地方文化特色数据库语义检索模型构建*

王智刚
（鄂阳师范高等专科学校图书馆，十堰 442000）

摘要：地方文化特色数据反映了某一特定区域独有的文化信息资源，众多信息情报部门陆续建立富有地方特色的资源数据库。本文中笔者以教育部calis中心三期特色数据库“武当文化特色数据库”建设为例，将语义网技术运用其中，通过分析语义网关键技术，构建领域本体实例，提出了一个基于语义检索技术的特色数据库检索模型，最后对模型实现中的关键技术做了具体阐述。

关键词：特色数据库；语义检索；本体；武当文化
中图分类号：G354.4
DOI：10.3772/j.issn.1673—2286.2015.02.009

引言

地方文化特色数据库，是以挖掘地方特有信息资源以便长久保存，服务广大师生和促进地方文化建设、经济发展为目的，以自建为主的电子资源数据库。各部属院校以及地方院校图书馆积极参与建设的数据库资源极其丰富，本科院校基本上都有自建的特色数据库，calis三期一次承建特色库就达235个[1]，笔者所在高校也参与建设了具有地方文化特色的《武当文化专题特色数据库》。

然而，在该数据库建设的过程中发现，传统的数据库检索都基于关键词，这种机械匹配忽略了词与词之间的语义关系，查准率和查全率都很低，很难满足用户的需求。人们对此问题进行分析与研究，提出了两个改进的方向：一是对检索技术和服务进行改良；二是将数据库的信息资源标注成计算机能够理解的内容。1998年，互联网创始人蒂姆•伯纳斯-李(Tim Berners-Lee)提出了“语义网（Semantic Web）”概念，它的目的是通过给数据添加能够被计算机理解的语义，从而使人机交互变得更有效率。目前，国内外对语义检索的研究，多是对信息资源加入语义结构化标注，将基于关键词的检索提升到基于概念级的检索，而对用户查询的语义解读，以及对检索结果排序方面则无太多成果，在特色数据库的语义网建设方面更是鲜有研究成果。

本文通过研究武当文化领域本体构建、语义标注、查询请求语义扩展算法以及查询结果算法排序等，提出了一种基于领域本体的特色数据库语义检索模型，以提高武当文化研究领域信息检索的查准率和查全率，为地方文化特色数据库语义网建设提供参考。

1 语义检索技术概述

1.1 语义检索研究现状

语义检索，是为了生成更相关的结果，使用语义网络中的数据来帮助区分查询和网页的内容，所进行的在线检索过程[2]。最早的语义检索技术研究，是

* 本研究得到教育部“211工程”高等教育文献保障系统（CALIS）三期“专题特色数据库”项目（编号：4401-HUB-507）和鄂阳师范高等专科学校校立科研项目“基于语义网技术的武当文化特色库检索模型研究”（编号：2012B06）资助。

Voorhees 于1994年提出的基于本体的扩展查询技术，而后出现了基于本体结构的扩展查询、基于本体注释的扩展查询方法等[3]。近些年来，一些IT公司相继开发了语义搜索引擎，如Yahoo公司开发的Search Monkey，微软公司的Powerset，谷歌公司的图谱（Knowledge graph）等[4]，这些大型商业公司进入语义搜索领域，表明语义搜索有着巨大的市场。以上关于语义检索方面的研究，要么只是在本体的概念层次基础上进行的，要么就是只对信息资源进行语义处理，而商业公司的语义搜索引擎，都是基于自然语言的，强化的是搜索技术，对本体知识与检索过程的结合工作略显不足。目前的语义检索研究范围很广，但在专业特色数据库语义检索技术方面，鲜有研究成果。

1.2 语义检索系统的关键技术

语义网体系结构中，XML、RDF和Ontology被认为是三大核心技术，是目前语义网的研究热点。在语义检索研究领域，本体、语义相似度、查询扩展和语义标注是研究热点、重点和难点[5]。

1.2.1 本体

本体（Ontology）是源自哲学领域的一个概念，即“存在论”。在信息科学领域，本体是指一种形式化的，对共享概念体系明确而又详细的说明，它是对特定领域之中某概念及其相互之间关系的形式化表达[6]。本体是语义网的核心技术之一，在语义网的信息组织过程中，负责描述概念及概念之间的关系，因此它也是语义检索系统的关键技术之一。

