

doi:10.3772/j.issn.2095-915x.2016.02.007

文献计量分析的我国语义网研究综述

王珊珊, 肖明

(北京师范大学政府管理学院 北京 100875)

摘要: 采用文献计量分析方法, 利用 EXCEL、BICOMB、UCINET 等工具软件, 对 CNKI 中国学术期刊网络出版总库中收录的我国学术期刊在 2001 年 1 月至 2015 年 9 月期间发表的语义网研究论文进行统计分析, 包括时序分析、基金资助分析、期刊来源分析、研究机构分析、关键词分析等, 旨在了解我国语义网研究的最新现状, 为进一步研究提供理论参考。论文还指出, Web 服务、数据关联、RDF、数字图书馆、大数据、语义检索将会是现在及未来的研究热点。

关键词: 语义网, 文献计量分析, CNKI, 文献综述

Review on Semantic Web Research in China Based on Bibliometrics Analyzing Method

WANG ShanShan, XIAO Ming

(School of Government, Beijing Normal University, Beijing 100875, China)

Abstract: This paper employed the bibliometrics analyzing method to analyze the academic articles in China Academic Journal Network Publishing Database(CNKI), which are related to the semantic web research during January 2001 to September 2015. The analysis included time series analysis, fund projects analysis, periodical sources analysis, research institutions analysis, and keywords analysis. Through the analysis, this study is aim

基金项目: 本文系高等学校计算机教育研究 2015 年度项目“国内外计算机网络课程教育调查与分析(项目编号: ER2015007)”
国家社科基金一般项目“基于多方法的 LIS 知识图谱实证研究(项目编号: 11BTQ019)”的中期研究成果。

作者简介: 王珊珊 女, 北京师范大学政府管理学院, 硕士研究生。E-mail: shanshanwang@mail.bnu.edu.cn。肖明, 男, 北京师范大学政府管理学院, 博士, 教授。E-mail: ming_xiao@bnu.edu.cn。

to understand the status of the latest research on semantic web in China and provide theoretical reference for the further research. It also proposes that web service, data association, RDF, digital library, big data, semantic retrieval might be the research hotspots for current stage and future.

Key words: Semantic Web, bibliometrics analysis, CNKI, literature review

1 引言

2001年,万维网之父——Tim Berners-Lee首次提出语义网(Semantic Web)的理论框架,并将其作为下一代互联网的发展趋势^[1]。目前,我国针对语义网的研究主要集中在计算机和图书情报两大领域。

在计算机领域,针对语义网的研究大体上可以进一步细分成以下两大类:(1)理论研究。例如,杜小勇等人对语义Web领域中的研究热点、Ontology研究状况、本体的定义和描述语言、建设方法和工具以及主要研究机构等进行了全面介绍,为后来者研究语义网提供了帮助^[2];龚洪泉等人通过讨论Semantic Web的层次框架模型,指出了各个层次扮演的角色,并着重分析了Semantic Web的重要研究领域^[3];崔华等人概述了语义Web服务组合的研究内容和目标,根据语义Web服务组合中使用的方法学对其进行分类,并且分析了方法的实现过程和特点^[4]。(2)应用研究。例如,唐杰等人基于贝叶斯决策理论,提出了风险最小化的本体映射方法Ri-MOM,与同类方法相比它具有更高的查准率和查全率^[5];孙萍等人采用Petri网作为Web服务过程描述的形式化工具,提出了一个新的语义Web服务发现框架,提高了服务的发现效率和查找的精确度^[6];金燕等人提出了一个基于推理的语义网检索模型,并且介绍了实现该模型的关键技术,该模型能够提

高信息检索的语义性,并且取得了满意的信息检索结果^[7]。

在图书情报领域,目前已有部分学者基于期刊论文数据库对我国语义网研究论文进行了相关分析。例如,南京大学的陆佳莹等人基于CNKI全文数据库,对我国图书情报领域的语义网研究进程和研究热点进行了分析和总结,得出语义网的关键技术和研究核心,为相关领域科研人员提供参^[8];王俊红对2001年至2011年期间我国语义Web研究论文的时间分布、地区分布、期刊分布、著者分布、研究领域分布等进行了统计分析,总结了我国语义Web的研究现状^[9]。

