基于关联数据的书目语义检索

邹鼎杰
（国防大学政治学院，上海 200433）

摘要：本文提出通过书目语义检索提升用户书目检索效率的方法。首先，总结国内外开展书目关联数据研究现状，认为现有书目关联数据研究基础足以支撑书目语义检索系统开发；其次，介绍在关联数据运动下发展起来的谷歌语义检索的功能特点和基本组成，并将谷歌语义检索作为开展书目语义检索的最佳实践；最后，基于关联数据现有研究，提出书目语义检索的设计思路和方法，为图书馆设计书目语义检索提供参考和借鉴。

关键词：关联数据；语义检索；知识图谱
中图分类号：G254.9
DOI：10.3772/j.issn.1673-2286.2018.04.009

关联数据目的是构建一个互联的数据网络，计算机拥有一部分数据即可访问整个网络数据[1]。W3C结合开放数据与关联数据，推出开放关联数据[2]（Linking Open Data，LOD）项目，并在LOD影响下积累大量高质量知识库，如Zhishi.me[3]、Freebase[4]、DBpedia[5]和Yago[6]等。谷歌以高质量知识库为基础推出谷歌知识图谱（Knowledge Graph，KG）[7]，并在KG的基础上设计谷歌语义检索产品。谷歌语义检索作为谷歌关键词检索的有效补充，弥补了传统关键词检索无法解决的问题，已经成为谷歌信息检索工具的重要组成部分。图书馆领域作为关联数据的参与群体之一，是否应基于当下关联数据开发书目语义检索，以补充传统关键词书目检索带来的不足值得探讨。

关联数据提出以来，W3C专门成立图书馆开放数据孵化小组以促进图书馆开放数据运动的发展[8]。图书馆界在近十年内广泛开展书目关联数据的理论研究和实践探索，目前已经有大量书目数据以关联数据形式发布，有关关联数据的发布、消费和应用技术也得到充分研究，为图书馆基于关联数据构建书目语义检索创造良好条件。谷歌语义检索是现有语义检索产品中比较出色的产品之一，为图书馆设计书目语义检索提供了最佳实践方案。

首先，本文介绍图书馆书目关联数据的研究现状，认为目前的理论和实践研究已经为开展书目语义检索研究奠定了良好基础；其次，介绍谷歌语义检索的功能特点和结构组成，作为开展书目语义检索的参考；最后，参照谷歌语义检索的设计思路及图书馆关联数据特点，提出图书馆开展书目语义检索的思路和方法。

1 书目关联数据为书目语义检索奠定基础

1.1 书目数据与关联数据

书目数据是关于书目的元数据，是图书馆进行书目控制的重要基础，也是揭示馆藏、开展服务的主要工具，还是传统图书馆对知识进行组织和整序最有价值的贡献[9]。一直以来，图书馆界都强调开放书目数据以提升图书馆信息资源利用率。在提出关联数据以前，书目数据的开放主要基于Z39.50协议，该协议仅支持单次检索，无法有效地支持书目数据的爬取，也无法获取书目之间的关联性。该协议本质上只是提供查询服务，没有做到真正意义上的数据开放。关联数据的提出，为书目数据的进一步开放创造机会。

Christian等[10]在语义Web基础上提出关联数据，希望依托关联数据构建一张数据网络，使用户拥有一部分数据即可通过链接访问网络中的所有数据，从而提高数据利用率，实现真正意义上的数据开放。关联数据被提出以来，国外图书馆陆续将本馆数据以关联数

据规范的形式发布[11]，我国学术界也开展了大量书目数据关联化、关联数据发布技术等理论方面的探讨，为开展书目数据关联化提供很好的参考和借鉴。

1.2 书目关联数据发展现状

美国国会图书馆将主题词表以SKOS表示并以关联数据形式发布[12]，截至2017年，已发布20余个数据集；瑞典国家图书馆将瑞典联合目录约600万条书目记录、2 000万条馆藏记录及20万条规范记录发布为关联数据[13]，并介绍了将图书馆书目数据纳入万维网的技术和方法[14]；德国国家图书馆将200万条个人名称规范记录、19万条主题规范记录、130万条团体名称规范记录，以及51 458个类概念、110条DDC的主题目标记录发布为关联数据，用户可通过网络门户访问这些数据[15]。此外，匈牙利国家图书馆、法国国家图书馆等欧洲国家的图书馆均不同程度地推进书目数据关联化进程。我国虽然尚未实现真正意义上的书目关联数据，但在学术界对书目数据的关联化、语义化展开广泛研究。

