文章编号:1000-5862(2019)06-0655-06

MOOC 统一检索平台的自动构建与应用

肖清泉,李云清*
(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要:随着 MOOC 平台增多以及同一平台下学习资源剧增,如何实现跨平台的高效语义检索成为目前 MOOC 亟待解决的问题之一. 该文通过网络爬虫工具获取多个知名 MOOC 平台的学习资源数据,进行相关预处理后存储到 Mysql 数据库,并根据数据库与本体之间的映射关系自动构建 MOOC 本体,使用 Jena 将 MOOC 本体解析成 RDF 3 元组,并将 RDF 3 元组存储至 HBase 数据库,最终构建出一个 MOOC 统一检索平台,为学习者推荐符合其检索需求的学习资源. 实验结果表明: 构建的 MOOC 统一检索平台可有效地提高检索的查准率和查全率.

关键词:统一检索平台; 自动构建; 本体; HBase 数据库; 大规模在线开放课程
中图分类号:TP 311 **文献标志码**:A **DOI**: 10.16357/j.cnki.issn1000-5862.2019.06.16

0 引言

MOOC(Massive Open Online Courses)即“大规模在线开放课程”,以开放、共享的独特优势促使教育机会和教育公平逐渐成为现实. MOOC 突破了传统教学的课堂地域和时间的限制,使得教育不再受时空的限制,丰富了学习者的学习方式. 其中学堂在线、中国大学 MOOC、Coursera、Edx 和 Udacity 等 MOOC 平台已受到众多学习者的认同与赞赏. MOOC 颠覆了传统的教育模式,产生了新的教育理念. 教学资源的“精品化”和“精细化”是 MOOC 的优势之一,MOOC 的教学资源一般是由该领域最优秀的教师来进行内容的编排和设计,较传统课程而言,MOOC 平台上的课程教学质量更能得到保障.

随着 MOOC 的发展,MOOC 平台日益增多、课程资源也迅速增长,导致学习者难以在众多 MOOC 平台中快速获取符合其偏好与学习兴趣的课程. 如何从众多的 MOOC 平台上快速获取符合学习者需求的课程资源信息成为一个亟待解决的问题.

语义 MOOC 亦称 LOOC(Linked Open Online Courses),即“关联开放在线课程”[1],在 2014 年由 Michael Hover 与 Max Muhlhause 首次提出. 语义 MOOC 借助本体中的语义网技术,诸如 RDF(资源描述框架)、OWL(Web 本体语言)、SPARQL(简单 RDF 查询语言)等[2],对 MOOC 上的学习资源进行语义化表示,以语义理解和语义推理为基础实现 MOOC 的语义检索. 语义 MOOC 最大的特点是通过语义化表示各个 MOOC 平台上的学习资源,使得这些 MOOC 平台上的学习资源关联起来,实现跨平台的精准化和智能化的课程检索. S. Dietze 等[3]提出将 Web 上的课程资源关联起来的思想,即将 Web 上的课程资源进行关联从而达到课程资源的开放和共享的目的. 文献[4-5]提出借助本体工具构建学习资源本体,使得教育资源共享并开展语义检索和学习资源标注. 由于其采用的是人工构建学科本体的方式,所以随着本体复杂度增加相应的人工构建本体的成本也增加. 丁国柱等[6]提出使用本体构建学习对象元数据知识库,虽然有几种数据模型用于构建教育数据以及采用这些模型的知识库,但是关联数据驱动的教育应用仍然寥寥无几. 也有学者提出一种学习资源语义关联,借助学习资源丰富的关联关系实现学习资源的语义化,为学习者进行智能化检索[7],但随着学习资源不断增加,学习资源的存储和检索效率降低.

将各个 MOOC 平台的学习资源语义关联化并构建一个 MOOC 统一检索平台,可以实现在多个 MOOC 平台之间跨平台检索,能更好地满足学习者的检索需求,促进 MOOC 的发展与应用.

