·工作研讨·

# 文献计量及共词分析视角下的国内云计算热点研究

王贵娟,李秀霞,陈 强
(曲阜师范大学信息技术与传播学院,山东 日照 276826)

[摘 要] 对从中国知网(CNKI)期刊全文数据库中提取的云计算领域的文献进行定量分析,选取其中的高频关键词进行共词分析,采用 SPSS 软件进行聚类分析,并将国内有关云计算的研究划分为:云计算的服务层次及安全问题、基于云计算架构的数据中心的存储及安全、基于云计算的高校信息化及信息服务研究。

[关键词] 云计算;文献计量;内容分析;共词分析
[中图分类号] TP393 - 3 [文献标志码] B [文章编号] 1005 - 6041(2012)01 - 0046 - 05

云计算是近年来出现的一个技术名词。很多专家认为,云计算会改变互联网的技术基础,甚至会影响整个产业的格局。正因为如此,很多大型企业都在研究云计算技术和基于云计算的服务。[1] 公共媒体也对云计算表现出极大热情,云计算文献已经出现在 BBC、《经济学家》等重要媒体上。

作为科技研究成果的科技论文,既是科学研究的重要手段,又是科技人员交流学术思想和科研成果的工具。[2] 对某一领域科技论文发表情况进行定量与定性分析,可以间接反映出这一学科的研究成果和水平,为客观公正的评价学科发展提供依据。[3] 为探究我国云计算研究的发展现状及热点趋势,本文综合应用了文献计量和内容分析两种情报学研究方法对相关文献进行定量与定性分析。

## 1 研究方法与数据来源

### 1.1 研究方法

文献计量分析法是以文献信息为研究对象、以文献计量学为理论基础的一种研究方法。内容分析法是对文献内容进行客观、系统和量化描述与分析的研究方法,是社会科学研究中普遍使用的一种科学方法。前者以定量分析为主,侧重外部表征;后者以定性分析为主,侧重内部特征。为了确保研究结论的准确性和可信度,本文将文献计量分析法与内容分析法结合起来对相关文献进行分析。[4] 本文先采用文献计量分析法统计、分析出关于云计算研究的核心期刊,然后对相关文献的关键词进行共词聚类分析,找出知识间的关联及变化趋势。

共词分析是内容分析法中常用的方法。一般以文本中的关键词或主题词为分析单元,词汇对在同一篇文献中出现的次数越多,说明这两个主题的关系越紧密。统计出文献集中关键词或主题词在同一篇文献中两两出现的频率,便可形成一个由这些词对关联所形成的共词网络,网络内节点之间的远近可以反映主题内容的亲疏关系。利用 Ochiia(包容)系数、聚类分析等多种统计分析方法,把词汇之间错综复杂的共词网状关系简化并以数值、图形直观地表现出来。

### 1.2 数据来源

以中国知网(CNKI)期刊全文数据库为目标源,以“云计算”作为检索词,分别采用关键词和主题两种检索途径进行检索。为了保证检索质量,对检索到的文献做了进一步的处理:剔除会议通知、会议报道、刊物征稿等消息类文献;只保留带有关键词的文献,以便后续利用关键词进行主题分析;尽量剔除重复文献和一稿多投文献;去掉无署名的文献。具体检索结果见表 1。

表 1 2007—2010 年“云计算”学术论文发表情况

(单位:篇)

| 检索途径 \ 年度 | 2007 | 2008 | 2009 | 2010 | 合计 |
|---|---|---|---|---|---|
| 关键词 | 0 | 15 | 170 | 602 | 787 |
| 主题 | 6 | 154 | 600 | 1 390 | 2 150 |

对检索到的文献进行分析后发现:我国云计算方面的论文虽在 2007 年开始出现,但对其进行真正的研究始于 2008 年,而且近两年呈剧增的趋势;通过“主题”检索途径检出的论文存在大量不相关的以及消息类文献,其中还包括一些不规范的关键词。基于以上两点,本文最后采用的检索策略为:关键词 = “云计算”(精确匹配),并勾选了“中英文扩展”;检索时间限定在 2008 年 1 月 1 日到 2010 年 12 月 31 日;检索日期为 2011 年 4 月 11 日。