1.3 书目关联数据是书目语义检索的基础

语义检索建立在语义Web基础上[16]，书目语义检索系统的开发依赖于书目语义Web的构建。语义Web由Web创始人Berners-Lee提出，其目标是构建一个机器可读的Web，并为此设计了复杂的语义Web七层结构。Berners-Lee提出轻量级简单化的语义Web，即关联数据。在开放数据等的推动下，关联数据运动得以发展，并推动图书馆领域书目关联数据的进一步发展。书目关联数据的本质即书目语义Web，可直接作为书目语义检索系统的一部分。书目数据作为开展用户服务的重要工具，核心价值在于帮助用户尽快检索到所需书籍。书目数据关联化发布经过十年发展，其理论和实践研究均较充分，但用户似乎并没有感受到关联数据带来的益处。以书目关联数据作为开发书目语义检索系统的基础，使用户真正获得书目关联数据带来的益处。

2 谷歌语义检索作为书目语义检索最佳实践

传统检索以关键词匹配为基础，这种检索策略不能真正理解用户表达的语义，只能从统计意义上提供用户最可能需要的信息。语义Web的目标是构建一个机器可读的数据Web[17]。语义Web的提出，为机器准确理解用户需求提供可能，因此Guha等[16]提出将语义检索作为传统关键词检索的补充，以提升用户信息检索体验。随后的语义检索研究集中在基于传统搜索的增强型语义搜索和基于本体推理的知识型语义搜索[18]，并设计出Hakia、Kngine、Kosmix等语义检索系统。我国图书馆界也对语义检索展开理论研究，包括语义检索的原理[19]、设计和测评[20]等。真正意义上面向大众用户的语义检索是谷歌开发的语义检索系统。本文将介绍谷歌语义检索的功能和组成，为开展书目语义检索的参考。

2.1 谷歌语义检索功能

谷歌语义检索首先分析用户提交的检索请求，如果检索请求中包含人名、地名等实体信息，谷歌将在搜索结果页面右侧以可视化形式展现实体信息，左侧页面显示关键词检索返回的结果。

2.2 谷歌语义检索组成

谷歌语义检索建立在KG基础上，该项目于2012年在谷歌黑板报上发布[21]，其目标包括获取最准确的信息、最全面的摘要，以及进行更深入地拓展。最准确的信息指KG能够区分包含歧义的词汇。如“Taj Mahal”既是歌手名字也是地名，如果只进行关键词检索，作为地名的信息很可能被淹没，谷歌可帮助用户发现“Taj Mahal”作为地名的网页信息。最全面的摘要指谷歌能够从散乱的网页中抽取和整合有价值的信息。更深入地拓展指谷歌能够帮助用户发现具有关联性的信息，并通过链接引导用户对话题内容作进一步拓展。在书目检索系统尤其注重两点，一是为用户提供准确、无歧义的书籍信息，二是在现有检索结果的基础上帮助用户进行深度和广度地拓展，发现更有价值的书籍。KG本质上是一个语义Web，语义Web的组成较复杂，但其核心组成主要包括本体和实体。

（1）本体构建。哲学界认为存在一个独立于现实世界客观存在的原理系统，其通过概念和逻辑推理来描述现实世界[22]，并将这个原理系统称作本体。人工智能专家受此原理的启发，认为可以开发一个用于描述现实世界的系统[23]，并以概念化为逻辑起点，将本体定义为一系列具有预定义含义的概念词表[22]。KG通过这一

系列概念及概念间关系构成的系统来描述现实世界。

传统的本体设计通常由领域专家完成。领域专家对其所属领域具有系统性认识，是开发全面描述该领域本体的最佳人选。但领域专家无法定义面向通用领域的客观世界，且领域专家难以直接开发出具有较强通用性的本体。因此，KG通过半自动化方式构建本体。KG维护记录计算机自动抽取概念的数据结构Collection[24]，Collection中包含成千上万个初始概念。有的初始概念生命周期较短，第一天被生成后第二天就被删除；有的初始概念生命周期较长，能在Collection中长期保留。对于Collection中得以长期保留且符合特定规范的初始概念，KG将交由专业人员进行决策和命名，最后成为一个正式概念，保存在KG中。本体抽取还包括术语抽取、概念抽取、关系抽取等任务[25]。