收稿日期:2019-01-19
基金项目:国家自然科学基金(61877031)资助项目.
通信作者:李云清(1964-),男,江西南昌人,教授,主要从事本体及知识工程等研究. E-mail: 1377832870@qq.com

1 MOOC 统一检索平台的自动构建

构建一个 MOOC 统一检索平台,使学习者可以在众多 MOOC 平台上快速准确地检索,更好地满足用户体验. 本文采用本体技术将各个信息源 MOOC 平台上的学习资源进行语义化表示,应用 HBase 数据库实现 MOOC 统一检索平台的本体存储,便于实现语义查询,最终构建出一个 MOOC 统一检索平台. MOOC 统一检索平台与各个信息源 MOOC 平台关系如图 1 所示.

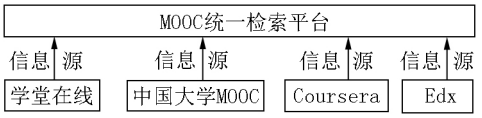

图 1 MOOC 统一检索平台与现有 MOOC 平台关系

MOOC 统一检索平台通过采集有关信息源 MOOC 平台上的学习资源的基本信息并进行语义表示,提供服务接口,用户在该平台上进行检索,即可以检索到该平台下所集成的各个信息源 MOOC 平台的学习资源.

1.1 本体和 HBase

1.1.1 本体 本体是一种形式化的,对于共享概念体系既明确又有详细的说明,由概念、概念属性以及属性间的关系构成[9]. 本体作为一种语义知识表示的方法,用形式化的语言表示各种资源,将解决问题的知识存储在计算机中,以便计算机处理,促使各个应用平台能够精确智能地处理网页中的资源. SPARQL 作为本体查询语言已被 W3C 列为推荐标准,它支持多种数据格式的检索. 运用最为广泛的是 RDF 3 元组形式,其采用基本图模式进行查询方式. 因此,对于各个 MOOC 平台的学习资源数据可以采用本体进行表示,根据数据表与本体之间的映射规则,将各个 MOOC 平台的学习资源数据表自动映射成本体文件,进而完成 MOOC 本体的自动构建. 对于 MOOC 本体的语义检索,可采用 SPARQL 查询语言进行语义检索,其检索出的结果符合学习者的查询偏好,从而为学习者提供更加优质的检索服务.

1.1.2 Hbase 对于领域本体的存储,现有的领域本体存储方式主要有文本存储和关系型数据库存储. 文本存储方式操作简单,但是存储效率较低而且存储规模受限. 关系型数据库存储技术已经发展成熟,但是这种存储方式可扩展性差、使用成本高、对数据的操作效率低,若需求增加或更改,则需要对数据库进行重构,不利于数据库的横向扩展和维护,特别是当表与表之间存在互相关联时,更改表结构十分繁琐. 由于 MOOC 本体是随着现有 MOOC 平台的学习资源数据的变化而改变的,所以对于 MOOC 领域本体的存储和查询应采用一种非关系型数据库,以适应在 MOOC 数据量不断增加的情况下仍能够实现学习资源的高效存储和实时查询,而 HBase 是最佳选择.

HBase 是一个构建在 HDFS(Hadoop 分布式文件系统)上的分布式、面向列的 NoSQL(不仅仅是 SQL)数据库[9],类似于 Google 文件系统的 Big-Table,可建立在 Hadoop 集群平台上实现海量数据的高效存储和实时查询操作. 陆婷等[10]设计出一个基于 HBase 的交通流数据实时存储系统,设计的实时存储系统的存储性能比 Oracle 实时存储系统更高. 针对如何将领域本体存储到 HBase 数据库中,A. Chebotko 等[11]提出使用 S_PO、P_SO、O_SP、PS_O、SO_P、PO_S 6 张 HBase 表来存储数据,以满足所有的 RDF 3 元组查询. 该方法可以高效匹配所有的 RDF 3 元组,但是却额外增加了存储空间. 在 J. Sunp等[12-13]提出 RDF 存储方案的基础上,提出使用 SP_O、PO_S、OS_P 3 张表来存储本体的方式,同样可覆盖所有的 RDF 3 元组查询,其中 SP_O 表的行键为(Subject,Predicate),列族存放 Object; PO_S表的行键为(Predicate,Object),列族为 Subject; OS_P 表的行键为(Object,Subject),列族存放 Predicate,该方法可大大减少存储空间. 针对领域本体的查询,可采用基于 HBase API 的 SPARQL BGP(SPARQL 基本图模式)的查询方法[14],该方法是基于 Hadoop 分布式集群平台上采用 MapReduce 并行计算框架来实现的,可有效提高查询效率.

