

基于计量分析的馆藏资源语义化理论研究^{*}

邱均平 余 凡

摘 要 目前的馆藏资源语义化方法存在一定的问题。本文引入信息计量学分析方法,结合语义网的相关理论和方法,构建了基于计量分析的馆藏资源语义化理论模型。通过对信息计量学和语义网进行理论和可视化应用的类比研究,发现:①信息计量学从定义、主要规律、馆藏资源元数据的关系研究到元数据的可视化展示都与语义网的理论、应用非常类似;②基于计量分析的馆藏资源语义化方法在语义关系深度、计算机自动处理程度、可重复性和实用性方面比基于元数据和领域本体的馆藏资源语义化方法更好;③把计量分析方法应用到馆藏资源语义化中,不仅可以提高馆藏资源的应用效率,解决基于元数据和领域本体语义化的不足,还能促进交叉研究和学科发展。图5。表3。参考文献10。

关键词 信息计量学 馆藏资源 语义化
分类号 G250

Theoretical Research on Semantization of Library Resources Based on Informetric Analysis

Qiu Junping & Yu Fan

ABSTRACT On the basis of revealing the problems existing in the present library resource semantization methods, this paper constructs a theoretical model of library resources semantization based on informetric analysis methods and semantic web theory. Through analogical comparison we discover that the definition, main laws, metadata relation and visualization research of informetric analysis are similar with that of semantic web. Library resources semantization method based on informetric analysis performs better than that based on metadata and domain ontology in the aspects of semantic relation degree, computer automatic processing degree, repeatability and practicality. Introducing informetric analysis to library resources semantization can not only improve the use efficiency of library resources and solve the shortages of semantization based on metadata and domain ontology, but also promote interdisciplinary research and discipline development. 5 figs. 3 tabs. 10 refs.

KEY WORDS Informetrics. Library resources. Semantization.

自从互联网之父 Tim Berners-Lee 提出语义网的概念后,语义网得到了学术界的广泛关注。随着研究的深入,语义网逐步被引入医学、生物学、地理科学、计算机科学等多个学科。在知识爆炸的年代,图书馆作为高校师生科研、学习的主要知识来源,为了提高用户在检索馆藏资源

时的满意度,数据的高效规范组织显得尤为重要。本文在分析目前馆藏资源语义化方法的基础上,对计量分析方法与语义网方法进行类比,构建了基于计量分析的馆藏资源语义化理论模型,为馆藏资源语义化方法提供参考。

^{*} 本文系国家社科基金重大项目“基于语义的馆藏资源深度聚合与可视化展示研究”(批准号:11&ZD152)的研究成果之一。
通讯作者:邱均平,Email:jpqiu@whu.edu.cn

1 语义网在馆藏资源中的应用现状

按照数据存储形式划分,馆藏资源包括资源内容和元数据。资源内容是用户最终需要的资源,包括各种类型的数字化文档,例如论文、著作等。元数据用来描述馆藏资源的属性特征,包括题名、作者、期刊名等。通过元数据的描述,图书馆员能对每一个资源进行入库、索引处理,用户也能利用元数据进行查寻。针对两种不同的数据源,形成了基于领域本体的馆藏资源语义化研究和基于元数据的馆藏资源语义化研究(图1)。

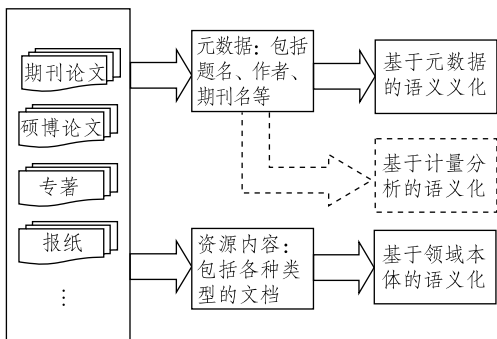


图1 馆藏资源的语义化种类

利用领域本体对馆藏资源进行语义化包括三个步骤^[1]：第一步,确定领域本体的范围。不同的领域能够提取的概念、关系、属性和构建的方法是不一样的。第二步,构造本体库结构。领域本体库是一个树型结构,首先要确定核心概念,然后以核心概念为基础,逐层细分,每一个节点都可以再分成若干个子节点,最后形成一个完整的本体库结构。第三步,填充实例。实例就是存储在本地库结构里的具体内容。本体库结构只是本体库的一个空架子,没有实质性内容,只有填充实例后才能应用。

