文献的知识挖掘: 概念、关键技术与应用[A]. 情报学进展(2006-2007年度评论)[C].北京: 国防工业出版社, 2008: 271-306.
5 郑华川,于晓欧,辛　彦. 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.
6 张　勤,马费成. 国外知识管理研究范式[J]. 管理科学学报, 2007,(6): 65-74.
7 赖茂生,王　琳,杨文欣,等. 情报学前沿领域的确定与讨论[J]. 图书情报工作, 2008, (3): 15-18.

(责任编辑:徐　波)