因此,可设计 S_P_O、P_S_O 和 O_S_P 3 张 HBase 表来存储 MOOC 领域本体并基于 HBase API 的 SPARQL BGP 查询方法对 MOOC 本体进行查询.

1.2 MOOC 统一检索平台的自动构建

MOOC 统一检索平台的自动构建,需要依次经过数据采集及预处理、MOOC 本体自动构建和本体存储等才能实现语义查询. 其中最核心的是 MOOC 本体的自动构建和 MOOC 本体存储.

MOOC 统一检索平台的自动构建,主要分为 3 个模块. 第 1 个模块为数据采集及预处理,该模块采用网络爬虫工具获取学堂在线、中国大学 MOOC、Coursera 和 Edx 等 MOOC 平台信息源的学习资源数据,通过相关预处理后存储到 Mysql 数据库. 第 2 个模块为 MOOC 本体自动构建,根据数据库与本体之间的映射关系将 Mysql 数据库中的表自动映射成 MOOC 本体,并将映射的 MOOC 本体上传至 HDFS

中. 第 3 个模块为 Jena 解析本体,并将解析后得到的 RDF 3 元组存储到 HBase 数据库中. MOOC 统一检索的平台自动构建的详细流程图如图 2 所示.

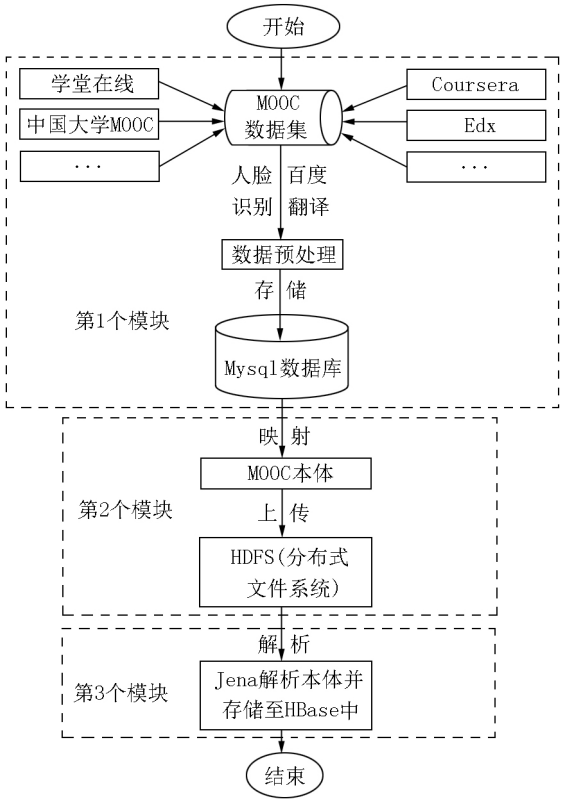

图 2　MOOC 统一检索平台的自动构建

1.2.1　**数据采集及预处理**　(i) 该实验数据主要从学堂在线、中国大学 MOOC、Coursera 和 Edx 等现有 MOOC 平台上采用爬虫工具获取,共采集 7 000 多条原始数据. 其中,采集的 MOOC 原始数据以学堂在线和中国大学 MOOC 平台为主,将采集的原始数据处理成 2 维表的形式并存储至 Mysql 数据库中.