尽管现有的语义网研究成果可以在一定程度上帮助后学者了解我国语义网的研究现状,但它们大多是基于学术经验定性总结的经验之谈,只有少数论文采用定量方法进行分析。另外一个重要原因,语义网研究近年来在我国发展十分迅速,研究热点不断更新,应用研究不断细化。因此,非常有必要系统地分析我国语义网近年来的研究现状,以方便语义网领域的专家和学者对此进行更深入的研究。

2 数据来源与研究方法

2.1 数据来源

本文采用的数据来源于中国知网的中国学术

期刊网络出版总库,主要检索策略是:主题=“语义网” or 主题=“语义Web” or 主题=“Semantic Web”,检索时间为2015年10月1日。最初得到的检索结果为3521篇学术论文,经过去重、筛选等处理之后,最终得到3152篇适用的学术论文。由于本文只是基于中国知网的学术期刊网络出版总库进行检索,因此国外发表的文献以及国内其他数据库收录的文献未在分析范围之列。

2.2 研究方法

本文主要利用相关数理统计方法,采用共词分析法等文献计量分析方法,以上述3152篇论文作为数据分析样本,统计分析工具主要选用了以下三种:微软公司开发的EXCEL软件、中国医科大学开发的书目共现分析系统BICOMB以及社会网络分析工具UCINET。首先,抽取相关条

目,形成共现矩阵。然后,对我国语义网研究论文进行定量分析,主要包括:时序分析、基金资助分析、期刊来源分析、研究机构分析、关键词分析等。

3 我国语义网期刊研究论文概述

3.1 期刊论文时序分析

研究论文发表的数量在一定程度上可以反映该领域的研究状况、研究水平以及发展速度。截止到2015年9月,我国共计发表了3152篇有关语义网的研究论文,如果按每年的发文量进行分年汇总,则可得到我国语义网研究论文的增长时间序列图(如图1所示)。

从图1可以看出,我国语义网研究大体上可



图1 我国语义网研究论文的增长时间序列图

以细分为以下四个阶段。

(1) 早期萌芽阶段(2001年-2003年)

2001年,万维网之父——Tim Berners-Lee首次提出语义网(Semantic Web)的理论框架,并将其作为下一代互联网的发展趋势。同年,我国聂培尧教授发表题目为“XML及语义Web技术”^[10]的论文。该文是笔者在中国学术期刊网络

出版总库检出的以“语义网”作为主题的第一篇文献。聂培尧教授在这篇论文中对语义网进行了简要介绍,并且描述了XML和RDF,但仅停留在介绍层面上,未进行更深入的研究。在早期萌芽阶段(2001年-2003年),我国学者只是对语义网进行理论介绍,但为此后的应用研究奠定了坚实的基础。

(2) 快速发展阶段(2004年-2006年)

在快速发展阶段,论文数量呈现快速持续增长趋势,研究范围从理论研究拓展到应用研究,语义网在Web服务、P2P网络、知识管理、E-learning、智能信息检索、语义Web挖掘、网格计算等多个领域的应用研究不断增加。2006年,亚洲语义网大会在北京举行,掀起了我国语义网研究热潮,该年度的语义网研究论文总数为289篇,较2005年增长了134篇。

(3) 稳定发展阶段(2007年-2010年)

在稳定发展阶段,我国语义网研究的论文数量增长速度放缓,语义网研究进入稳定发展期。在此阶段,有关语义Web服务的研究论文数量大幅度增加,对服务发现、服务匹配、服务组合、服务基础技术、服务知识模型等的研究热度不断增加。

(4) 应用拓展阶段(2011年-2015年)

在应用拓展阶段,尽管每年发表的语义网研究论文数量略有下降,但语义网的应用领域却在不断扩大,语义网的理论研究越来越细化。例如,知识管理、语义检索、P2P、电子商务、电子政务、关联数据、数字图书馆、大数据、资源描述框架、领域本体等领域的应用研究论文数量在不断增加。

