

基于语义 Web 的科技期刊网络信息检索及其应用

谢文亮¹, 王石榴²

(1. 广东财经大学学报编辑部, 广东广州 510320;

2. 嘉应学院学报编辑部, 广东梅州 514015)

摘要: 针对目前科技期刊数字出版过程中内容表示格式多样化和多数据源带来的异构和跨库检索难、基于关键词匹配无法达到精确匹配、基于关键字的关系数据库检索数据挖掘能力有限以及无法用自然语言进行检索查询等问题, 提出构建基于语义 Web 的科技期刊网络信息检索模型, 并讨论科技期刊领域本体库、应用本体库的创建, 给出该检索模型在科技期刊的创新其服务模式、数据挖掘、知识关联和发现等方面的应用。

关键词: 科技期刊; 信息化建设; 数字出版; 语义 Web; 信息检索

中图分类号: G255.2; C39; G254.9

文献标志码: A

文章编号: 1000-7695 (2015) 02-0196-05

Semantic Web – based Network Information Retrieval for Scientific Periodical Journal and Its Application

XIE Wenliang¹, WANG Shiliu²

(1. Editorial Office of Journal, Guangdong University of Finance & Economics, Guangzhou 510320, China;

2. Editorial Office of Journal, Jiaying University, Meizhou 514015, China)

Abstract: Currently, in the scientific periodical journal publishing, multi-format presentation and multi database sources bring difficulty for heterogeneous inquiry and cross-database inquiry: keyword-based methods can not realize the goal of precise matching; there is limited data mining capability for keyword-based relation database; users can not use natural language to carry out inquiry, etc. In this paper, we propose a semantic web-based network information retrieval model for scientific periodical journal, and discuss the construction of scientific periodical journal field ontology and application ontology. Finally, we give some example applications in scientific periodical of this network information retrieval model, such as the innovation of the service model, data mining, knowledge correlation and knowledge discovery.

Key words: scientific periodical journal; information construction; digital publishing; semantic web; information retrieval

1 科技期刊数字出版中的信息检索现状

数字出版是科技期刊出版的趋势, 并且出版的方式日趋多样化, 如开放存取出版、网络出版、数据库出版、移动出版等。不同的数字出版方式的数据表示方式不同, 给内容挖掘、知识发现与关联、跨库检索等带来了极大难度; 另一方面, 基于关键词匹配的方法无法解决一词多义, 或一义多词的情形, 无法精确的检索到结果。在我国, 以中国知网、万方数据、维普资讯、龙源期刊网为代表的期刊数据库, 以中国科技论文在线为代表的开放存取期刊, 以及一些科技期刊官方网站所形成的互联网出版等, 都使用基于关键字的关系数据库服务, 其检索功能仅限基本查询, 无法从数据库中挖掘和发现更多有意义、相关联的信息; 各种数据库使用的数据库语言、架构原理等也不同, 这对跨库检索、异构数据

库之间互操作等带来一定的挑战。国外发达国家的科技期刊主要朝集团化发展、集约化经营, 虽然在资源聚合方面做得比较好, 然而在语义出版方面还没重视研究。在国际科技期刊相比, 我国科技期刊竞争力处于弱势地位, 语义出版为我国数字出版转型过程中掌握话语权、提高竞争力提供了契机。

鉴于此, 语义 Web 技术成了研究热点。基于语义 Web 的网络信息检索有利于科技期刊在内容检索过程中进行深入的知识挖掘, 知识关联, 创新其服务模式, 能提高科技期刊资源整合能力, 有利于科技期刊竞争力的提高^[1]。语义 WEB 检索模型在计算机领域早已是成熟的技术, 并且在计算机应用领域^[2-5]、图书馆数字化建设^[6-8]中已经有成熟的应用, 但国内外学术出版研究机构及大型出版商在有关语义出版方面的研究和应用较少, 仅有一些文献都提出关于某个学科的本体的构建, 如 CSSCI 本

收稿日期: 2014-06-24, 修回日期: 2014-10-16

基金项目: 广东省科技计划项目“在编科技期刊语义化及关键技术研究”(2012B060200013)