(ii) 数据预处理. 获取的原始数据不能直接使用,需要采取一致性检查、缺失值处理和异常数据清除等方法对其进行预处理,从而得到完整可用的数据集. 数据一致性检查通过使用 SPSS 统计分析软件来检查数据格式是否合乎要求,检查并纠正非正常、逻辑上不合理或者互相矛盾的数据; 对某些数据的缺失,需要进行补全; 对于一些无效数据和异常数据需要进行清除. 另外,针对一些无法直接用于研究的原始数据,需要经过人工处理才能使用. 如针对教师性别则无法直接获得,对此需要采取一定的方法进行处理,采用人工识别教师人脸从而判断性别的方式虽可以准确识别,但当数据量庞大时显然不可取,而通过使用百度的人脸识别 API 可以自动识别教师的性别,这可大大节省人工识别人脸的成本. 此外,为了保证国内外 MOOC 平台的一致性,将一些字段值翻译为中文,比如将课程名字、课程所属学校、教师级别等字段值翻译成中文,这里使用百度翻译 API 将字段值翻译成中文.

原始数据经过处理后,得到的部分学习资源数据信息见表 1.

表 1　学习资源信息

c_{Name}	c_{Platform}	c_{School}	t_{Language}	t_{Name}	t_{Level}	t_{Sex}	…
Web 开发技术	学堂在线	重庆大学	中文	王成良	教授	男	
公正	堂在线	哈佛大学	英文	Michael Sandel	教授	男	…
悖论和无穷	学堂在线	麻省理工学院	英文	Agustin Rayo	教授	男	…
职业与创业胜任力	学堂在线	南京大学	中文	费俊峰	副教授	男	…
宇宙新概念	中国大学 MOOC	武汉大学	中文	赵江南	副教授	男	…
世界文化地理	中国大学 MOOC	北京大学	中文	邓辉	教授	男	
数据结构基础	中国大学 MOOC	北京大学	中文	张铭	教授	女	
全球贫困挑战	Coursera	麻省理工学院	英文	Abhijit Vinaya	教授	男	…
数据结构	Edx	加州大学圣地亚哥分校	英文	Daniel Kane	教授	男	…
…	…	…	…	…	…	…	…

注: c_{Name} 为课程名字,$c_{\text{Platformn}}$ 为课程所属平台,c_{School} 为课程所属学校,t_{Language} 为授课语言,t_{Name} 为教师名字,t_{Level} 为教师职称,t_{Sex} 为教师性别.

1.2.2　**MOOC 本体的自动构建**　将上一步骤获得的关系数据表进行本体映射,自动构建 MOOC 领域本体. 通过设立一套完善的映射规则以抽取本体构建所需要的基本元素: 概念、属性、层次、公理以及实例[15]. 数据库和本体之间的具体映射关系见表 2.

根据映射规则,将上一步骤处理的数据库的表自动映射成 MOOC 领域本体,将映射的本体文件导入到本体构建工具 Protégé4. 3 中,对 MOOC 本体进行进一步修改,从而完成 MOOC 本体的自动构建. 最终构建的 MOOC 本体包括 3 个概念(教师、课程资源、学校)、12 个数据属性(课程名字、课程英文名字、课程描述、课程所属机构、课程所属学校英文名字、课程所在学校级别、课程所在平台、授课语言、教师名字、教师级别、教师级别英文名字、教师性别)

和 4 个对象属性(教、被教、属于、被属于). 根据数据库和本体之间的映射规则实现 MOOC 本体的自动构建,极大地节约了构建 MOOC 本体的成本.

表 2 数据库和本体映射

关系型数据库(Relational Database)	本体(Ontology)
表名(Table)	类或概念(Concept)
字段(Fields)	属性(Attributes)
元组(Tuple)	实例(Instance)
数据库中的约束(Constraint)(主键、非空、唯一)等	基数限制、定义域、值域(owl: Cardinality、Domain、Range)等