1.2.2 语义相似度计算

传统的检索技术，是基于关键词的机械式匹配，其相似度计算只针对关键词。而语义相似度计算，是在概念层面上，对源和目标词汇进行相似度匹配，需要考虑语境和语义等信息[7]。语义检索中引入相似度计算，有助于提高检索准确率和检索效率。基于本体的语义相似度算法主要有：基于距离的语义相似度计算（Edge Counting Measures），基于内容的语义相似度计算（Information Content Measures），基于属性的相似度计算（Feature-based Measures），和混合式语义相似度计算（Hybrid Measures）。

1.2.3 查询扩展

在检索过程中，运用计算机语言学和信息学等多种学科技术，将与检索词相关的词语、语义概念扩充为新的检索词，以提高查全率和查准率的方法，称之为查询扩展。而查询扩展技术，指的是实现查询扩展的方法和手段。检索词的扩展来源主要有三种：一是来自于初检中被认为相关的词；二是用某种技术，如聚类技术、文本挖掘技术等，从文献集或查询日志中找出与原查询相关的语词作为扩展词；三是来自某种包含词与词间相关信息的资源，这种资源可以是人工生成的，也可以是利用大规模语料通过统计的方法自动生成[8]。

1.2.4 语义标注

语义标注就是将实例与本体的概念相联系的过程，它用本体对网页数据进行标引，让动态变化中的实例与本体结合在一起，使网页实现智能化。语义标注工作类似于给数据库添加记录，通过对普通网页文档进行语义标注，把网页文档信息和推理规则联系起来，使计算机根据本体得到信息的含义，从而找到精准的结果。目前，语义标注主要有三类[5]：一是人工标注；二是利DTD（文献类型定义）和Schema（文献模式）进行概念集映射和标注；三是利用词汇语义分析进行标注。

2 武当文化领域本体构建

构建武当文化领域的语义检索模型，第一步，就是需要对武当文化领域的知识、概念、术语、具体实例等进行组织，也就是完成武当文化领域本体知识建模。类、关系、属性和实例是本体构成四要素[9]，图1是本体要素构成图。

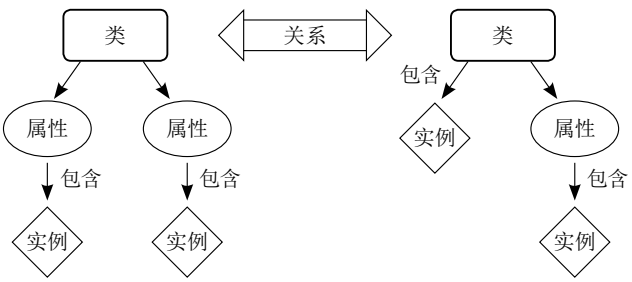

图1 本体要素构成

2.1 类的获取

类，即概念、对象类型，它描述了具有共同特征或某些方面相似的资源的集合。武当文化领域本体中的概念，主要指武当文化研究领域的抽象化定义和术语。在武当文化领域中，“武当”、“人物”、“单位组织”是本体领域的核心概念，也是领域中最顶层的概念。武当文化领域本体概念的获取，主要从这三个概念展开。

“武当”是本领域的核心概念，从这个顶层概念向下展开，以获取下级概念。从内容上来展开，主要有“武当道教”、“武当武术”、“武当医药”、“武当养生”、“武当旅游”、“武当建筑”等；从文化艺术形式上来展开，主要有“武当碑刻”、“武当书画”、“武当音乐”、“武当雕塑”、“武当传说”等；从文化载体上来展开，主要有“武当书刊”、“武当图片”、“武当视频”、“武当网源”等。图2，是以“武当”为顶层概念的树状概念分枝图的一部分。