基于元数据的馆藏资源语义化过程与领域本体的语义化过程非常类似^[2-3],它们的差别主要体现在:领域本体的库结构、关系和属性随着领域的不同而不同,比如医学领域的本体库结构、关系和属性与生物学的存在很大差异。

领域本体在确定核心概念、关系和属性阶段需要领域专家长时间分析、判断和讨论。馆藏资源的元数据具有统一的标准,比如 MARC 数据或者 DC 元数据,字段都是固定的,不存在领域差异性。本体库的核心概念、关系和属性不需要专家的参与,可以由计算机自动完成。

2 馆藏资源语义化方法存在的问题

基于领域本体的和基于元数据的馆藏资源语义化研究,都成功地把语义网理论、方法和技术引入到图书馆领域,对馆藏资源组织、检索效率的提高提供了新的思路。但是,两种馆藏资源的语义化方法也存在一些问题。

基于领域本体的馆藏资源语义化存在以下问题^[4-5]：

(1)领域本体的不可重复性问题。构建领域本体之前,领域专家对某领域进行知识概念、关系、属性的抽取。不同的领域专家对同一领域的认识很难完全一致。虽然都是某领域的专家,但是他们对该领域的熟悉程度有强弱之分。即使是同一位专家,随着年龄的增长,知识的积累,对于同一领域的认识也可能大相径庭。这就好比德漠克利特所说的“人不能两次跨进同一条河流”。这种现象导致的结果是对于同一个领域构建的领域本体,每一次都会出现不同的结果。

(2)本体重用问题。每一个学科领域都拥有自身独有的知识点和知识关联,我们称之为知识网络。小规模领域本体无法构建某一个学科的知识网络,往往会截取学科里的一小部分知识网络作为研究对象。从理论上讲,只要把整个学科的知识网络用本体语言构建起来,就可以实现学科知识的语义化。而实际上,目前还没有哪个学科的知识网络被完整地构造出来。不同的研究机构总是会选择一个学科里的小领域进行研究,采用同样的语言、方法和工具构建领域本体库,只是数据源有所改变,其中很有可能存在重复构建的领域本体库。

(3)本体粒度问题。本体的应用经历了从基于内容的本体到基于元数据的本体过程。本

体表示的知识单元的粒度由小至大。粒度越小,语义关系发掘越深越窄,主要依靠人工处理构造本体库,计算机起着辅助作用。粒度越大,语义关系越浅越宽,主要依靠计算机自动构造本体库,人工处理起着辅助作用。本体粒度并不是越小越好。粒度太小,关系过于复杂,无从下手。构建本体库前还需要考虑可行性问题。如果关系错综复杂,构造本体库的时间也长。

基于元数据的馆藏资源语义化能够借助计算机构造本体库,最大的问题是揭示的语义关系太浅显,与关系数据库构造的语义关系差不多,语义关系的揭示方面还有待提高。

基于以上两种馆藏资源语义化存在的问题,本文提出了一种新的馆藏资源语义化方法——基于计量分析的馆藏资源语义化(图1),试图通过引入计量学的相关分析方法来融合两种语义化方法的优势,尽可能提高计算机自动处理能力,同时提高语义关系揭示能力。