1.2.3 Jena 解析本体并将解析后得到的 RDF 3 元组存储到 HBase 中 对于本体的解析,采用本体解析工具 Jena 将构建好的 MOOC 本体解析成标准的 RDF 3 元组模式[16],如将表 1(学习资源信息表)映射成 MOOC 本体后,经过 Jena 解析得到的结果如下:(C_1,t_{ype},C_{ourses}),(C_1,c_{Name},学习工程师),(C_1,$c_{Platform}$,学堂在线),(C_1,c_{School},重庆大学),(C_1,$t_{Language}$,中文),(C_1,t_{Name},战德臣),(C_1,t_{Level},教授),(C_1,t_{Sex},男),…,(C_2,c_{Name},公正),(C_2,$c_{Platform}$,学堂在线),(C_2,c_{School},哈佛大学),(C_2,$t_{Language}$,英文),(C_2,t_{Name},Michael Sandel),(C_2,t_{Level},教授),(C_2,t_{Sex},男),…,将解析后的 RDF 3 元组文件上传至 HDFS 中.

将 HDFS 上的 RDF 3 元组文件存储至 HBase 数据库中. 采用基于分布式数据库 HBase 的 RDF 本体存储架构,通过设计 S_P_O、P_S_O、O_S_P 3 张表分别存储 RDF 3 元组. 3 张 HBase 数据表的存储模型分别为: Subject→{Info→{Predicate→{Object}}},{Predicate→{Info→{Subject→{Object}}},{Object→{Info→{Predicate→{Subject}}},3 张表的列族均为 Info,S、P、O 分别表示 3 元组的 Subject(主语)、Predicate(谓语)、Object(宾语),S_P_O 表示以 Subject 作为表的行健、Predicate 作为列限定符、Object 作为单元值; P_S_O 表示以 Predicate 作为表的行健、Subject 作为列限定符、Object 作为单元值; O_S_P 表示以 Object 作为表的行健、Subject 作为列限定符、Predicate 作为单元值. 采用 MapReduce 并行框架将 RDF 3 元组格式化为 HFile 文件,之后采用 BulkLoad 将 HFile 导入 Hbase 进行批量处理以提高存储速度.

1.3 语义检索模型

通过构建一个语义检索模型,供学习者检索课程资源并推送相关课程. 计算查询的课程对象与 MOOC 本体中的课程对象之间的语义相似度,按照其相似度大小为学习者推荐符合其检索条件的课程,使用 HBase API 的 SPARQL 基本图模式查询,可在 MOOC 资源数据不断剧增的情况下实现高效语义检索.

根据相似度的可量化原则,课程的各个特征属性权重取值为 0 ~ 1 之间. 将 MOOC 平台中课程的属性分为: 课程所属学校(985/211,非 985/211,国外大学,其他),依据学校所属层次将其权重依次设置为(0.4,0.3,0.2,0.1); 教师类别(教授,副教授,讲师,其他),依据教师职称级别将其权重依次设置为(0.4,0.3,0.2,0.1); MOOC 平台(学堂在线,中国大学 MOOC,Coursera,Edx),依据该平台的注册基数将其权重依次设置为(0.3,0.3,0.2,0.2); 课程授课语言(中文,英文),其权重依次设置为(0.5,0.5); 教师性别(男,女),其权重依次设置为(0.5,0.5); 教师教学风格,由于暂未考虑教学风格,因此其权重可设置为 0. 即 $A_{ttr}(C_{ourse}) = \{c_{Scholl}, t_{Level}, c_{Platform}, c_{Language}, t_{Sex}, t_{Style}\}$. 所以,通过以下公式计算学习者查询的课程对象与 MOOC 本体中的课程对象之间的相似度:

$$f(\text{Degree of Interest Course}) = \alpha/(\alpha + \sqrt{f(x)}), \quad (1)$$

其中 $f(x) = \beta_1 c_{School} + \beta_2 t_{Level} + \beta_3 c_{Platform} + \beta_4 c_{Language} + \beta_5 t_{Sex} + \beta_6 t_{Syle}$.