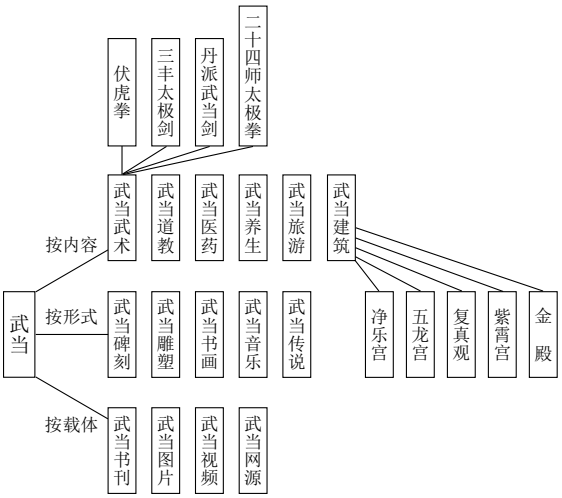

图2 树桩概念分支图示例

使用同样的方法，可以从另外两个核心词汇“人物”、“单位组织”入手，将其作为顶端概念，向下逐步展开，进一步获取下级概念，直至将武当文化领域本体的所有概念抽取出来。

2.2 关系抽取

概念获取完毕之后，就要对概念之间的关系进行抽取。因为有关系的描述，就能将一个个独立的概念串联起来，以构成领域知识网状结构。关系是领域本体中的脉络，负责支撑领域本体知识机构。关系的抽取，可以从顶层概念入手，采取从上而下，层层递进的方法，获取概念之间的关系。关系大体上可以分为自然关系和构建关系，在获取概念的过程中，概念层次之间存在一种自然的父子关系，这种关系称之为自然关系。自然关系主要包括：继承关系、整体与部分的关系、属性关系、实例关系等。同时，在不同的核心概念之间事实上存在着某种联系，领域专家将这些关系抽取出来，再对这些关系进行定义或构建，这种关系称之为构建关系。如“武当人物”与“武当武术”这两个核心概念之间就存在着一种创造关系，这就是构建关系的一种。在武当文化领域研究专家的帮助下，我们依照此方法抽取了领域本体的各种关系，图3是武当文化领域核心概念与概念之间关系图的一部分。

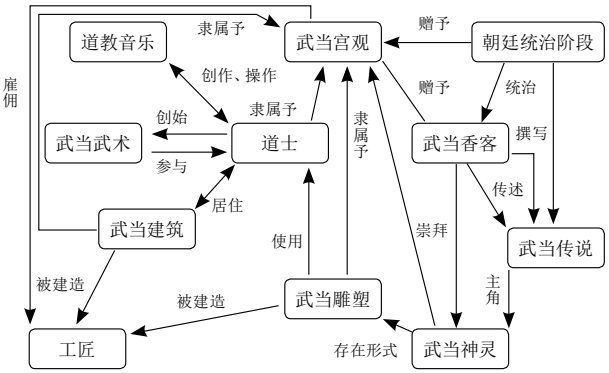

图3 概念间关系示例

2.3 属性填充

获取武当文化领域本体概念和抽取概念之间的关系之后，就构筑了该领域本体知识模型的顶层框架，而接下来的任务，就是将这些概念和关系用具体化的知识以属性和实例的方式进行填充。属性处于本体层次结构的下层，用以描述顶层框架中抽象知识的具体化扩展，它具有传递性，子概念的属性一般继承父概念的属性。在属性的填充过程中，常常会发现一些新的概念，本体中加入这些新概念之后，概念之间的关系也随之加入，然后新的概念属性也被提取出来，重新填入本体之中，整个过程是在迭代填充中完成的。表1是武当文化领域概念及其属性关系的一个具体实例。

2.4 实例构建

概念、关系、属性构成武当文化领域知识中的骨架

和脉络，接下来就需要按照这一架构，将领域中庞大的具体信息资源用本体知识实例加以描述。实例填充是整个武当文化领域本体构建过程中工作量最大的步骤。领域信息资源的收集是构建本体实例的基础，信息收集可采取人工收集和利用软件从网络智能抓取相结合的方法。而信息资源的分析和整理有一定的主观性，因此，邀请武当文化领域专家参与实例构建也是构建本体的基础。

表1 概念和属性举例

父概念	子概念	属性
武当	武当书刊	著者 书名 出版年 出版社 ISBN
	武当武术	名称 流派 创始人 武术器械
人物	神仙	尊称 别称 神像 供奉宫观
	道士	性别 尊号 生卒年 特长

2.5 利用本体编辑器构建本体

以上是本体的构建过程，基本上涵盖了武当文化领域本体构建所有步骤和要点，但是在实际构建过程中，一般会用到一种本体编辑器，作为构建本体的平台，实现可视化操作。本体编辑器，是指设计用于辅助本体之创建、编辑、保存或处理等操作的应用程序[10]，它提供本体编辑和管理的环境。目前，被广泛使用的本体编辑器主要有语义火鸡Semantic Turkey、Swoop、OBO-Edit、Protégé，其中Protégé是美国斯坦福大学开发的基于Java的开源软件，在国内已得到广泛应用。图4是我们用Protégé构建武当文化领域本体过程中，构建类和抽取关系的截图。