（2）实体生成。实体是知识库中数量最多的部分。第一个大规模中文开放关联数据集Zhishi.me有9个自定义本体概念，截至2015年11月24日已经拥有上亿个实体。DBpedia本体包含685个概念，拥有400万个实体。KG根据本体模式从各类结构化、半结构化数据中抽取实体。KG的实体层主要从以下数据来源中抽取：①维基百科、百度百科等在线数据资源；②特定主题的在线资源，如从专门的天气预报网站获取天气信息、从世界银行获取经济统计数据；③开放的知识库，如Freebase、DBpedia、Yago等高质量知识库；④谷歌搜索数据。

KG是一个面向大众、适用于所有领域的语义Web，其构建难度超过特定领域的语义Web构建。书目语义Web处于图书馆领域，面向特定用户群体，其建设难度相较KG建设小得多。谷歌语义检索实践证明，通过构建语义Web能有效地为语义检索提供支撑，提升用户信息检索满意度。书目语义检索实体数据源于书目关联数据，建设难度相对较小，但能提升用户的检索效率。

3 基于关联数据的书目语义检索

从2000年开始，Web领域的检索工具从以基于目录检索的雅虎逐渐转变为基于关键词匹配的谷歌，图书馆领域的书目检索策略也逐渐由以目录检索为主，发展为以关键词匹配为主；2015年以来，Web领域的检索工具再次发生改变，谷歌语义检索成为谷歌关键词检索的有效组成部分，书目语义检索逐渐成为书目关键词检索的有效组成部分。

3.1 书目语义检索功能设计

作为关键词检索有效补充的书目语义检索，至少在两个方面带来书目检索体验的提升：①准确理解用户检索词汇表达语义，提供更精确的检索结果；②利用关联关系帮助用户进行在广度和深度上的检索拓展。书目语义检索系统需分析用户提交的检索请求，如果检索请求中包含作者、书名、出版社、编辑、丛书等实体信息，检索系统在搜索结果页面右侧以可视化形式展现实体信息，并提供进一步拓展检索的链接，搜索结果页面左侧显示关键词检索返回的结果。

3.2 书目语义检索组成原理

随着关联数据的推动，目前图书馆界已经有比较成熟的书目本体作为各馆设计书目本体的参考。图书馆书目数据具有较强的结构性和较高的质量，从书目数据中提取实体也有较成熟的工具。

（1）本体构建。本体的功能是描述和揭示现实世界。著录是在编制文献目录时对文献形式特征和内容特征进行分析、选择和记录的过程，著录的结果款目是反映文献形式特征和内容特征的著录项目组合[26]。从功能角度看，著录发挥的作用是描述和揭示书目形式及内容特征，著录规范实质是设计用于各图书馆书目统一描述的本体。基于上述考虑，有学者以著录规范MARC为基础设计书目本体，典型的本体有MarcOnto、Dublin Core及BibTeX等[27-29]。这些本体的优点是从不同角度揭示书目特征，在特定场景下均能发挥功能；缺点是只关注书目自身的描述，缺乏对作者等对象的描述及各类对象间关系的建立。

书目记录功能需求（Functional Requirement for Bibliographic Records，FRBR）弥补了MARC时代书目本体的不足[30]，从实体、属性和关系角度描述书目及相关信息，以更全面的视角揭示书目特征，帮助用户发现实体、识别实体、选择实体和获取实体。美国国会图书馆在FRBR的影响下提出书目新格式BIBFRAME。BIBFRAME的设计动机在于满足书目描述的同时，实现书目信息最大范围的交换。本体为各信息系统实现信息交换提供基本设计思路，BIBFRAME采用知识本体对书目数据建模，使用关联数据的原则来组织、展示和分享数据[31]。美国国会图书馆在其官方网站发布BIBFRAME模型和词汇表[32]，其实质是面向所有书目

和文化载体的本体模型，希望在世界范围内构建一个书目语义Web，实现图书馆界内部的书目信息交流及图书馆与外部信息世界的书目信息交流。BIBFRAME是现阶段图书馆领域最完备的书目本体，可作为各图书馆设计书目本体的参考标准和最佳实践。

因此，本文建议书目语义检索系统的本体构建遵循BIBFRAME规范，其优势在于：①减少图书馆聘请专家设计本体耗费的时间成本和人力成本，有利于书目语义检索的推进；②本馆语义检索系统与世界范围内图书馆书目语义检索系统具有良好的兼容性，有利于后续开展各图书馆书目信息集成、图书馆馆际互借等活动。