对采用此检索策略检索到的 787 篇文献采用文献计量内容分析法分阶段进行分析。第一阶段:将文献导入 EndNote 软件中,分析文献的年代分布、文

献期刊分布,并确定核心期刊。第二阶段: 对关键词进行共词分析。首先是关键词的预处理。由于所选关键词不是标准的受控词,所以在进行词频统计之前进行了一定程度的人工干预,如将“软件即服务”统一为“SaaS”,‘安全策略’‘安全问题”统一为“安全”等。接着计算关键词出现的频次,并从学科领域知识出发选择频次在 4 以上的 26 个关键词作为代表云计算研究方向的高频词。统计这些词在同一篇文章中两两出现的频次,并制作共词矩阵。最后利用 SPSS 对共词矩阵进行聚类分析。第三阶段: 研究结果的分析和讨论,对聚类结果进行解析。

## 2　云计算研究文献的计量分析

### 2.1 文献年代分布

由表 1 可以清晰地看到关于云计算研究的论文自 2007 年后数量在不断地增长,而且每年的增幅都很大,可谓突飞猛进。这说明云计算的出现迅速引起了国内学者的高度关注。考虑到普赖斯文献指数增长规律以及文献逻辑增长规律: 当一个处于诞生与发展阶段的主题出现时,会引发许多不同思想的交流。学科内容的相互渗透、交叉丰富了云计算的研究内容。

### 2.2 文献期刊分布

据统计,787 篇论文散布于 280 种期刊中。依据布拉德福定律,按期刊的实际载文量,将所有期刊分为 3 个区,依次是核心区、相关区以及边缘区,每个区的论文数量大致相等,大约是 258 篇。由于排名前 14 位的期刊发文总数为 258,所以可以确定排名前 14 位的期刊为云计算研究领域的中文核心区期刊。

## 3　云计算研究的内容分析

### 3.1 文献关键词分析

关键词是对文献内容的深度提炼,能直观、快捷、鲜明地反映文献的主题。同一文献的 3 到 6 个关键词之间存在着一定的内在联系,共同表述所在文献的主题。而同一主题的多篇文献中的关键词相互交叉反映这一主题的主要内容。因此,本文采用词频统计、共词分析、聚类分析对关键词进行分析,以期准确地把握云计算研究的热点、趋势。

我们分 2008 年、2009 年、2010 年三个时间点以及 2008—2010 年整个时间段分别统计文献中出现的关键词的词频,得到关键词词频的排列见表 2( 部分) 。

表 2　关键词词频

| 2008 年关键词 | 频次 | 2009 年关键词 | 频次 | 2010 年关键词 | 频次 | 2008—2010 年关键词 | 频次 |
|---|---|---|---|---|---|---|---|
| 云计算 | 8 | 云计算 | 55 | 云计算 | 197 | 云计算 | 260 |
| 网格技术 | 2 | 网格计算 | 8 | 虚拟化 | 14 | 虚拟化 | 21 |
| ICT | 1 | SaaS | 4 | SaaS | 11 | SaaS | 15 |
| 刀片系统 | 1 | 虚拟化 | 4 | IaaS | 10 | 网格计算 | 13 |
| 教育信息系统 | 1 | PaaS | 3 | PaaS | 9 | PaaS | 12 |
| 绿色创新 | 1 | 云服务 | 3 | 图书馆 | 8 | IaaS | 11 |
| 绿色存储 | 1 | 云存储 | 2 | 云服务 | 7 | 互联网 | 11 |
| 网格计算 | 1 | Google 协作平台 | 2 | 物联网 | 7 | 云服务 | 10 |
| 网络学习 | 1 | Hadoop | 2 | 安全 | 7 | 安全 | 9 |
| 网络应用 | 1 | SOA | 2 | 云安全 | 6 | 图书馆 | 9 |
| 节能减排 | 1 | 云 | 2 | 数字图书馆 | 6 | 数字图书馆 | 8 |
| 虚拟化技术 | 1 | 云计算技术 | 2 | 云存储 | 6 | 云存储 | 8 |
| 计算力 | 1 | 互联网 | 2 | 互联网 | 6 | 云安全 | 7 |
| 软件复用 | 1 | 信息服务 | 2 | 高校 | 6 | 物联网 | 7 |