3.2 期刊论文基金资助分析

通过对我国语义网研究论文的基金资助情况(如表1所示)进行分析后发现,该领域发表的论文获得国家自然科学基金资助的有829篇(占论文总数的48%),排名第一;获得省级自然科学基金资助的有364篇(占论文总数的21%),排名第二;获国家高技术研究发展计划(863计划)、国家重点基础研究发展计划(973计划)、国家社会科学基金、国家科技支撑计划资助的研究论文数量紧随其后。由此可以看出,无论是国家还是省

市(自治区)均对语义网研究提供了较大的支持力度,对语义网研究给予了高度关注。

表1 我国语义网研究论文的基金资助情况

排名	基金来源	论文篇数	所占比例(%)
1	国家自然科学基金	894	49%
2	省级自然科学基金	364	20%
3	国家863计划	258	14%
4	国家973计划	135	7%
5	国家社会科学基金	101	6%
6	国家科技支撑计划	74	4%

3.3 期刊来源分析

目前,我国语义网研究主要集中在计算机科学和图书情报两大领域。因此,语义网研究相关论文也大多发表在计算机、图书情报这两大领域的专业期刊上。表2中显示的是发文量排名前15的期刊概况。其中,10种期刊属于计算机领域,5种期刊属于图书情报领域。由此可以推断出,《计

表2 发表论文数量排名前15的期刊

排名	期刊	论文篇数
1	计算机科学	102
2	计算机工程	98
3	计算机工程与应用	71
4	图书情报工作	68
5	计算机工程与设计	67
6	计算机技术与发展	66
7	情报杂志	58
8	计算机应用研究	55
9	现代图书情报技术	54
10	计算机应用	51
11	计算机应用与软件	48
12	情报科学	45
13	情报理论与实践	41
14	电脑知识与技术	40
15	计算机研究与发展	36

计算机科学》、《计算机工程》、《计算机工程与应用》、《计算机工程与设计》、《计算机技术与应用》、《计算机应用研究》这6种期刊是计算机领域语义网研究的核心期刊；《图书情报工作》、《情报杂志》、《现代图书情报技术》、《情报科学》、《情报理论与实践》这5种期刊是图书情报领域语义网研究的核心期刊。这就为后学者研究该课题或是发表相关论文提供了指南，为文献收集管理者提供了参考依据。语义网作为互联网的扩展形式，已经受到计算机和图书情报两大领域学者们的高度重视。

通过对比分析以后，笔者发现：在计算机领域，语义网研究主要关注的是语义网的关键技术、算法创新以及系统设计；而在图书情报领域，语义网研究更多关注的是语义网的应用研究，该领域学者致力于将语义网相关技术应用到图书馆，以便对其服务进行改进。

高被引文献通常被视为某一研究领域的经典文献。在计算机科学领域的高被引文献中，《基于模糊多属性决策理论的语义Web服务组合算法》一文提出一种基于模糊多属性决策理论的语义Web服务组合的优化选择算法，该算法能够评价以实数、区间数和语言型数据描述的QoS信息，从而进行综合决策，具有优越性和有效性^[11]；《基于本体概念相似度的语义Web服务匹配算法》一文提出了一种基于该相似度的Web服务的精确匹配算法，该算法与经典的OWL-S/UDDI匹配算法比较，不仅在等级上保持一致，而且使同一等级或不同等级之间的服务匹配都达到精确的程度^[12]。图书情报领域主要侧重于基于语义网的数字图书馆关键技术应用研究。在图书情报领域被引频次较高的语义网研究论文中，《基于语义Web和Jena插件的语义检索系统实验研究》一文中利用语义Web及知识本体的相关标准和软件工具，介绍和分析了Jena插件的调用接

口和处理机理，在此基础上开发了一个语义检索的实验系统^[13]；《基于语义网的网络信息检索相关性研究》一文中对Mizzaro相关性理论以及语义网理论进行了介绍，并且提出一种语义信息标引方法和语义查询扩展方法来改善查全率和查准率，还通过一个计算机科学领域本体实例来详细阐述其实现过程^[14]；《基于语义网的e-knowledge组织框架与内容研究》一文中则探析了基于语义网的e-knowledge组织框架和内容^[15]。

