基于“语义主题模型”的知识系统框架设计及应用研究*

□ 李颖 / 中国科学技术信息研究所 北京 100038
张毅 / 北京外国语大学 北京 100089

摘要：构建语义化知识服务系统是数字图书馆在语义Web环境下的发展方向。针对目前各种语义Web技术在数字图书馆开发应用的现状和问题，本研究融合国际基于Topic的语义知识组织技术，结合国内的需求，提出了“语义主题模型”，目标为构建可实施的语义化知识系统。文章首先对以“概念化主题、主题关联、语义标签”为内涵的“语义主题化模型”概念进行定义；其次，设计基于此模型的语义化知识系统框架；再次，对相关应用进行了探讨；最后，对未来开发工作进行了规划。

关键词：语义主题模型，知识系统框架，主题
DOI：10.3772/j.issn.1673—2286.2012.04.005

背景及研究动机

主题（Topic）与图书馆学的主题（Subject）为同义词，本研究侧重人脑的概念主题在计算机上的实现，故采用Topic这一术语。由于主题直接表现概念，主题著录和主题元数据描述、主题索引和导航是图书馆信息学永恒的研究课题。针对全球性的知识服务需求，面对语义Web技术所带来的机遇与挑战，如何实现传统主题理论的继承与升华？相关研究和技术在国内外日趋凸显。国际上，从上世纪90年代起，Topic和Ontology等与主题相关的概念进入了知识系统。主题图（Topic Maps）和DITA（达尔文信息类型化体系结构）、资源描述框架（RDF）、关联数据（Linked Data），以及资源描述与检索（RDA）等国际标准技术，都深化了主题的理念。某种程度上，这些国际标准技术满足了用户对资源的知识获取需求，也满足了知识重用的客观环境需求。然而，这些技术的焦点主要集中于从元层来描述形形色色的资源，很少将资源自身的结构化构建或重建纳入系统的整体框架设计之中，很难真正满足用户意图检索的需求。可以想象，任何强大的语义技术和自然语言处理技术，都不可能将无序化的海量资源处理为理想的结构化知识系统。“Garbage In, Garbage out”，为获取关联的知识，系统入口的底层资源本身的结构化处理必不可缺。尽管针对期刊、专利和图书等各种资源，有基于XML Schema的各种结构化标准，但它们没有在对象的粒度、语义主题化组织及主题关联上进行系统的考虑。而这些是知识的主要属性，不能割裂地考虑。为此，本研究提出了“语义主题模型”的概念（详见第1部分内容），就是为了弥补该方向的研究空白，用以构建全新的知识系统。在国内，有关知识系统的构建研究，跟踪和效仿国际领先技术的综述较多，实用化大规模的知识服务系统很少。尤其在面向企业创新需求的知识系统构建方面，几乎没有成型的架构体系来指引企业的知识系统构建。比起底层资源和用户导航层的构建，研究多集中在元数据层，从系统的角度来看，不是一种协同发展的良好模式。

本研究的动机如下：

理论层面：“语义主题模型” 是针对构建知识系

* 基金项目：“十二五科技支撑计划——科技知识组织体系共享服务平台建设”资金支持（编号：2011BAH10B03—2）；中国科学技术信息研究所“汉语科技词系统建设与应用工程”重点工作和国家科技支撑计划“面向外文科技文献信息的超级科技词表和本体建设”子任务支持。

统而首次提出的数据模型，是传统理论在语义环境下的延伸，可适用任何粒度和任何层面的对象，比如资源底层、元数据层及用户导航层。它没有领域的限制，适用范围广。

应用层面：文献信息系统的发展方向为知识系统。由于其理论的复杂性、劳动密集型特征以及技术工具支持的滞后，理想的知识系统在我国极其匮乏。本课题以满足用户意图检索和资源重用为导向，研究通用的知识系统框架及应用开发的技术路线，具有现实的意义。基于国际标准技术，可采用敏捷式开发，门槛低，系统可以不断地扩展和进化。

1 “语义主题模型”的定义

同其他Web系统目标一致，提供满足用户检索意图的服务，是数字图书馆应当具备的主要功能。本研究以经典主题理论和本体理论为基础，研究语义Web技术、主题图技术和DITA技术等知识系统的相关技术，分析其理论与功能的局限性，提出了以“主题化描述、语义主题标注、主题关联”为中心内涵的“语义主题模型”的概念，设计基于此概念模型的新型知识系统。