体^[9]、医学领域本体^[10-11]、农学本体^[12-13]、地理本体^[14]、爆破领域本体^[15]、工业自动化仪表领域本体^[16]等, 本体仅是构建语义 Web 模型的组成部分之一, 并且其本体仅适合特定的领域, 没有涉及到科技期刊本体的构建, 本文作者曾提出了一个基于 Web 的科技期刊数字化模型^[1], 但没有对模型中的本体解析和领域本体领域构建作深入的研究。文献[2]中提出的基于本体的语义网检索模型的对象是计算机应用领域, 其应用原理虽适合科技期刊, 但必须对领域本体、元数据、标注等作修改, 以便适合科技期刊。本文意对科技期刊的语义 Web 网络信息检索模型建立作有益的探讨。

2 基于语义 Web 科技期刊网络信息检索模型

基于本体的语义 Web 科技期刊网络检索系统是一种脱离具体硬件平台的服务模式, 能运行于跨平台出版机构, 并能使用自然语言检索。这对多种出版方式的资源整合、跨库检索、知识挖掘与关联等提供了可行性。在基于本体的语义 Web 科技期刊数字出版服务模式中, 要解决几个关键的问题: 一是自然检索语言输入能被计算机读懂, 查询结果能以人性化的语言返回给读者; 二是不同数据类型的数据源的互操作问题。关键问题的解决, 有赖于科技期刊语义 Web 数字化出版网络信息检索模型的构建。

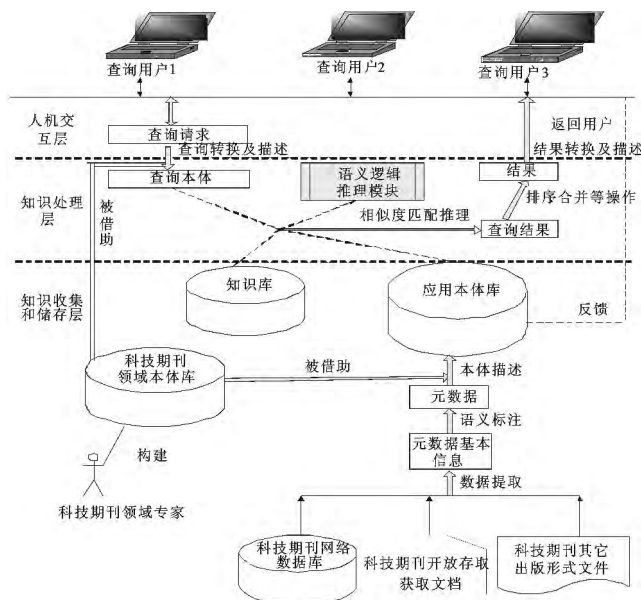


图1 基于本体的语义 Web 科技期刊网络信息检索模型

2.1 检索模型概念描述

本文提出一个基于本体的语义 Web 科技期刊网络信息检索模型(图1)。其中一些重要的概念解释如下:

(1) 科技期刊网络数据库: 以网络出版数据库形式存在的科技期刊数字出版方式, 如中国知网、

万方数据、维普资讯以及龙源期刊网等;

(2) 科技期刊开放存取文档: 在网站开放电子文档供读者免费下载的数字出版方式, 一般是在期刊各自的官方主页上发布, 还有中国科技论文在线之类的开放存取电子期刊也是以主页发布的方式;

(3) 科技期刊其它版形式文件: 一些科技期刊直接使用 Web 出版的形式, 如网页、博客发布等; 另外还有移动出版等形式;

(4) 元数据: 元数据 (mega data) 是组成期刊全文的基本元素, 如题名、作者信息、作者机构等。不同的期刊有不同的元数据, 这给检索造成一定的难度; 一些机构抽出一些比较常用的元数据组成元数据集, 如 DC (Dublin Core); 中国知网也有自身的元数据集, 共集中目前一共有 17 个元素, 并利用这个 17 个元素进行检索操作;

(5) 本体: 在本文中, 本体是科技期刊资源概念化的、明确的、规范说明。即科技期刊本体每一个概念要经过定义, 并可被计算机识别处理, 本体必须有一套严格、无歧义的关系演算规范和概念描述方法;