（2）实体生成。与谷歌语义检索的实体一样，书目语义检索的实体也在书目语义检索中占据绝大部分内容；与谷歌语义检索实体生成不同的是，书目数据实体来自图书馆质量良好的结构化数据，实体的生成难度要远低于谷歌语义检索。实体生成指根据书目本体生成与书本一一对应的书目描述信息，其本质与图书馆著录工作一致。开放书目数据可以作为实体直接生成数据源。针对各图书馆原有系统MARC格式的书目数据，美国国会图书馆开发了将MARC数据转换为书目本体的工具，并制定了一系列规范。

（3）结果展现。建议针对不同类型实体设计相应网页模板，采用混搭技术嵌入传统关键词检索结果的右侧，以实现对现有书目检索系统的最小入侵。混搭技术是图书馆领域常用的技术之一，其理论研究和实际应用较为成熟，开发成本较低且对原有系统入侵最小。用户书目检索的主要需求由基于关键词的书目检索满足，语义检索只是从更准确的检索结果和更深入的检索拓展两个方面提升用户检索体验。从图书馆开发的成本角度考虑，也应尽可能减少对现有系统的改动。

4 结论

本文首先回顾了图书馆领域书目关联数据在国外的实践状况和在国内的理论研究状况，认为当前书目关联数据的研究基础足够支撑书目语义检索的理论研究和实践探索，书目语义检索也可能成为图书馆关联数据研究直接服务用户的研究成果；其次介绍语义检索的发展背景，重点介绍谷歌语义检索的功能和组成，以此作为语义检索的最佳实践；最后，参照谷歌语义检索的基本组成从本体构建、实体生成和结果展现角度提出书目语义检索系统的初步设计思路。

图书馆书目检索技术的发展随着信息检索技术的发展而不断向前推进。自谷歌、百度等基于关键词的搜索引擎成为用户检索互联网信息的主要入口，关键词检索已经成为用户检索信息的主要工具。图书馆书目检索系统也应跟随谷歌、百度等搜索引擎公司采取关键词检索技术。语义网技术的提出和不断完善，为开展语义检索提供了良好的数据基础和技术基础，经过十多年的发展，语义检索技术已经被谷歌、百度等搜索引擎公司发展成为成熟的产品，为传统关键词检索作补充。随着语义技术的不断发展，语义检索在信息检索领域的地位和作用将更加凸显。图书馆在继续学习谷歌、百度等搜索引擎技术的道路上，应将语义检索技术作为书目检索系统的组成部分，提高用户书目检索效率。

从外部环境看，语义检索已经成为谷歌、百度等搜索引擎的重要组成部分，并逐渐改变用户检索习惯；从内部环境看，图书馆界已经积累充足的关于书目关联数据的研究，成为图书馆研发书目语义检索的基础。在内、外部环境影响下，书目语义检索将成为未来图书馆书目检索的重要组成部分。

参考文献

[1] W3C. Linked Data [EB/OL]. (2006-08-01) [2018-01-30]. https://www.w3.org/wiki/LinkedData.

[2] W3C. Linking Open Data [EB/OL]. (2017-03-12) [2018-01-18]. https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

[3] NIU X, SUN X R, WANG H F, et al. Zhishi.me-weaving Chinese Linking Open Data [C] // The Semantic Web-ISWC 2011: 10th International Semantic Web Conference. Berlin: Springer, 2011: 205-220.

[4] BOLLACKER K D, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge [C] // Proceedings of the 2008 ACM SIGMOD International conference on Management of data, June 10-12, 2008, Vancouver. New York: ACM, 2008: 1247-1250.

[5] AUER S, BIZER C, KOBILAROV G, et al. DBpedia: a nucleus for a web of open data [C] // The Semantic Web. [S.l.]: Springer, 2007: 722-735.

[6] MAHDISOLTANI F, BIEGA J, SUCHANEK F. Yago3: a knowledge base from multilingual wikipedias [C] // Proceedings of the 7th Biennial Conference on Innovative Data Systems

Research，January 4-7，2015. California：CIDR Conference，2015.

[7] Google official blog. Introducing the knowledge graph：things，not strings [EB/OL]．(2012-03-16) [2018-01-30]．https://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html.

[8] W3C. Library Linked Data Incubator Group wiki [EB/OL]．(2013-08-06) [2018-01-30]．https://www.w3.org/2005/Incubator/lld/wiki/Main_Page.

[9] 刘炜，夏翠娟. 书目数据新格式BIBFRAME及其应用 [J]．大学图书馆学报，2014 (1)：5-13.