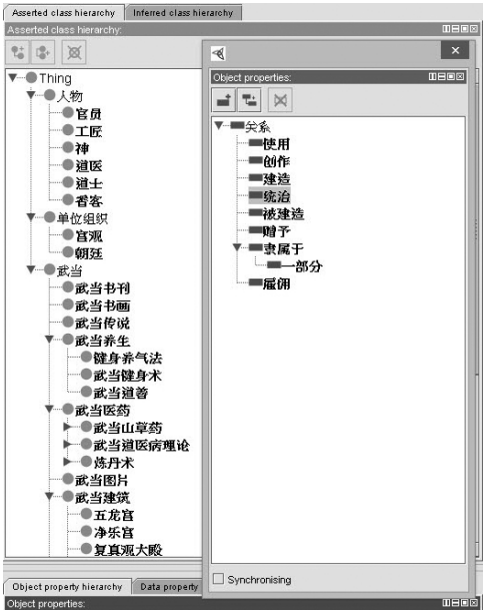

图4 类和抽取关系截图

3 武当文化语义检索系统模型构建与实现

3.1 武当文化语义检索系统模型构建

基于以上对语义检索和武当文化领域本体的构建研究，我们设计了一个武当文化语义检索系统模型。如图5所示，该模型由人机交互层、逻辑层和数据层所组成。人机交互层是本检索系统与用户的接口，负责接受用户的查询请求，以及将结果呈现给用户；逻辑层是本系统的核心层，负责查询请求的语义扩展，和对检索结果排序等工作；数据层是本系统的基础层，它包含武当文化领域本体知识库、问题库和语义相似度计算，系统会将用户每次检索的问题搜集起来，建立一个问题库，其他用户检索同样问题时，就直接从问题库抽取结果。

3.2 语义检索系统模型的实现

武当文化语义检索系统的模型的实现，除了本体知识库的构建外，还涉及到语义查询扩展、语义相似度计算和结果排序等几个关键过程。

3.3.1 查询扩展

语义检索的第一步，就是对用户的查询请求进行语

义解读，即查询语义扩展，用来扩展一组语义相关的概念集合，再进行进一步检索。目前常用的语义查询扩展技术主要有：全局分析、局部分析、基于关联规则和基于用户查询日志的查询扩展技术。其中，基于关联规则的查询扩展，是通过数据挖掘技术挖掘词间关联规则，然后将结论作为扩展词的来源。本系统采用了以局部分析为出发点，对初检查询文件得到的局部文档进行关联规则挖掘，以得到查询扩展概念的方法，算法流程图如图6所示。

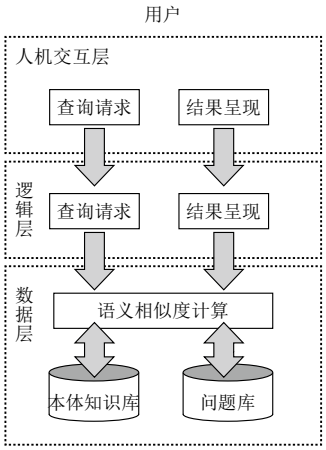

图5 语义检索系统模型

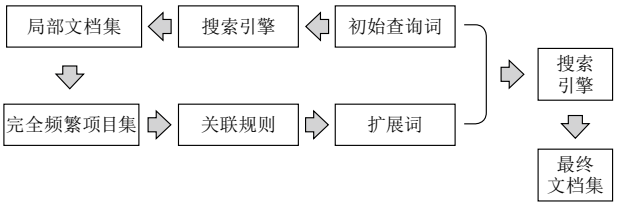

图6 查询扩展流程

3.3.2 语义相似度计算

语义检索系统将传统的基于关键词的检索，提升到了概念级匹配，也就是计算词语之间的语义相似度。语义相似度计算，是对源和目标词语在概念层面上相似程度的度量，需要考虑词语所在的语境和语义等信息[7]。国内外被广泛使用的语义相似度计算算法主要有4种，武当文化语义检索系统模型，采用了混合式语义相似度计算方法，算法模型为：

$$Sim(a^p, b^q)=W_wS_w(a^p, b^q)+W_uS_u(a^p, b^q)+W_nS_n(a^p, b^q) \quad (1)$$

式中，W_w表示同义词集，W_u表示特征项，W_n表示语义邻节点相似度对整体相似度的贡献大小，并且$W_w+W_u+W_n=1$。a^p表示概念词a属于本体p，b^q表示概念词b属于本体q，而函数S_w、S_u、S_n用来计算概念词a和b的同义词集、特征项、语义邻节点相似度。武当文化语义检索模型中，通过语义相似度计算后，可解决概念级匹配的问题，从而在语义理解的基础上，从本体库和问题库抽取检索结果。