(6) 科技期刊领域本体库: 科技期刊有不同于其他科技领域的元数据, 以及元数据与元数据之间的关系定义、概念描述、规范说明等。以分类主题词为例, 科技期刊的分类主题词不同于社科期刊, 即使同样是科技期刊, 各种学科的分类主题词也不一样。因此, 对科技期刊来说, 必须建立自身的领域本体库才能实现更好的检索查询;

(7) 查询本体: 查询本体将作者检索查询的输入转化成与领域本体形式相同的本体, 用于查询指令的转化;

(8) 应用本体库: 应用本体库将从各种科技期刊文件源获取的期刊元数据通过语义标注、本体描述等方法, 构造起与领域本体库形式相同的本体;

(9) 知识库: 知识库定义和存储一套语义逻辑的推理原则, 包括排序、合并、分类、聚类; 还有科技期刊关联的知识, 包括文章引用与被引用关系、参考文献的相似度原理、相关关键字匹配、主题匹配、引文网络、相关作者文献聚类等一套与期刊相关的知识体系, 实现查询本体和领域本库的推理匹配等。

2.2 信息检索模型分层

从图1中看出, 所提出的科技期刊语义 Web 网络信息检索模型分为三个层次, 分别是人机交互层, 知识处理层, 知识收集和存储层。

2.2.1 知识收集和存储层。知识收集和存储层负责建立科技期刊领域本体库, 并根据检索的需要, 从各种期刊出版数据源中抽取元数据信息, 进行语义标注, 形成元数据; 再借助科技期刊领域本体库对所抽取的元数据进行本体描述, 形成应用本体库。

该层是整个网络信息检索系统中最关键、最重要的部分,涉及到科技期刊对数据源的元数据提取、标注,以及科技期刊领域本体库、应用本体库的构建等。

(1) 科技期刊数据库的来源

数据的来源必须是权威的,目前,互联网出版由于审核还不够严格,数据的权威性不够,因此,数据的来源一般是科技期刊出版单位提供的;数据库服务商是一个特殊的数据提供商,聚集了大部分的科技期刊出版单位的历年来的数据;还有一些科技期刊其它的出版方式,也形成了数据库的来源。由于不同的数据源的格式不同,如有数据库文件格式、HTML 格式、以及 XML 格式,在不同的数据库出版商中,其对应的数据格式也不同;同样,不同的期刊主页使用的格式也不同。不同的数据格式给检索带来异构问题;不同的数据库来源给跨库检索问题。因此,必须构建基于本体的语义 Web 网络信息检索系统才能更好的解决问题。

(2) 科技期刊领域本体的构建

科技期刊本体作为一个特定领域本体,必须在科技期刊专家的指导下构建的。领域本体有其自身的一些属性和操作行为。本体的类型有多种,如文献 [7] 中指出的,本体有单本体、多本体以及混和本体。科技期刊本体可由全局本体和局部本体组成,全局本体可由权威机构或由国家主导建成;局部由各个数据库服务商以及各大科技期刊出版单位自建。各个单位自建的本体的规模、大小不一样,这是同他们出版的内容决定的,如光学期刊所创建的本体就仅与光学有关。

①科技期刊本体的开发工具

本体开发工具有多种,本体编辑器有 Protégé、SWOOP、OntoEdit 等,本体开发语言 OWL、RDF 等,本体推理机有 Jena、Racer、Pellet 等,目前,有 75.9% 的开发者使用 OWL 本体开发语言开发本体^[8]。

②科技期刊本体的建设

科技期刊本体有其自身的特点,不能照搬其它本体的构造方法,本文吸取文献 [9-16] 中的本体的构建原理,考虑科技期刊的元数据特点,总结出科技期刊本体构建的七步骤,即:

一是确定本体范畴:不同的科技期刊数据源有其自身的特征,不能一概而论,如《科技管理研究》要求给出作者出生年月、性别、祖籍、研究方向等,文章要求给出英文摘要和关键词等,在确定本体范畴时要根据期刊数据源的特点来决定。

二是本体复用:本体的构建是一个较大的工程,如果能对已有的本体进行复用,将起到资源共享,减少工作量等目的。期刊数字化可因地制宜,采用多种数据格式,然后通过基于 XML 的 RDF 对资源