3 计量分析方法引入到馆藏资源语义化的可行性

本节从理论和可视化应用两个方面对计量分析方法和馆藏资源语义化方法进行类比分析,探讨计量学与语义网之间存在的关系。

3.1 理论类比

Tim Berners-Lee 提出的语义网七层模型^[6]中,本体处于中间层,底层的 URI、XML 和 RDF 作为 OWL 的语言基础,本体的推理、验证需要以本体库作为数据支撑,本体是模型中最关键的一层。Studer 对本体进行了定义^[7]:“本体是共享概念模型的明确的形式化规范说明。”从定义中可以看出,语义网的首要工作是要找出概念和关系,然后通过语义网语言、工具构建概念模型。概念和关系是构成语义网的最小单元。

表1 信息计量学中的主要规律

规 律	关 系	数学模型	提出者
文献指数增长规律	发表文献数量与发表时间的关系	$F(t) = ae^{bt} (a > 0, b > 0)$	普赖斯
文献老化规律	被利用文献数量与发表时间的关系	$C(c) = Ke^{-at}$	贝尔纳和 Brookes
		$y = 1 - \left(\frac{a}{e^x} + \frac{b}{e^{2x}} \right)$	Burton 和 Kebler
		$Y(t) = C_0 (e^{-at} - e^{-mt}), m > a$	阿弗拉米斯库
文献集中与离散分布规律	论文数量与期刊数量的关系	$n_1 : n_2 : n_3 : \cdots 1 : a : a^2 : \cdots$	布拉福德
文献词频分布规律	词的频次与序号的关系	$F_r \cdot r = C$	齐普夫
文献作者分布规律	论文数量与作者数量的关系	$f(x) = C/x^n$	洛特卡

信息计量学主要以元数据作为研究对象。信息计量学的定义是^[8]:应用数学、统计学等方法从定量的角度分析和研究信息的动态特性,

并找出其中的内在规律。表1列出了普赖斯、布拉福德等人提出的规律、数学模型以及文献数据之间的关系。规律是指事物之间内在的必然

联系,用数学模型表示,实质上是指自变量与因变量之间的函数关系,或者称之为映射关系。这种关系与语义网定义中概念之间的关系非常类似。

如果说表 1 中列出的规律是比较宏观的语义关系,那么表 2 列出的基于元数据的语义关系则是比较微观的关系。表中列出的是几种比较常见的计量学的关系研究。从最简单的作者合作揭示的一层关系到引文耦合揭示的二层关系,再到作者关键词耦合揭示的三层关系,可以发现,运用计量学的方法对元数据进行分析,能够发掘出非常复杂的关系。关系层次越浅,利用元数据联系馆藏资源的覆盖范围最大,反之亦然。那么这些关系是语义网中的语义关系吗?

表 2 信息计量学中基于元数据的语义关系

关系类型	语义关系	关系层次
作者合作	研究合作发表论文的作者之间的关系	一层关系
引文耦合	不同论文的作者引证某篇或某几篇完全相同的论文,这些论文之间的关系	二层关系
同被引 (共被引)	某篇或某几篇论文被其他论文同时引用,这些论文之间的关系	二层关系
关键词共 词(共现)	不同论文作者会使用相同的关键词,研究出现相同关键词的论文的关系	二层关系
作者关键 词耦合	不同论文作者会使用相同的关键词,研究出现相同关键词的论文的作者关系	三层关系

语义网中本体的定义有四层含义^[9]:概念模型、明确、形式化、共享。“概念模型”是指通过抽象客观世界中的现象形成的概念,利用这些概念构建的模型。计量学的研究数据本身就是基于元数据的,元数据是描述数据的数据,是对文献信息的概念抽象。“明确”是指使用的概念及概念的约束都有明确的定义。表 2 中列出的关系以及没有列出的关系都是通过信息计量