4 我国语义网研究机构及合作网络分析

4.1 语义网研究机构分析

从2001年以来，国内先后有40个研究机构发表了多于30篇的语义网论文。表3中显示的是发表语义网论文数前10的研究机构概况。其中，中国科学院发表的语义网研究论文高达144篇，位列第一；武汉大学、吉林大学、浙江大学、东南大学紧随其后，由于本文只是基于中国知网的中国学术期刊网络出版总库进行检索，因此国外发表的文献以及国内其他数据库收录的文献未包含在内。

表3 发表论文数量排名前10的研究机构

排名	单位	论文篇数
1	中国科学院	144
2	武汉大学	129
3	吉林大学	83
4	浙江大学	73
5	东南大学	60
6	大连海事大学	54
7	清华大学	53
8	南京大学	50
9	上海交通大学	50
10	同济大学	47

从表3可知,中国科学院、武汉大学、吉林大学、浙江大学、东南大学是国内语义网研究的重要机构,值得学者们高度关注。笔者对武汉大学所发表的语义网研究论文进行深入分析后发现,武汉大学的学者们在2006年以前主要集中在本体研究上。例如,李智等通过研究RDF和RDFS以及语义Web对信息语义的组织方式,阐述了基于OWL DL在语义Web中建立本体模型方法,在此基础上探讨了本体语义模型的结构,并且搭建了本体语义模型的框架^[16];虞为等人提出了一种针对语义网上的本体进行检索和排序的新方法ARRO,解决了在传统的基于关键字的信息检索中只能从句法上对关键字进行分析,无法将推理和检索相结合,互相促进的问题^[17]。2007年以后,学者们针对Web服务的研究更加深

入。例如,冯在文等人提出一种基于情境感知和推理的Web服务发现方法,实现了对语义Web服务查询结果的精化和优化^[18];吴金红等人针对OWL-S对服务质量描述的不足,提出了有助于提供Web服务操作的自动化程度的语义Web服务质量描述框架^[19]。

4.2 语义网研究合作网络

研究合作网络通常是某一特定领域理论发展过程中的一大重要影响因素。目前,国内已有许多研究机构开展了语义网研究合作。笔者首先选取的是发文数在10以及以上的研究机构,然后分析它们之间的合作关系,最终得到我国语义网研究机构合作网络图(如图2所示)。其中,节点的大小代表研究机构论文数量的多少。

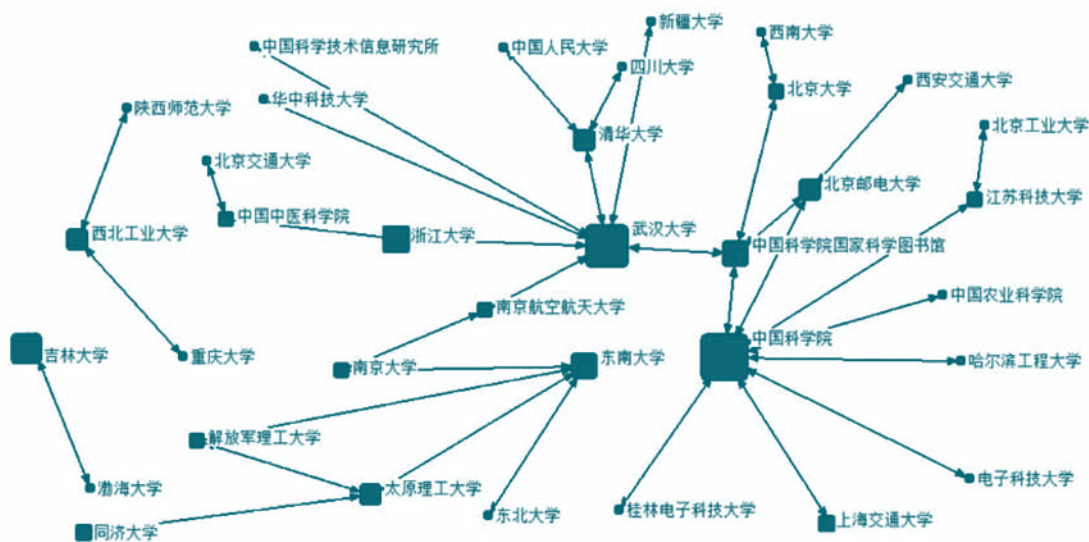


图2 研究机构合作网络图

从图2可以看出,在语义网研究领域比较活跃的研究机构中,中国科学院和武汉大学处于合作网络的核心,它们与清华大学、北京邮电大学、浙江大学、北京大学等都有良好的合作关系。此外,东南大学和南京大学、太原理工大学,西北工业大学、重庆大学和陕西师范大学之间也都有良好的合作关系。