从表 2 中可以看出,随着时间的变化云计算研究的主题越来越明确。虚拟化一直是云计算研究的热点,2009 年开始对网格计算与虚拟化进行比较的主题则研究得比较多。

选择 2008—2010 年间出现的频次大于 4 的 26 个关键词两两配对,统计他们在文献中出现的次数,

得到一个 26×26 的矩阵见表 3(部分)。

表 3 关键词频次矩阵

| | 虚拟化 | SaaS | 网格计算 | PaaS | IaaS | 互联网 | 云服务 | 安全 | 数字图书馆 | 云存储 | 云安全 | 物联网 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 虚拟化 | 21 | 2 | 4 | 2 | 1 | 3 | 0 | 2 | 1 | 0 | 0 | 0 |
| SaaS | 2 | 15 | 3 | 5 | 3 | 2 | 1 | 2 | 0 | 0 | 0 | 1 |
| 网格计算 | 4 | 3 | 13 | 0 | 0 | 4 | 0 | 0 | 1 | 0 | 0 | 0 |
| PaaS | 2 | 5 | 0 | 12 | 5 | 3 | 0 | 2 | 0 | 0 | 1 | 1 |
| IaaS | 1 | 3 | 0 | 5 | 11 | 3 | 1 | 2 | 0 | 0 | 0 | 1 |
| 互联网 | 3 | 2 | 4 | 3 | 3 | 11 | 1 | 5 | 1 | 1 | 2 | 1 |
| 云服务 | 0 | 1 | 0 | 0 | 1 | 1 | 10 | 1 | 0 | 0 | 0 | 0 |
| 安全 | 2 | 2 | 0 | 2 | 2 | 5 | 1 | 9 | 1 | 0 | 0 | 0 |
| 数字图书馆 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 8 | 0 | 0 | 0 |
| 云存储 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 8 | 1 | 0 |
| 云安全 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 1 | 7 | 0 |
| 物联网 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 7 |

由于频次悬殊会给统计结果造成影响,所以用 Ochiia 相似系数将共词矩阵转换成相关矩阵。相关矩阵表明的是两个关键词之间的相关程度,数值越大,表明关键词之间的相关程度越高,相似度越好。转换的公式为 Nij/sqrt(NiNj)。统计后发现,相关矩阵中 0 值过多,这会导致统计时误差过大。为了后面的进一步处理,用 1 和相关矩阵上的全部数据做相减运算,得到表示两词间相异程度的相异矩阵,见表 4(部分)。

表 4 关键词相异矩阵

| | 虚拟化 | SaaS | 网格计算 | PaaS | IaaSP | 互联网 | 云服务 | 安全 |
|---|---|---|---|---|---|---|---|---|
| 虚拟化 | 0 | 0.887 | 0.757 909 | 0.784 012 | 0.934 205 | 0.802 614 | 1 | 0.854 521 |
| SaaS | 0.887 313 | 0 | 0.785 166 | 0.627 322 | 0.766 450 | 0.844 300 | 0.918 350 | 0.827 867 |
| 网格计算 | 0.757 909 | 0.785 166 | 0 | 1 | 1 | 0.665 503 | 1 | 1 |
| PaaS | 0.874 012 | 0.627 322 | 1 | 0 | 0.564 806 | 0.738 884 | 1 | 0.807 550 |
| IaaS | 0.934 205 | 0.766 450 | 1 | 0.564 806 | 0 | 0.727 273 | 0.904 654 | 0.798 992 |
| 互联网 | 0.802 614 | 0.844 300 | 0.665 503 | 0.738 884 | 0.727 273 | 0 | 0.904 654 | 0.497 481 |
| 云服务 | 1 | 0.918 350 | 1 | 1 | 0.904 654 | 0.904 654 | 0 | 0.89 4591 |
| 安全 | 0.854 521 | 0.827 867 | 1 | 0.807 550 | 0.798 992 | 0.497 481 | 0.894 591 | 0 |