学领域的学者实证分析得出的,显然都有明确的定义,不会引起歧义。“形式化”是指计算机可读,具体是指用基于 XML 的语言来描述概念、关系和属性,以达到计算机的良好识别性。信息计量学中的概念、关系只要通过本体语言进行再组织,也可以达到这一要求。“共享”是指定义的概念是相关领域公认的概念集。正如上文所说,计量分析方法总结出来的概念、关系是科学的、公认的,不存在不同学者对同一概念、关系有着不同理解的现象。以上分析得出,信息计量学中基于元数据的关系是一种语义关系。

3.2 可视化应用类比

语义网中主要以 RDF、OWL 作为描述语言。RDF 使用 URI 标识资源,XML 作为数据结构,解决了资源唯一性问题和人机可读性问题。RDF 用陈述来描述知识概念及其关系,陈述由主体、谓词、客体组成,三者的关系如图 2 所示。Subject、Predicate 和 Object 分别表示主体、谓词和客体,箭头表示关系的指向。图 2 展示了 RDF 语言描述的最小知识单元。

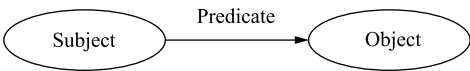


图 2 RDF 中的陈述

从图 2 看出,语义网中的一条关系由两个节点和一条带有属性的边组成,所有关系则由无数个节点和无数条边组成,构成一个网状结构。语义网的可视化工具非常多,例如: TGVizTab、OntoViz、OWL2Prefuse、Protégé、TouchGraph 等,以 OWL2Prefuse 为例,如图 3 所示。图中包括两个部分:矩形框和带箭头的直线。矩形框的黄色表示概念,红色表示实例。带箭头的直线表示概念—概念,概念—属性,概念—实例之间的关系及方向。这些可视化软件生成的语义网图形有三个共同特征:网状结构,节点表示概念,边表示关系。因此,我们可以给出以下定义:以节点表示概念、边表示关系的网状结构图称作语义网的可视化图形。