进行描述,实现各期刊资源的互操作、共享以及知识关联。国际上比较流行的有关期刊的本体有 DC (Dublin Core),可根据期刊的需要对其中的字段进行取舍,也可以对其进行扩展,以适合期刊的需要。另外,一些科技期刊已经建立了其学科本体,如医学领域本体^[10]、中医药领域本体^[11]、农业学本体^[12-13]、地理本体^[14]、爆破业领域本体^[15]、自动化仪表领域本体^[16]等,这些本体可直接复用。对于没有现有的本体的科技期刊,可根据文献 [9] 中所描述的方法进行创建。

三是本体类目及等级体系:期刊本体类目可按照中国知网的 17 个检索字段^[17],也可按照 CSSCI 所定义的 17 个字段^[9],不过,不同的科技期刊数据源可按照的自身的需要,尽可能选择有代表性的字段。选取的原则是能最大可能的反映期刊资源的关系的字段,并可以根据所需要体现的知识关联的需要,添加一些字段,作为本体类目。本体类目有等级体系,如一级类目下面有二级类目,二级类目下面又有三级类目等。期刊本体类目构建时要对类目的等级体系设计周全,不遗漏任一类目。

四是定义类的属性:为了使得检索更多期刊的知识、发现更多作者、期刊、文章、机构、引用、被引等的关联,类的属性定义应该考虑特定的类的关系,父类与子类之间的关系等,并在类的属性中尽最大可能的体现出来;一个最好的办法是尽可能将类名——类名——属性用表格列出来,并遍历类与类之间的关系。

五是添加实例:根据所定义类目,分类对数据源检索,对检索结果进行筛选、去重、提取等操作,最后将检索的条目储存,作为类的实例。

六是定义实例属性:在语义 WEB 中,实例与实例之间的关系是通过语义相似度计算出来的,其计算方式有以下几种:基于特征的计算方法^[17]、基于距离的计算方法^[18]、基于信息论的计算方法以及混合方法^[19]等。

七是对本体进行编码、形式化:各个期刊根据自身的需要构造了期刊本体。为了使计算机能读懂所构造的期刊实例,需要通过本体描述语言,对本体进行编码和形式化。目前,一般都是使用 OWL 本体描述语言^[20],实现期刊本体的编码。

(3) 科技期刊应用本体库的构建

科技期刊应用本体库与领域本体库的区别是,领域本体库是抽象的本体库,是某个学科资源特点的抽象描述;而应用本体库是领域本体库的具体化、实例化。应用本体库构建时,需要经过以下的步骤:一是元数据信息抽取:从各种科技期刊数据源中根据一定的原则进行分词操作,提取出所有的元数据信息;二是元数据信息的标注:所提取出的元数据信息经过词性标注、词义标注、语义标注等形成元

数据; 三是应用本体库构建: 参考领域本体库的构建方法, 对元数据进行本体描述, 形成期刊资源的本体对象, 所有的本体对象构成了科技期刊应用本体库。

2.2.2 知识处理层。知识处理层将来自人机交互层的用户查询使用本体描述语言转化为查询本体, 查询本体结合知识库推理规则, 调用语义逻辑推理模块, 将查询本体与应用本体库进行相似度匹配推理, 得到一系列元始查询结果, 并将结果进行排序、合并等操作, 形成最终的用语义表示的结果返回给人机交互层处理。

2.2.3 人机交互层。人机交互层是处理查询用户的查询请求, 将查询的内容参考领域本体库的构建进行转换、标注, 转成计算机能识别的本体查询。知识处理层从知识库中提取相关的关联知识, 并调用语义逻辑推理模块, 将查询本体与应用本体库进行相似度匹配, 对查询结果进行排序、合并、分类、聚类等操作, 并将相关的结果提交给人机交互层进行语义转换及描述, 最后的结果由人机交互层发送给查询用户。