笔者又对武汉大学发表的129篇语义网研究论文进行深入分析后发现,其中研究本体的文章有27篇,研究Web服务的文章有20篇,研究关联数据、馆藏资源的文献各5篇。特别值得一提的是,武汉大学的崔华等人对语义Web服务相关的基本概念进行了归纳和总结,概述了语义Web服务组合的研究内容和目标,根据语义Web服务组合中使

用的方法学对其进行分类,并且分析了这些方法的实现过程和特点,指出了下一步研究方向^[20]。此外,武汉大学的常亮和中科院的史忠植等人提出了一种基于DDL(X)的语义Web服务自动组合方法,该方法充分发挥了DDL(X)在表达能力和计算性能等方面的优势,为语义Web服务自动组合提供了一套行之有效的理论工具^[21]。

5 我国语义网研究的主题分析

5.1 高频关键词分析

(1) 排名前10的高频关键词

关键词是论文核心内容的浓缩和精炼。通过对论文中的关键词进行分析,可以发现该领域的研究热点和发展动向。

笔者对语义网研究论文中的关键词进行了统计和词频分析,通过合并同义词等数据清洗方法建立了关键词库,获得排名前20的高频关键词(如表4所示)。

表4 部分关键词及其词频

排名	关键词	词频	排名	关键词	词频
1	语义网	1545	11	关联数据	86
2	本体	1073	12	领域本体	81
3	Web 服务	494	13	数字图书馆	77
4	服务发现	135	14	资源描述框架	70
5	语义检索	134	15	语义推理	66
6	OWL	122	16	本体映射	66
7	RDF	120	17	元数据	66
8	描述逻辑	101	18	服务匹配	62
9	OWL-S	98	19	服务组合	53
10	XML	87	20	语义网格	50

从表4中可以看出,除了与检索直接相关的关键词以外,本体、Web服务、服务发现、语义检索、OWL、RDF、描述逻辑等均是语义网中的重要研究领域。

为了进一步分析最新研究热点,笔者对2013年至2015年的关键词词频进行了统计,词频超过12的关键词共计有12个(如表5所示)。据此可知,近三年我国的语义网研究热点主要集中在本体、Web服务、数据关联、RDF、数字图书馆、大数据、语义检索上。其中,针对Web服务的研究主要集中在解决Web服务中的发现、组合和执行等问题上。当前,针对语义Web服务发现的研究主要集中在服务描述和服务匹配上;针对语义Web服务组合的研究,工业界侧重于业务过程的研究,通过构建描述Web服务组合的语言,开发相关编辑工具和执行引擎,实现交互式的服务组合系统;学术界则从人工智能规划和形式化方法等方面研究以语义推理为核心的Web服务的自动组合^[20]。

表5 2013年至2015年词频超过12的关键词

排名	关键词	词频	排名	关键词	词频
1	语义网	241	7	大数据	15
2	本体	139	8	语义检索	13
3	Web 服务	57	9	OWL	13
4	关联数据	56	10	资源描述框架	13
5	RDF	20	11	领域本体	13
6	数字图书馆	17	12	描述逻辑	13