用 SPSS 对所得相异矩阵进行聚类分析,所得见图 1。

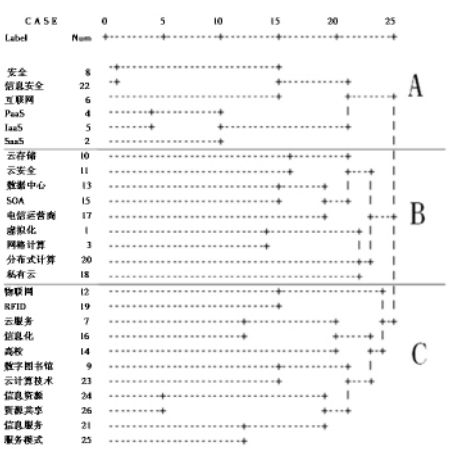


图 1 相异矩阵聚类分析结果树形图

图 1 为将相异矩阵区分为 3 大类进行分层聚类形成的聚类结果树形图。本研究将这 3 个研究主题自上而下分别命名,即 A 类为云计算的服务层次及安全问题,B 类为基于云计算架构的数据中心的存储及安全,C 类为基于云计算的高校信息化及信息服务研究。

### 3.2 文献主题分析

3.2.1 云计算的服务层次及安全问题。在云计算受到众多企业追捧的同时,它所带来的安全问题也引起了业界的重视。Gartner 公司于 2008 年发布了一份关于云计算安全的风险分析,列举了 7 项安全风险,包括特权管理、数据位置、数据隔离、数据恢复、审计与法律调查、服务延续性等。谢四江、冯雁于 2008 年发表的《浅析云计算与信息安全》是国内首

篇介绍云计算与信息安全的学术论文,文章在介绍云计算的相关概念、系统架构和主要形式的基础上,简要分析了云计算给现代信息安全带来的机遇与挑战。[5] 葛慧在2009年发表的《云计算的信息安全》也探讨了由云计算带来的信息安全问题及解决这些问题的方法。[6] 2009年,一贯关注技术进展的大学图书馆学报刊载了胡小菁、范并思的《云计算给图书馆管理带来挑战》,该文介绍了云计算给图书馆管理带来的包括可替代性问题、标准问题、数据安全和保密问题、知识产权问题等挑战,提出了图书馆界应该近一步解决的问题: 云计算的基础理论问题,云计算在图书馆应用的可行性,图书馆云计算政策、标准与协议,基于云计算的图书馆管理体制,云计算案例等。[7] 根据云计算所提供的服务类型,将其划分为3个层次: 应用层、平台层和基础设施层。相应地,各自对应着一个子服务集合: 软件即服务(SaaS, Software as a Service)、平台即服务(PaaS, Platform as a Service)和基础设施即服务(IaaS, Infrastructure as a Service)。这一方面的典型代表作是2009年黎春兰和邓仲华发表的《论云计算的价值》一文,文章通过介绍各主流厂商(Google、微软、IBM等)的云计算的理念及其共同特点,从内外部架构来分析云计算的潜在和现实价值,文章最后还提出了云计算模式所面临的关于安全性、竞争性等的挑战。[8] 2010年范并思在《图书情报工作》上发表的《云计算与图书馆: 为云计算研究辩护》一文探讨了云计算在图书馆中几种可能的应用,包括: 软件即服务、图书馆集成系统、云存贮、平台即服务或基础设施即服务等。[9] 这是将云计算服务类型与图书馆的具体应用结合起来进行研究的比较成功的文章。

3.2.2 基于云计算架构的数据中心的存储及安全。随着"分布式计算""网格计算"和"SOA""虚拟化"等新技术、新理念的进一步发展推动了计算机产业的发展,云计算运动随之产生。作为承载企业、单位业务和应用基础的数据中心,云计算的重要性正在凸显,更多的用户开始接受各大"电信运营商"提供的基于云计算的"数据中心"的"云存储""云安全""私有云"等服务。2009年张敏、陈云海发表的《虚拟化技术在新一代云计算数据中心的应用研究》对虚拟化技术的概念、特性、发展过程和现状以及电信运营商目前的现状和发展瓶颈进行了描述,考察了虚拟化技术给电信运营商带来的新的发展契机,其中包括虚拟化技术的优势、如何应用虚拟化技术建设云计算数据中心开展新型增值业务等,分析了各云计算数据中心运营商的优势和劣势,并建立了一个以电信运营商为核心的生态圈模型。[10] 同年,石屹嵘、段勇的《云计算在电信IT领域的应用探讨》一文重点介绍了云计算的演进过程和相关概念,阐述了云计算在电信IT领域的应用模式,分析了云计算的架构,并对电信内部数据中心初步实现云计算的过程进行了简要的分析。[11]