“语义主题模型”（Semantic Topicized Model）面向构建新型的知识系统，拟从系统的整体框架设计到面向知识服务的应用和示范进行全面的研究。“语义主题模型”的内涵是对任意粒度和任意层面（参见图1）的对象进行“主题化描述（topicized-description）；语义的主题标注（semantic topic-tagging）；主题关联（topic-associating）”处理，从而实现完整性知识系统的构建。基于“语义主题模型”的概念，本课题将统一对文献信息系统的各个层面进行语义主题化组织，并针对不同层面的重点问题，进行具体的实施应用。也就是说，①对底层信息资源（resource layer）进行基于语义主题的组织构建；②对资源的元数据层（meta layer）进行语义化的主题元数据描述；③在用户访问层（access layer）上建立语义主题关联的知识导航，以填补数字图书馆技术在文献信息系统的语义主题化资的源构建上空白，同时通过语义主题元数据描述和基于语义主题关联导航来增强系统的功能，提高文献信息系统整体上的语义整合度，最终满足用户的意图检索（intended research）、实现信息资源的主题构建和高度重用，对文献信息系统向知识系统的迈进、文献信息的知识化服务提供一种可实施的技术路线。

2 基于“语义主题模型”的知识系统框架设计

基于语义主题模型知识服务系统框架设计如图1所示。

本框架主要包含有三层：语义主题化资源层、语义主题元数据层、主题关联知识导航层。

图1 基于语义主题模型的知识系统框架

（*：包括词系统、领域本体、自动分类、自动标签工具、知识模块加工工具等）

如图1所示，融合现有的各种知识工具的基础上，在三个层面实施基于“语义主题模型”的知识系统的具体设计。三个层次具体设计任务如下：

1) 语义主题化资源层：研究文献信息资源本身的结构特征、现有文献XML Schema的局限性。研究如何基于“语义主题模型”对资源进行构建或重组。

2) 语义主题元数据层：研究和分析现有各种元数据体系，尤其是主题元数据。研究如何基于“语义主题模型”进行语义主题元数据的构建。

3) 语义主题关联知识导航层：为满足知识导航和获取，研究现有导航体系。研究如何构建语义主题关联的知识导航图。

3 基于“语义主题模型”知识系统的应用探索

本研究设计目标是针对知识系统的开发应用，是传统和当代Subject理论与Topic语义技术的延伸，又融入了语义Web技术和内容知识管理等技术。可适用任何

粒度和任何层面的对象，比如资源底层、元数据层及用户导航层。理论上，只要是知识系统，都可适用。

基于目前我们在知识工具上的积累以及国家重点支持的科技创新任务的需求，我们将重点在金融和医药领域进行应用探索。将基于图2所示的流程实施应用。

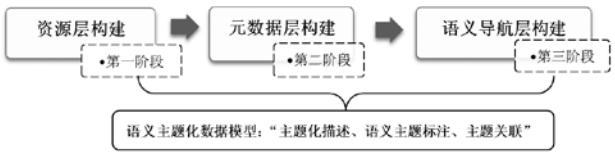

图2 基于"语义主题模型"知识系统的应用实施步骤

基于"语义主题模型"金融和医药知识服务系统应用研讨的主要任务分解如下：

1) 基于语义主题模型知识服务系统的架构设计

在设计针对金融和医药知识服务系统的架构时，需要传承主题理论的精华，融合数字图书馆的高端技术、语义Web技术、自然语言处理技术。既要考虑不同领域与不同资源的共性与特性，还要符合国际标准，以此实现适应多领域、多类型资源和多语种资源的知识服务体系。因为知识是多领域构成的，多类型格式及多语种也是知识的特征。