(2) 高频关键词年度演化分析

笔者还对排名靠前的高频关键词进行了年度演化分析。在2003年以前,我国语义网研究人员对语义网各个分枝领域的研究相对较少(如图3所示)。

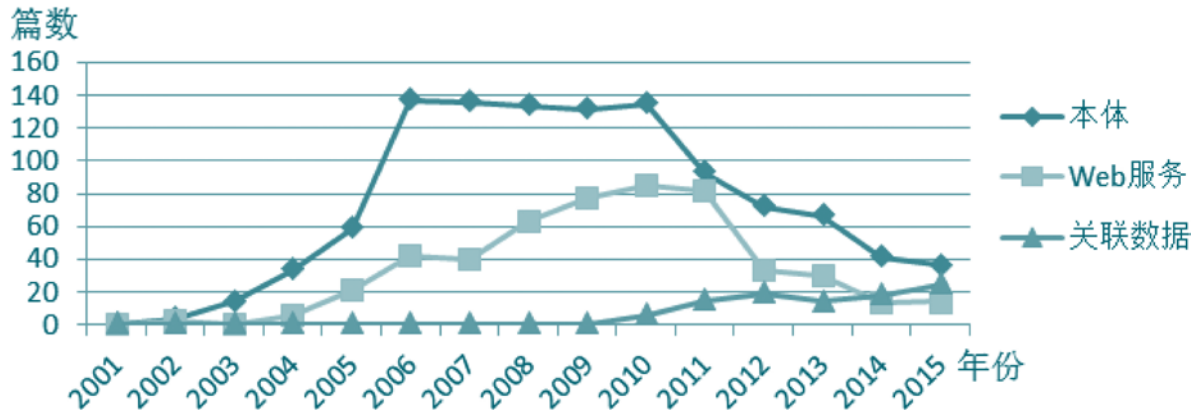


图3 主要关键词的演化

2003年以后，针对本体和Web服务的研究论文数量快速增长。2006年以本体为关键词的论文达到峰值，并在2006年至2010年保持相对稳定；2010年以Web服务为关键词的论文达到峰值，此后对该领域的关注度逐年降低。特别值得一提的是，我国针对关联数据的研究论文直到2010年才开始出现，此后关联数据研究论文快速增长，热度没有丝毫减弱，说明它是语义网研究中的一个新兴研究热点，值得关注。关联数据的概念由Tim Berners-Lee于2006年提出，我国有多位来自图书情报领域的学者对此进行了专题研究。例

如，张晓林等人对关联数据的核心技术进行了分析，并且介绍了其典型应用，指明了该技术的应用前景^[22]；刘炜等人从概念、技术、应用前景等方面对关联数据进行了全面分析，并且呼吁学界重视关联数据技术，投入人力和资源进行研究开发和应用推广^[23]。

其他次要关键词的演化概况如图4所示。从总体上来看，2010年以后，我国学者对服务发现、语义检索、OWL、描述逻辑、OWL-S的研究热度均有所下降。

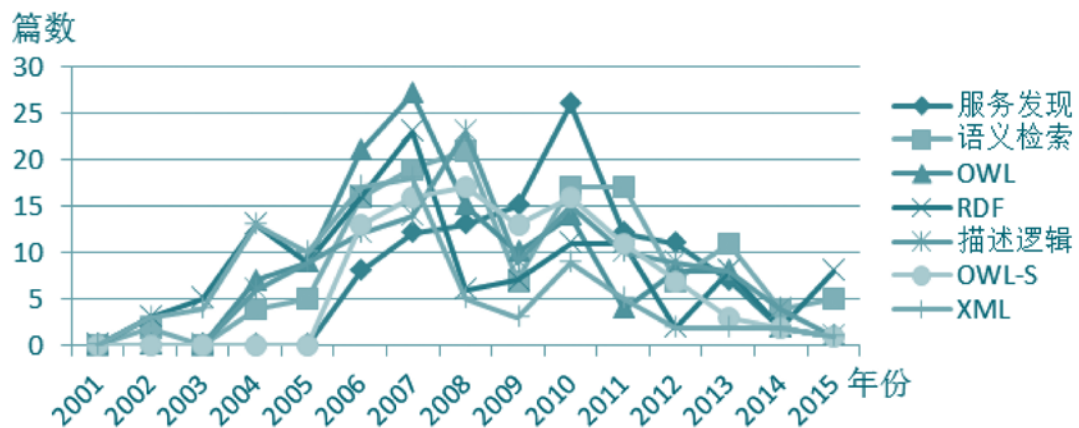


图4 次要关键词的演化

5.2 关键词共现关系网络

为了分析语义网的重要研究主题，需要构建

语义网关键词共现网络图。笔者先将阈值设定为27，剔除频率小于27的关键词，共计得到33个关键词。然后，利用BICOMB工具生成33×33的高

构进行交流。因此,今后有必要加强跨机构、跨学科、跨地区的语义网合作与交流。

(5)从研究热度来看,国内学者对于服务发现、语义检索、OWL、描述逻辑、OWL-S的研究热度均有所下降,语义网应用研究正在不断细分和拓展。近年来,关联数据方面的研究论文数量增长迅速,研究热度没有丝毫减弱,说明它是语义网研究中的一个新兴研究热点,值得关注。