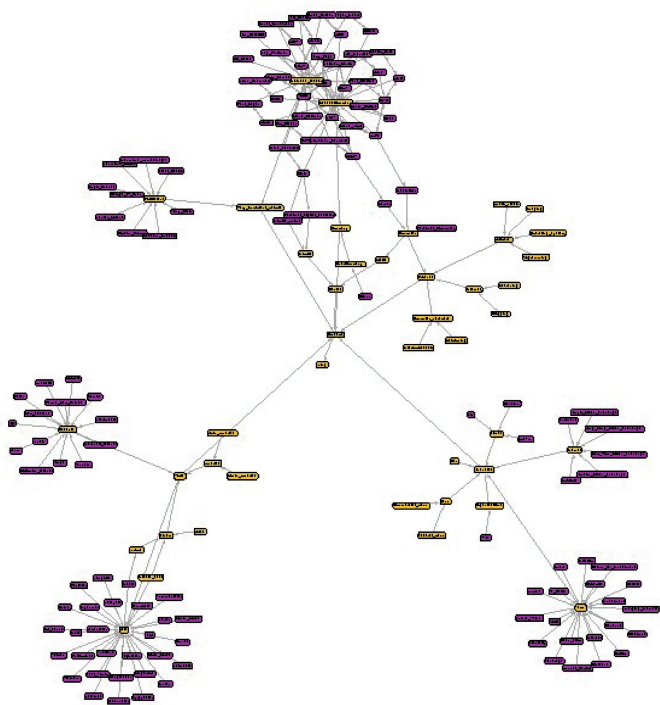


图3 OWL2Prefuse 可视化

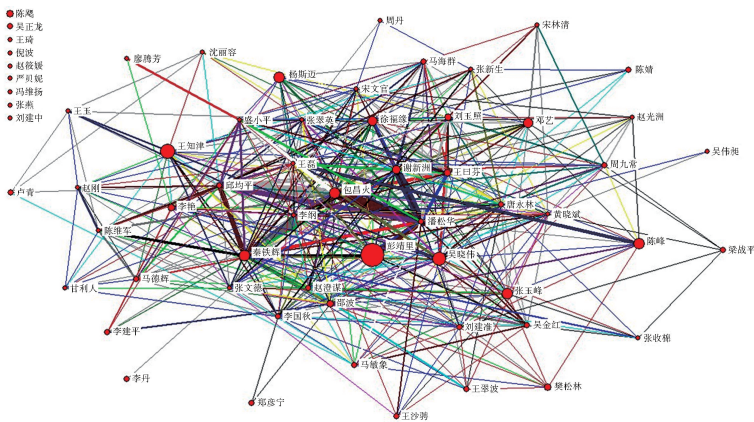


图4 作者关键词耦合关系

语义网还有一个非常重要的特征——语义推理。它利用工具对生成的本体库进行推理,用于丰富本体库和检测逻辑错误。语义推理是在可视化之前的步骤。计量学分析出来的概念、属性、关系使用语义网语言进行描述也可以进行语义推理。

计量学的结果可视化是从最简单的折线图、柱状图开始的。这些图一般由 EXCEL 等比较常用的分析软件生成。图的横轴和纵轴分别表示一个变量,整张图反映了两个变量之间的关系。近几年,社会网络成为计量学的研究热点,各种社会网络的可视化软件纷纷出现,例

如:Ucinet、NetDraw、Pajek 等。我们只需要将元数据导入软件,软件就能够生成非常绚丽的网络图。类比发现,这些图形同样具有语义网的三个特征。图 4 是以作者关键词耦合为例,由 NetDraw 生成的社会网络图,图中节点表示文献作者,边表示作者关键词耦合关系,整张图形形成了一个复杂的语义网^[10]。

通过理论和应用的类比,我们发现计量学与语义网的研究内容形成了交叉,计量学的研究理论、方法完全可以应用于语义网领域。

4 计量分析方法引入馆藏资源语义化的必要性

计量学的分析方法引入语义网是非常必要的,主要包括两点:

(1)提高基于语义的馆藏资源的应用效率。表 3 对三种馆藏资源语义化的指标进行了简单对比。好坏程度用高、中、低区分,四个指标都是值越高越好。基于计量分析的语义化吸收了其他两种方法的优势,其语义关系深度高于基于元数据的语义化,计算机自动处理程度高于基于领域本体的语义化。计算机自动处理程度不高是基于领域本体的馆藏资源语义化应用的瓶颈,语义关系揭示过于简单是基于元数据的馆藏资源语义化应用的软肋,基于计量分析的馆藏资源语义化牺牲了部分领域本体的语义关系,但大幅度提高了计算机自动处理的程度。计算机自动处理程度能够反映出该项技术的应用广度。显然,基于计量分析的语义化更容易产生规模效应,走向市场化。

表 3 三种馆藏资源语义化的指标对比

	元数据的语义化	计量分析的语义化	领域本体的语义化
语义关系深度	低	中	高
计算机自动处理程度	高	中	低
可重复性	高	高	低
重用性	高	高	低

(2)促进交叉研究和学科发展。目前,计量学的研究一般步骤包括选定数据源、选定分析方法、进行处理、得到结果、对结果进行定性分析。实际上,计量学的分析结果是基于馆藏资源元数据完成的,仅仅对结果进行分析未免觉得可惜,完全可以把结果与语义网相结合,应用到计算机领域。反过来,语义网在应用到馆藏资源领域时,完全可以考虑借用以馆藏资源为研究对象的相关研究方法。计量学的研究结果能够应用到馆藏资源的语义化研究中,而馆藏资源的语义化研究正需要计量学这样的研究方法。两个学科的交叉研究,既能够促进学科的发展,扩展学科的研究内容,又可以共同解决馆藏资源的语义化问题。