3.3.3 结果排序

通过查询语义扩展以及语义相似度计算等语义检索技术处理之后，所得到的查询结果已经非常接近用户查询的目的。在武当文化语义检索系统模型框架内，接下来将对查询结果，结合原始查询请求和扩展查询请求，使用基于权重分析的结果排序算法排序，进一步提高检索质量。公式（2）是一种基于权重的结果排序算法模型：

$$WS_i^n = ae^{-\beta(n-1)} + \frac{R_i^{N-1}}{\sum_{j=1}^{M} R_j^{N-1}} \quad (2)$$

检索过程中，如果 R_i^N 是用户执行的第n次搜索请求，则遍历的搜索结果中来自搜索平台的结果数量为 $R_i^N = R_i^1 + R_i^2 + \cdots + R_i^{n-1}$。式中α和β为初始可调因子，$\sum_{j=1}^{M} R_j^N$ 为检索n次之后，用户遍历的搜索结果总数，WS_i^n 即为检索结果的权重。通过对检索结果进行权重计算之后，然后依据权重对检索结果进行排序，将最接近用户要求的结果呈现在用户面前。

4 结语

语义网技术作为Web3.0的核心技术之一，已经被越来越多地应用于互联网和数据库行业，各图书馆以及信息情报部门建设的特色数据库，也很有必要适应这一趋势。本文所论述的武当文化特色数据库检索模型构建，结合了语义网中领域本体构建、查询请求语义扩展、语义相似度计算以及查询结果排序等技术方法，有别于单一的语义检索技术手段，提升了语义检索效果，也为特色数据库语义检索模型建设提供了一定的参考价值。然而，语义检索的质量评价决定了其结果没有最好，而只有更好，所以对于语义检索系统模型的探讨，还会不断出现创新的模型构造。而且，基于语义检索的特色数据库建设，决不是某个领域技术和专家能够解决的问题，因此，还需要从优化领域专家团队建设角度，加强领域

之间的合作，来提升语义检索模型的检索结果质量。

参考文献

[1] CALIS.CALIS三期专题特色库立项评审结果通知[EB/OL].[2011-5-12]. http://www.calis.edu.cn/educhina/viewnews.do?newsid=49

[2] 维基百科.语义检索 [EB/OL].[2011-5-12].http:// http://zh.wikipedia.org/wiki/语义检索

[3] Navigli R,Velardi P,An analysis of ontology-based query expansion strategies[C]. In Workshop on Adaptive Text Extraction and Mining (ATEM 2003), in the 14th European Conference on Machine Learning(ECML 2003)

[4] 郭卫宁,司莉. 国外语义搜索引擎调查与分析[J]. 图书情报工作, 2013,57(23):121-129.

[5] 亢丽芸,王效岳,白如江. 国内语义检索研究计量分析[J]. 现代情报, 2012,32(5):104-109.

[6] 维基百科.本体 [EB/OL]. [2015-1-21].http://zh.wikipedia.org/wiki/本体

[7] 刘宏哲, 须德.基于本体的语义相似度计算方法研究综述[J].计算机科学,2012,39(2):51-56.

[8] 黄名选,严小卫,张师超. 查询扩展技术进展与展望[J]. 计算机应用与软件,2007,24(11):1-4,8.

[9] 维基百科.本体构成要素 [EB/OL].[2015-1-21]. http://zh.wikipedia.org/wiki/本体构成要素

[10] 维基百科.本体编辑器 [EB/OL].[2015-1-21]. http://zh.wikipedia.org/wiki/本体编辑器

作者简介

王智刚，男，网络工程师，馆员。

Local Cultural Characteristics Database Semantic Retrieval Model Building

WANG ZhiGang
（Yunyang Teachers′ College, Shiyan 442000, China）

Abstract: Data reflect local cultural characteristics unique to a particular area of cultural information resources, information and intelligence departments have established numerous local characteristics of the resource database. In this paper, the author of the Ministry of Education calis center Phase III Database "Wudang Culture Database", for example, the use of semantic web technology which, by analysis of the semantic gateway key technology to build domain ontology instance, proposed a semantic retrieval technology based on the characteristics of database retrieval model, the final model to achieve the key technologies specifically addressed.

keyword: Features Database; Semantic Retrieval; Ontology; Wudang Culture

（收稿日期：2014-01-26）
编辑：王立学