(6)从研究趋势来看,有关本体、Web服务、数据关联、RDF、数字图书馆、大数据、语义检索的研究将会是语义网领域现在及未来研究热点。

参考文献

- [1] Berners-Lee T, Hendler J, Lassila O. THE SEMANTIC WEB.[J]. Scientific American, 2009, 284(5):34-43.
- [2] 杜小勇,李曼,王大治.语义 Web 与本体研究综述[J]. 计算机应用, 2004, 24(10):14-16.
- [3] 龚洪泉,张敬周,钱乐秋,等. Semantic Web 研究综述[J]. 计算机应用与软件, 2005, 22(2):1-6.
- [4] 崔华,应时,袁文杰,等. 语义 Web 服务组合综述[J]. 计算机科学, 2010, 37(5):21-25.
- [5] 唐杰,梁邦勇,李涓子,等. 语义 Web 中的本体自动映射[J]. 计算机学报, 2006, 29(11):1956-1976.
- [6] 孙萍,蒋昌俊. 利用服务聚类优化面向过程模型的语义 Web 服务发现[J]. 计算机学报, 2008, 31(8):1340-1353.
- [7] 金燕,王志华. 基于推理的语义网检索模型及关键技术研究[J]. 计算机工程与设计, 2013, 34(7):2585-2589.
- [8] 陆佳莹,赵宇翔. 我国图情领域语义网的研究现状和热点分析[J]. 情报杂志, 2014(8):99-104.
- [9] 王俊红. 基于 CNKI 文献计量分析我国语义 Web

的研究[J]. 福建电脑, 2012, 28(10):76-77.

- [10] 聂培尧,安世虎. XML 及语义 Web 技术[J]. 计算机科学, 2001, 28(5):34-36.
- [11] 李祯,杨放春,苏森. 基于模糊多属性决策理论的语义 Web 服务组合算法[J]. 软件学报, 2009, 20(3):583-596.
- [12] 彭晖,史忠植,邱莉榕,等. 基于本体概念相似度的语义 Web 服务匹配算法[J]. 计算机工程, 2008, 34(15):51-53.
- [13] 颜端武,丁晟春,李岳蒙,等. 基于语义 Web 和 Jena 插件的语义检索系统实验研究[J]. 情报理论与实践, 2006, 29(3):349-352.
- [14] 何绍华,宫兆晖. 基于语义网的网络信息检索相关性研究[J]. 情报杂志, 2007, 26(12):120-123.
- [15] 田倩,肖红琳. 基于语义网的 e-knowledge 组织框架与内容研究[J]. 图书情报工作, 2010, 54(2):11-15.
- [16] 李智,唐胜群,杨青. 语义网中基于 OWL DL 的本体模型研究[J]. 武汉大学学报:理学版, 2005(S2):163-166.
- [17] 虞为,陈俊鹏,曹加恒. 一种对语义网上本体进行检索和排序新方法[J]. 小型微型计算机系统, 2007, 28(6):1044-1048.
- [18] 冯在文,何克清,李兵,等. 一种基于情境推理的语义 Web 服务发现方法[J]. 计算机学报, 2008, 31(8):1354-1363.
- [19] 吴金红,殷之明,王翠波. 基于 OWL-S 的语义 Web 服务质量描述框架[J]. 情报杂志, 2007, 26(10):75-77.
- [20] 崔华,应时,袁文杰,等. 语义 Web 服务组合综述[J]. 计算机科学, 2010, 37(5):21-25.
- [21] 常亮,刘进,古天龙,等. 基于动态描述逻辑的语义 Web 服务组合[J]. 计算机学报, 2013, 36(12):2468-2478.
- [22] 沈志宏,张晓林. 关联数据及其应用现状综述[J]. 现代图书情报技术, 2010(11):1-9.
- [23] 刘炜. 关联数据:概念、技术及应用展望[J]. 大学图书馆学报, 2011, 29(2):5-12.