5 基于计量分析的馆藏资源语义化理论模型

本节结合语义网相关技术和计量学相关分析方法构建馆藏资源语义化的理论模型(图 5)。模型中使用了资源本体的概念,之所以没有使用领域本体这个词,是因为两者存在差异。领域本体会把范围限制在某一个领域,资源本体的数据没有领域之分,包括所有学科馆藏资源的元数据。借用本体的定义,本文把资源本体定义为馆藏资源共享概念模型的明确的形式化规范说明。

基于计量分析的馆藏资源语义化分为六个步骤:

第一步,数据采集入库。从期刊库、论文库、图书馆等馆藏资源数据库中采集数据。比如:CNKI、CSSCI、WOS 等。这些数据无法大批量自动下载,只能手工下载。

第二步,标准化处理。采集的数据并不符合数据库的要求,需要对数据进行处理。采用 Eclipse 集成开发环境和 Java 语言提取所需字段,并存入数据库,形成知识单元库,包括作者表、引文统计表、主题词表和关键词表等。值得注意的是,不同数据库提供的下载字段并不相同,如果需要整合不同数据库的元数据,需要对表示相同意思的不同字段进行映射。

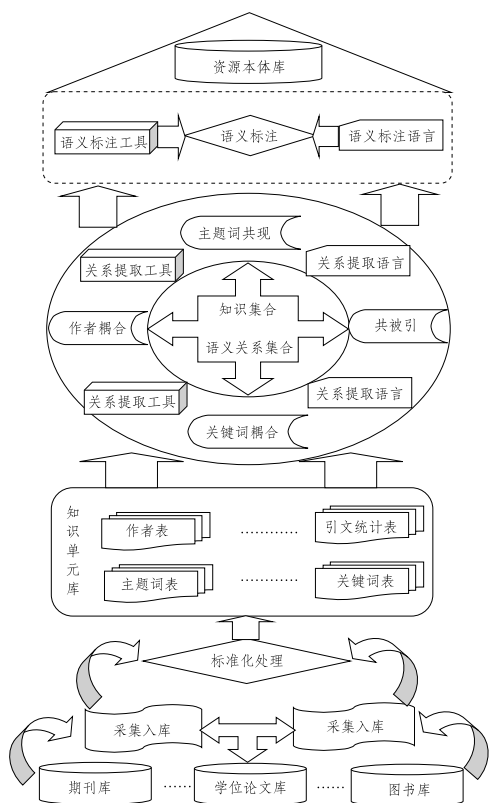


图5 基于计量分析的馆藏资源语义化模型

第三步,利用信息计量学的分析方法和 EX-CEL、VBA、SPSS 等工具生成矩阵,进行各种聚类操作,例如主题词共现、作者耦合、共被引和关键词耦合等。最后形成一个知识(概念)集合和语义关系集合。显然,信息计量学的分析方法分析得越全面,馆藏资源的概念和关系也揭示得越透彻。另外,不同的计量分析方法可以重叠使用。例如,可以利用作者耦合分析出发文作者之间的关系,以作者为出发点,可以继续分析关键词耦合蕴含的关系等。最终会发现,计量分析方法揭示出来的馆藏资源元数据之间的关系形成了一个非常复杂的语义网络。网络越复杂,用户能够通过更多的检索手段获取需要的信息,越有利于提高检索服务的质量。

第四步,确定资源本体库的库结构、概念之间的关系。这里的概念实际上是馆藏资源的元数据,而关系就是通过计量分析得到的关系。

这是运用语义网技术对上一步收集到的概念和关系的再组织。这一步与领域本体库结构的构建过程相比简便得多。所有的概念、关系都已确定下来,无需专家参与,可以完全计算机化。

第五步,利用 Protégé 工具构建资源本体库,添加实例,并推理检验。这一步并不是本模型的重点,添加实例和推理的过程与领域本体的过程基本相同。

第六步,生成资源本体库。资源本体库生成后,可以作为馆藏资源前台系统的数据支撑。

本节提出的理论模型的核心是把计量学分析的语义关系直接用语义网语言替换,形成描述馆藏资源元数据的关系。随着计量学研究的不断深入,元数据新的关系会不断被发掘出来,元数据的语义关系也会变得更加复杂,用户能够通过更加多样化的检索途径找到需要的资源。