3.2.3 基于云计算的高校信息化及信息服务研究。云计算技术的不断成熟以及基于RFID技术的物联网的快速发展使云服务成为了可能,并加快了高校的信息化进程,改变了信息资源的提供方式和信息服务模式。随着数字图书馆信息资源的日益增多,传统的信息服务模式逐渐难以满足知识经济的发展和知识创新的需求。信息服务模式的改变是图书馆服务发展的必然趋势,近年来信息服务及涉及的知识整合、组织、处理、检索、存储等受到越来越多的关注。2009年钱杨等人在《面向信息资源管理的云计算性能分析》一文中介绍了信息资源管理的中心内容,从信息资源管理的各个视角看云的性能要求,从信息交流的障碍出发阐述了云的安全性问题,并且将信息资源管理中信息组织和服务的相关原则应用到云的信息交流、信息组织、信息服务等方面的质量考察中,试图为云计算的性能评价找到合适的标准。[12] 同年,基于企业信息服务对社会性信息技术大平台要求的背景下李勇发表了《云计算对信息服务的影响及存在的问题》,对云计算的基本概念、类型和原理进行了探讨,介绍并分析了Google、微软、IBM、亚马逊等公司的云计算产品,总结了云计算对于企业信息服务的平台支持作用,同时也指出了当前云计算产品的问题和不足。[13]

## 4　总　结

本研究基于文献计量法和内容分析法,利用词频分析、共词分析、聚类分析并结合相关论文,较真实、客观地总结了近几年我国云计算研究的热点。但这种分析也存在一定的局限和不足,首先是数据来源问题。云计算目前在国内还处于起步阶段,大部分研究论文主要是对云计算概念、特点及其应用的研究,真正有影响力的成果还很少。在这种情况下,利用词频统计和高频词共词分析对相关数据进行处理,其结果与实际情况会有一定出入。这一点从本文云计算研究主题的分析可以看出来。其次是共词分析虽然是内容分析法中较常用的方法,但是在揭示信息内容方面还不够全面,可以考虑基于主题聚类的综合研究方法并结合作者聚类分析、机构聚类分析、主题－作者映射分析、主题－机构映射分析等来完善该实证研究。

[参考文献]

[1] 云计算[EB/OL]. [2010 - 04 - 20]. http://baike.baidu.com/view/1316082.htm.

[2] 科技论文[EB/OL]. [2010 - 04 - 20]. http://baike.baidu.com/view/4286893.htm#sub4286893.

[3] 曹冰凌,郑 瑜,王小雄. 我国数字图书馆安全问题研究综述[J]. 江西农业大学学报:社会科学版,2009(4):165—167.

[4] 邱均平,王曰芬. 文献计量内容分析法[M]. 北京:国家图书馆出版社,2008:1—11.

[5] 谢四江,冯 雁. 浅析云计算与信息安全[J]. 北京电子科技学院学报,2008(12):1—3.

[6] 葛 慧. 云计算的信息安全[J]. 信息科学,2009(4):42—43.

[7] 胡小菁,范并思. 云计算给图书馆管理带来挑战[J]. 大学图书馆学报,2009(4):7—12.

[8] 黎春兰,邓仲华. 论云计算的价值[J]. 图书与情报,2009(4):42—46.

[9] 范并思. 云计算与图书馆:为云计算研究辩护[J]. 图书情报工作,2010(21):5—9.

[10] 张 敏,陈云海. 虚拟化技术在新一代云计算数据中心的应用研究[J]. 广东通信技术,2009(5):35—39.

[11] 石屹嵘,段 勇. 云计算在电信 IT 领域的应用探讨[J]. 电信科学,2009(9):24—28.

[12] 钱 杨,代 君,廖小艳. 面向信息资源管理的云计算性能分析[J]. 图书与情报,2009(4):53—56.