[10] CHRISTIAN B，TOM H，BERNERS-LEE T，et al. Linked data：the story so far [J]．情報処理，2009，52 (3)：1-22.

[11] 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.

[12] SUMMERS E，ISAAC A，REDDING C，et al. LCSH，SKOS and linked data [C] // Proceedings of International Conference on Dublin Core and Metadata Applications-Metadata for Semantic and Social Applications 22-26 September 2008，Berlin. 2008：25-33.

[13] 李琳. 关联数据在图书馆界的应用与挑战 [J]．图书与情报，2011 (4)：58-61.

[14] MALMSTEN M，李雯静. 将图书馆目录纳入语义万维网 [J]．现代图书情报技术，2009，3 (3)：3-7.

[15] 张海玲. 图书馆书目数据的关联数据化研究——以德国国家图书馆为例 [J]．图书馆论坛，2013，33 (1)：120-125.

[16] GUHA R，MCCOOL R，MILLER E. Semantic search [C] // Proceedings of the 12th International World Wide Web Conference，May 20-24，2003，Budapest. New York：ACM，2003：700-709.

[17] BERNERS-LEE T，HENDLER J，LASSILA O. The semantic web：a new form of web content that is meaningful to computers will unleash a revolution of new possibilities [J]．Scientific American，2001，284 (5)：34-43.

[18] 文坤梅，卢正鼎，孙小林，等. 语义搜索研究综述 [J]．计算机科学，2008，35 (5)：1-4.

[19] 余传明. 语义检索的原理及其实现 [J]．情报理论与实践，2007，30 (2)：182-184.

[20] 何琳，杜慧平，侯汉清. 一种基于领域本体的语义检索系统的设计与实现 [J]．图书情报工作，2008，52 (8)：85-88.

[21] 杨萌，张云中. 知识地图、科学知识图谱和谷歌知识图谱的分歧和交互 [J]．情报理论与实践，2017，40 (5)：122-126.

[22] 李善平，尹奇韡，胡玉杰，等. 本体论研究综述 [J]．计算机研究与发展，2004，41 (7)：1041-1052.

[23] MCCARTHY J. Circumscription—a form of non-monotonic reasoning [J]．Artificial Intelligence，1980，13 (1)：27-39.

[24] 漆桂林，高桓，吴天星. 知识图谱研究进展 [J]．情报工程，2017，3 (1)：4-25.

[25] 任飞亮，沈继坤，孙宾宾，等. 从文本中构建领域本体技术综述 [J]．计算机学报，2017 (5)：1-26.

[26] 中华人民共和国质量监督检验检疫总局，中国国家标准化管理委员会. 文献著录 第1部分：总则 GB/T 3792.1 [S]．北京：国家标准出版社，1983.

[27] 白海燕，乔晓东. 基于本体和关联数据的书目组织语义化研究 [J]．数据分析与知识发现，2010，26 (9)：18-27.

[28] 宋琳琳，李海涛. 大型文献数字化项目图书书目本体的构建研究 [J]．图书馆建设，2013 (12)：19-25.

[29] 郭振英，赵文兵，魏育辉. 轻量级书目本体关联数据建设实践 [J]．现代图书情报技术，2015，31 (7)：139-143.

[30] FLA study group on the functional requirements for bibliographic records. Frbr：functional requirements for bibliographic records [R/OL]．[2018-01-28]．https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf.

[31] 夏翠娟. 面向语义网的书目框架 (BIBFRAME)：功能需求及实现 [J]．大学图书馆学报，2014，32 (6)：61-69.

[32] Library of Congress. Overview of the BIBFRAME 2.0 Model [EB/OL]．(2016-04-21) [2018-01-30]．https://www.loc.gov/bibframe/docs/bibframe2-model.html.

作者简介

邹鼎杰，男，1988年生，博士研究生，讲师，研究方向：图书情报与档案管理，E-mail：13661673864@163.com。

Book Semantic Search Based on Linked Data

ZOU DingJie

(College of Politic, National Defense University, Shanghai 200433, China)

Abstract: This paper introduce book semantic search to improve the efficient of searching book. Firstly, this paper surveyed the situation of linked data in library at home and abroad, and believe that the basement of research on library linked data can support the development of book semantic search. Secondly, this paper introduced the function and construction of Google semantic search, which is the best practice. Lastly, this paper introduced the technical methods to design book semantic search, which could be a reference for library.

Keywords: Linked Data; Semantic Search; Knowledge Graph

（收稿日期：2018-02-01）