3 语义 Web 的科技期刊信息检索应用

语义 Web 科技期刊信息检索能在知识挖掘, 知识关联以及创新其服务模式上发挥作用。

3.1 创新科技期刊服务模式

首先, 通过构建科技期刊语义本体, 无论是读者、出版单位或编辑人员都能够使用自然语言对想要查询的对象进行描述, 由信息检索系统转变成计算机能理解的本体, 实现了跨库检索, 这是一种创新服务的模式。语义 Web 科技期刊数字化出版服务中要解决不同的期刊内容提供者之间的互操作问题, 包括语义互操作和语法互操作。由于不同的内容提供者使用的知识体系语法不一样, 如有的使用 XML 语言描述, 有的使用关系数据库表示, 有的使用 HTML, 甚至有的使用纯文本文件, 因此, 各个知识体系间无法互联操作; 另一个方面是语义的互操作, 如中国知网数据库和万方数据、维普的数据库都是基于不同的开发方式, 对于相同的关键词, 他们表达的内容可能不一样; 对于不一样的关键词, 他们有可能表示的是同一个内容。因此, 利用语义 WEB, 可以解决跨库的检索, 可解决基于语义相似性的匹配, 提高检索的准确性, 另外, 用户端能够通过一种有意义的自然语言进行检索, 如输入查询指令“查找 5 年来以第一作者在《科技管理研究》上发文超过 5 篇的作者姓名、职称、单位、联系方式、文章篇名”, 而检索结果又能以一种有意义的自然语言形式返回。

3.2 科技期刊知识挖掘与知识关联

通过使用语义 Web 信息检索服务, 对所有的数

据都加标签, 使计算机能识别, 这就增加了被检索到的概率, 并且也增加数据与数据的关系, 能发现更多的关联, 能将一些看似毫无关系的数据通过某种方式关联起来, 达到聚类、预测、信息挖掘等目的。通过为科技期刊建立元数据, 能起到很多作用, 如选题挖掘、作者发现, 文献聚类和关联、学术不端检测的精确等。传统的科技期刊网页信息发布中, 其检索的数据项仅有 title, content, body 等项; 中国知网数据库发布的检索数据项作了扩展, 截止到目前 (2013-12-24) 为止, 有 SU (主题), TI (题名), KY (关键词), AB (摘要), FT (全文), AU (作者), FI (第一责任人), AF (机构), JN (文献来源), RF (参考文献) YE (年), FU (基金), CLC (中图分类号), SN (ISSN), CN (统一刊号), IB (ISBN), CF (被引频次) 等 17 个检索项 [21], 其检索功能比网页发布强, 但仅限于中国知网的查询系统, 其它的搜索引擎无法检索到。而在语义 Web 服务中, 利用本体的强大描述功能, 可定义任意的有关期刊的数据项, 并通过本体转成计算机能识别的元数据, 不仅提高匹配的准确度, 强化了知识挖掘和知识关联功能。

(1) 作者群以及团队的挖掘。强大的知识挖掘功能, 能运用于很多方面, 如通过对读者发文量、发文选题、发文刊物档次、作者机构等信息的挖掘, 能发现某学科的核心作者群, 以及由作者群所形成的一个个研究团队, 这有利于科技期刊出版单位和编辑人员的策划和组稿, 扩充了稿源 [22];

(2) 学术不端检测准确率的提升。目前的学术不端检测是基于关键词的匹配检测方法, 这种方法一方面经常发生误检 [23]; 另一方面, 由于没法对更隐蔽的学术不端进行关联, 发现更多的学术不端行为, 也经常导致漏检的情况发生。基于本体的语义 Web 扩大到对任何数据都能描述, 加强了元数据与元数据之间的关联, 应用于学术不端检测, 能发现更多的隐性学术不端;

(3) 选题挖掘。语义 Web 的跨库检索提高了知识挖掘和知识关联, 通过检索近期的研究热点、研究动态、作者的研究方向的转移等, 科技期刊能最大的把握学科发展的动态和趋势, 为选题策划指定了方向;

(4) 期刊资源的整合, 消灭信息孤岛, 最大提升期刊资源的利用率。通过利用元数据构建出版过程的本体, 一定程度上能促进出版过程的出版数据知识组织存储和检索。在建成的全局应用本体上, 不仅支持作者检索、文献检索、语义 Web 服务还支持选题发现、学科趋势跟踪等, 对办刊过程的各项指标现状以及发展趋势也能了然于胸。

4 结语

在出版信息化时代, 大量数据的收集和存储将

为语义 Web 的发现提供数据基础。随着大数据技术的发展, 将其与语义 Web 网络信息检索服务的结合, 将提高有用信息的挖掘能力, 知识关联进一步紧密, 科技期刊数字出版过程中的各种隐藏关系也能更准确的挖掘和预测出来, 这对提高科技期刊数字出版中的资源整合、营销能力、竞争力, 以及创新其服务模式将起到重要的作用。