6 总结

通过理论研究,我们看到,扩大基于语义的馆藏资源应用范围的最大障碍是概念、关系、属性的规范化、统一化。这也是语义网应用的最大障碍。规范化、统一化是语义网应用过程中的根本问题。正是因为没有规范化、统一化的概念集、关系集、属性集,才会出现前面提到的不可重复性、重用性和粒度的问题。理论上,如果所有的知识概念、关系都能够统一定义,建构本体库就是一件非常轻松的事情。但是,知识的增长永远超出我们的想象,不可能把出现的新知识、新关系系统收录到规范化的知识库。计量学的理论、方法与语义网的结合起到了抛砖引玉的作用,在构建领域本体时,首先可以考虑粒度比较大、相对统一、规范的概念;同时,需要一个国际化的组织专门定义各种知识概念,概念由宏观到微观,形成一个庞大的、规范的知识库,进行研究开发的组织机构不能构建新的概念、关系,只能调用这个知识库。如果这样做了,语义网的终极目标就不是遥不可及的了。

参考文献:

- [1] 董慧,余传明,杨宁,等.基于本体的数字图书馆检索模型研究(Ⅲ)——历史领域资源本体构建[J].情报学报,2006,25(5). (Dong Hui, Yu Chuanming, Yang Ning, et al. Research on the ontology-based retrieval model of digital library—History domain ontology building [J]. Journal of the China Society for Scientific and Technical Information, 2006, 25(5).)
- [2] L Ding, T Finin, A Joshi. Swoogle: A search and metadata engine for the semantic web[J]. D-Lib Magazine. 2004, 5(12).
- [3] McCmy Gallagher. Extending the role of metadata in a digital library system[C]//Proceedings of the IEEE research and technology advances in digital libraries, 1999.
- [4] 张云秋,冷伏海.领域本体整合的问题及对策研究[J].中国图书馆学报,2007,30(2). (Zhang Yunqiu, Leng Fuhai. Domain ontology integration: Problems and strategies [J]. Journal of Library Science in China, 2007, 30(2).)
- [5] 丁晟春,李岳盟,甘利人.基于顶层本体的领域本体综合构建方法研究[J].情报理论与实践,2007,(3). (Ding Shengchun, Li Yuemeng, Gan Liren. Research on integrated domain ontology construction method based on top-level ontology [J]. Information Studies: Theory & Application, 2007, (3).)
- [6] Tim Berners-Lee, J Handler, O Lassila. The Semantic Web[R]. Scientific American, 2001.
- [7] Studer R, Benjamins V R, Fensel D. Knowledge engineering, principles and methods[J]. Data and Knowledge Engineering, 1998, 25(122).
- [8] 邱均平.信息计量学(一)第一讲:信息计量学的兴起和发展[J].情报理论与实践,2000,23(1). (Qiu Junping. Informetrics (I) The first lecture: The rise and development of informetrics [J]. Information Studies: Theory & Application, 2000, 23(1).)
- [9] Guarino N. Formal ontology and information system[M]. Trento: IOS Press, 1998.
- [10] 邱均平,王菲菲.基于SNA的国内竞争情报领域作者合作关系研究[J].图书馆论坛,2010,30(6). (Qiu Junping, Wang Feifei. Analysis on author cooperation relationship of competitive intelligence research in China based on SNA [J]. Library Tribune, 2010, 30(6).)

邱均平 武汉大学信息管理学院教授、博士生导师。通讯地址:湖北武汉武汉大学信息管理学院。邮编:430072。

余凡 武汉大学信息管理学院博士研究生。通讯地址同上。

(收稿日期:2011-12-31;修回日期:2012-02-07)