[13] 李 勇. 云计算对信息服务的影响及存在的问题[J]. 情报理论与实践,2009(12):89—91,120.

[收稿日期] 2011 - 08 - 29

[作者简介] 王贵娟(1987—),女,曲阜师范大学信息技术与传播学院 09 级在读硕士,主要从事基于云计算的图书馆 IT 管理与服务的研究;李秀霞(1971—),女,硕士,副教授,主要从事实用电子系统设计、图书馆智能管理系统设计等方面的研究工作;陈强(1985—),男,曲阜师范大学信息技术与传播学院 09 级在读硕士,主要从事中小学生知识负荷的研究。

(上接第 37 页)

5.3 “童孙未解供耕织,也傍桑阴学种瓜。”抄本作“学卖瓜”

应依据抄本订正。“童孙未解供耕织,也傍桑阴学种瓜”是著名诗句,无人置疑。但细品原诗,可知此诗句写的是五月的田园生活。一般而言,“五月,瓜便熟”(元司农司《农桑辑要》卷五)。五月不是种瓜而是收瓜时节了。梅尧臣《五月七日见卖瓠者》(《全宋诗》5/3041)诗句谓“四月彼种瓜,五月此卖瓠。”五月卖瓜瓠了。据此理推断,范成大诗句应为“也傍桑阴学卖瓜”——在大人收瓜的时候,儿童们在树阴下玩耍,欢快地叫喊卖瓜,其情景因合时宜而美。如果此时他们“学种瓜”,那就不合时宜了。

5.4 “虫丝罥尽黄葵叶,寂历高花侧晚风。”抄本作“虫丝网尽黄葵叶,寂寞高花侧晚风。”

应依据抄本订正。“虫丝罥尽黄葵叶”中的“罥”字,在《汉语大字典》中列两个义项,一为捕取鸟兽的网,二为用绳索系取鸟兽。前者名词,后者动词。“罥”是人做的,作用对象是鸟兽,它的网眼较大。“虫丝”的网,其作用对象主要是小昆虫。黄葵叶上布满了虫丝网,是“寂寞”的意象。“罥尽”当为“网尽”之误。“寂历”与“寂寞”皆通,但后者更直白,更符合田园诗通俗易懂的风格。

5.5 “不知新滴堪篘未?今岁重阳有菊花。”抄本作“新酒”。

应依据抄本订正。“新滴”,意为新酒,《全宋诗》中仅见此诗句。《全宋诗》中,诗句有“新酒”者很多,范成大的诗中有 3 次提及“新酒”,如《春日杂兴十二首》(《全宋诗》41/25661):“见说市楼新酒美,杖头今日一钱无。”因此可以肯定,“新滴”应校正为“新酒”。这里的“新酒”,即前诗句中的“瓦盆新熟土茅柴”。

## 6 结语

以上从书法墨迹中辑录出宋诗 14 首,整合 1 首,校勘 5 首,这对于订补《全宋诗》是很有价值的,对宋代文学的研究也有意义。这些足以说明,书法墨迹具有独特的文献版本价值,堪称为“书帖本”,可以称为版本体系中的奇葩。历来的文献整理者或多或少注意在书法墨迹中征引资料,但没有把它作为一种版本来看待,也就是重视不够。通过以上的考察,笔者认为,可以确立书法墨迹在文献整理中的版本地位,有理由给予足够的重视了。

[参考文献]

[1] 谢稚柳. 中国历代法书墨迹大观·第七册[M]. 上海,上海书店出版社,1992.

[2] 启 功、王靖宪. 中国法帖全集·第八册[M]. 武汉,湖北美术出版社,2002.

[3] 赵孟坚. 赵孟坚自书诗稿[M]. 沈阳:辽宁美术出版社,2001.

[4] 谢稚柳. 中国历代书法墨迹大观·第八册[M],上海,上海书店出版社,1987.

[5] 梁诗正. 三希堂法帖[M]·第十七册,拓印本.

[6] 谢稚柳. 中国历代书法墨迹大观·第九册[M],上海,上海书店出版社,1987.

[收稿日期] 2011 - 06 - 04

[作者简介] 黄权才(1958—),男,研究馆员,文献学硕士生导师,广西师范学院图书馆。