参考文献:

- [1] 王石榴. 基于语义 Web 的科技期刊数字化 [J]. 中国科技期刊研究, 2013 (6): 1143 - 1145
- [2] 王珏, 向朝参, 王萌, 等. 语义 Web 服务发现研究现状与发展 [J]. 计算机应用研究, 2013 (1): 7 - 12
- [3] 王志华, 赵伟. 基于本体的语义网检索模型及关键技术研究 [J]. 计算机工程与设计, 2011 (1): 145 - 148
- [4] SRINIVASAN N, PAOLUCCI M, SYCARA K. Semantic Web service discovery in the OWL - S IDE [C]. Proceedings of the 39th Hawaii International Conference on System Sciences, 2006
- [5] MARTIN D, BURSTEIN M, HOBBS J. OWL - S Semantic Markup for Web Services [EB/OL]. <http://www.w3.org/Submission/OWL-S/>
- [6] 周静怡, 黄国彬. 2007 - 2008 年国外语义网研究与应用进展 [J]. 图书馆建设, 2009 (1): 19 - 23
- [7] 王蔚. 基于本体的图书馆数据集成研究 [J]. 情报探索, 2012 (2): 93 - 96
- [8] 周静怡, 黄国彬. 2007 - 2008 年国外语义网研究与应用进展 [J]. 图书馆建设, 2009 (1): 19 - 23
- [9] 邱均平, 楼雯. 基于 CSSCI 的情报学资源本体构建 [J]. 情报资料工作, 2013 (3): 57 - 63
- [10] 付强. 基于主题词表的医学领域本体的构建研究 [D]. 长春: 吉林大学, 2011: 38 - 58
- [11] 谢琪. 基于本体方法构建中医药概念信息模型的方法学示范研究 [D]. 北京: 中国中医科学院, 2011: 51 - 69
- [12] 何琳. 古农学本体的半自动构建及检索研究 [D]. 南京: 南京农业大学, 2007: 71 - 89
- [13] 李景. 本体理论及在农业文献检索系统中的应用研究——以花卉学本体建模为例 [D]. 北京: 中国科学院, 2004: 145 - 181
- [14] 陈素娟. 基于基础地理数据库的地理本体自动构建研究 [D]. 阜新: 辽宁工程技术大学, 2011: 38 - 56
- [15] 杜振兴. 面向爆破行业的领域本体自动提取技术研究 [D]. 广州: 华南理工大学, 2011: 28 - 41
- [16] 李丽亚, 李春梅, 薛中, 等. 工业自动化仪表领域本体的构建研究 [J]. 图书情报工作, 2009 (10): 111 - 115
- [17] TVERSKY A. A feature of similarity [J]. Psychological Review, 1977, 84 (4): 327 - 352
- [18] VALERIE C. Fuzzy semantic distance measures between ontological concepts [C]. IEEE Annual Meeting of the Fuzzy Information, 2004: 635 - 640
- [19] RODRIGUEZ M A, EGENHOFER M J. Determining semantic similarity among entity classes from different ontologies [J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15 (2): 442 - 456
- [20] 百度百科. OWL - S [EB/OL]. [2013 - 12 - 24] <http://baike.baidu.com/view/1954438.htm>
- [21] 中国知网. 文献专业检索表达式语法 [EB/OL]. [2013 - 12 - 24] http://epub.cnki.net/KNS/help/help.aspx?helpType=zhuanye&url=help_yufa.htm
- [22] 周国清. 出版产业化环境下编辑决策的作者意识 [J]. 湖南城市学院学报, 2011 (09): 32 - 37
- [23] 谢文亮, 李俊吉, 张宜军. 期刊学术不端文献检测系统误检分析 [J]. 中国科技期刊研究, 2013, 24 (6): 1126 - 1129

作者简介: 谢文亮 (1979—), 男, 广东潮州人, 编辑, 主要研究方向为编辑学、计算机应用; 王石榴 (1970—) 男, 编辑, 主要研究方向为期刊编辑出版和期刊管理。