2) 基于语义主题模型知识服务系统的实施技术路线

我们的最终目标是为了构建提供用户知识服务的系统，要设计出实施启动成本小，并可不断丰富和进化的系统。为此，规划可实施的技术路线对未来应用至关重要。

3) 基于语义主题模型知识服务系统的原型开发

选择金融和医药领域的部分数据作为示范应用的对象，在三个层面上应用语义主题模型，开发基于"语义主题模型"的原型知识系统。

4) 基于语义主题模型知识服务系统的实用化推广研究

评估原型系统，在此基础上，研究如何将"语义主题模型"概念应用于实用化的文献信息系统，研究多种资源类型和多语种资源的语义主题化组织，构建全球性通用知识系统。

抽象了上述基于"语义主题模型"知识服务系统的主要任务之后，本研究将首先以金融知识服务系统为例，构建具有如下功能模块的应用体系，见图3所示：

图3 "语义主题模型"在金融行业的应用

此体系最大优势是：

◆ 通过将原始资源的主题化创作或重组，满足了产品制作的组件化和重用的基本原则，可提高知识产品的生产力和质量；

◆ 通过语义主题元数据的描述，使用户容易发现关联知识，维护者也容易管理；

◆ 通过语义关联导航，用户可以跟踪知识路径。良构的知识图可以帮助用户快速发现所需信息。该部分模块相当于知识的GPS，可实现基于用户知识需求驱动的系统。

"语义主题模型"是基于大型现代化企业产品的组件化和重用的理念，对提升信息知识产业的生产力和质量，具有长远的意义。通过在金融领域的先导型应用，对其他行业将起到应用示范作用。之后，本课题将在另一个有关民计民生的行业，即医药，展开大规模的实用开发。

4 未来的开发规划

首先，对重点工作进行全面系统的调研，具体如下：

1) 有关语义主题模型知识服务系统框架的优化；

2) 有关语义主题模型知识服务系统实施技术路线的细化和具体化。

其次，攻关难点问题：

1) 如何基于语义主题模型组织信息资源为劳动密集型工作，需要自动化辅助处理技术，而现有的知识工具还很不成熟，需要统筹考虑。

2) 如何基于语义主题模型，来优化主题元数据，需要考虑的因素很多，包括传统各种元数据体系的继承与协同。

3) 在设计知识图时，如何将用户的目标检索需求与使语义化主题关联图建立映射，是一个涉及众多研究领域的问题，还很不成熟。本研究需要做深入的研究。

再次，开发基于语义主题模型医药知识服务系统的原型。

总结

本研究提出了以下主要观点：

1) “语义主题模型”提出，此模型具有“概念化主题、主题的语义标签、主题关联”的内涵。

2) 基于“语义主题模型”的知识系统。该系统将语义主题模型的理念，贯穿于资源层、元数据层及检索导航层等系统的三个层面，是一个完整的知识系统。

主要创新之处为：

1）“语义主题模型”为一种新型的知识系统构建的数据模型。

2）将知识系统分为三个层次处理，在不同层次上，都可以有效地应用语义主题化组织的概念，整合度高。

最后，如何通过医药和金融领域的开发应用来验证和评价我们的模型？如何持续该项研究？知识系统是人类对知识探究的一种手段，知识探索没有终结，希望我们的研究也将一直持续下去。

参考文献

[1] Topic Maps [OL]. [2012-02-12]. http://www.topicmaps.org/.
[2] DITA [OL]. [2012-02-12]. http://dita.xml.org/.
[3] Linked Data. Linked Data, TBL [OL].[2012-02-12]. http://www.w3.org/DesignIssues/LinkedData.html.
[4] RDA [OL]. [2012-02-12]. http://www.rdatoolkit.org/.
[5] LI Y, ISHIZUKA H. A Metadata Model Based on the Concept of Structured Digital Object (SDO) and Its Application in Digital Libraries - From Concept to Prototype System [C]// Proceedings of the International Conference on Dublin Core and Metadata Applications, 2004.

作者简介

李颖，信息系统专业博士。相关研究课题：语义知识组织、基于主题的知识组织技术的应用等。E-mail: liying@istic.ac.cn
张毅，北京外国语大学信息技术中心。

Design of a New Knowledge System Framework and Its Application Study Based on the Concept of "Semantic Topicized Model"

Li Ying / Institute of Scientific and Technical Information of China, Beijing, 100038
Zhang Yi / Beijing Foreign Studies University, Beijing, 100089

Abstract: Building a semantic knowledge service system is a new direction for digital libraries in the Semantic Web environment. According to the present situation and problems of various Semantic Web technologies in the digital library development and application, this study merges the international semantic knowledge organization technologies based on topic.Combined with domestic demand, it defines a concept of "Semantic Topicized Model", in order to build a semantic knowledge system that can be implemented. Firstly, this paper gives the concept of "Semantic Topicized Model", with the intensionof" topicized-description, semantic topic-tagging, topic-associating." Then,it designs of a semantic knowledge system framework based on this model. Finally, it discusses the relevant applications and planning of the future development works.

Keywords: Semantic topicized model, Knowledge system framework, Topic, Subject

（收稿日期：2012-03-15）