(1)式表示学习者查询的课程对象与 MOOC 本体课程对象之间的语义相似度,其中 α、β_1、β_2、β_3、β_4、β_5、β_6 为调节因子,经反复实验表明,将 α 设置为 $3\sqrt{6}$、β_1 设置为 0.5、β_2 设置为 0.2、β_3 设置为 0.1、β_4 设置为 0.1、β_5 设置为 0.1 较为合理,暂未考虑教师的教学风格,将 β_6 设置为 0. 以下对(1)式得到的相似度进行归一化:

$$f(\text{Search Course}) = \lambda(\text{Degree of Interest Course}), \quad (2)$$

(2)式表示对相似度进行归一化,值越大表示该课程越符合学习者的检索需求,$\lambda \in [0,1]$ 为调节因子,经过反复实验表明,将 λ 设置为 0.9 较为合适. 当输入"男教授教的 java"检索条件时,采用中文分词算法[17]得到"男"、"教授"、"教"、"的"、"java"等关键词,然后剔除停用词、低频次和无关词后得到"男"、"教授"、"java"等关键词. 针对上述的关键词,采用 HBase API 的 SPARQL 基本图模式查询算法对 HBase 数据库中的 RDF 3 元组进行语义检索,得到检索的课程资源对象集合. 由于已经确定了学校和教师级别的属性值,所以(1)式剩下课程所属 MOOC 平台、课程授课语言、教师性别等属性,通过计算已检索到的课程对象与 MOOC 本体中课程对

象之间的相似度,按照其值大小依次为学习者推荐课程资源.

2　MOOC 统一检索平台的应用

构建的 MOOC 统一检索平台的最终目的是为学习者提供更加智能化的应用服务,其中最主要的是实现语义检索服务. 通过 MOOC 统一检索平台的检索结果与现有的 MOOC 平台的检索结果进行对比、分析,得出构建的 MOOC 统一检索平台相较于现有 MOOC 平台其查准率和查全率更高. 在输入相同的检索条件下,MOOC 统一检索平台和现有 MOOC 平台的检索结果见表 3 和表 4.

表 3　MOOC 统一检索平台检索结果　　门

检索条件	学堂在线	中国大学 MOOC	Coursera	Edx	总计
男教授教的 java	5	6	4	3	18
985/211 大学数据结构	9	15	2	1	27
中文授课的 java	3	8	1	1	12

表 4　现有 MOOC 平台检索结果　　门

检索条件	学堂在线	中国大学 MOOC	Coursera	Edx
男教授教的 java	23	0	609	0
985/211 大学数据结构	33	0	0	0
中文授课的 java	25	0	328	0

以输入检索条件“男教授教的 java”为例,在构建的 MOOC 统一检索平台上,共检索到 18 门课程,课程按照语义相似度从大到小依次推荐给学习者,部分检索结果如表 5 所示.

表 5　MOOC 统一检索平台检索结果

课程号	课程名字	课程授课语言	MOOC 平台	所在大学	授课教师	授课教师级别	授课教师性别	语义相似度
C4710	Java and…	中文	Edx	香港科技大学	Ting-Chuen	教授	男	0.900
C4712	AP 计…	英文	Edx	PurdueX	PurdueX…	教授	男	0.840
C4705	Java 编程	英文	Edx	HKUSTx	Ting-…	教授	男	0.840
C4704	Java 编程	英文	Edx	HKUSTx	Ting-Chuen…	教授	男	0.840
C4691	Java 中的…	英文	coursera	莱斯大学	Vivek Sarkar	教授	男	0.837
C4690	Java 中的…	英文	coursera	莱斯大学	Vivek Sarkar	教授	男	0.837
C4689	Java 中的并行…	英文	coursera	莱斯大学	Vivek Sarkar	教授	男	0.837
C3030	函数式…	英文	学堂在线	荷兰代…	Erik…	教授	男	0.836
C2458	面向对象程…	中文	中国大学 MOOC	浙江大学	翁恺	教授	男	0.828

通过观测和分析表 3、表 4 以及表 5 可知,以查询条件“男教授教的 java”查询结果为例,MOOC 统一检索平台上共检索到 18 门相关课程,其中,有 5 门属于学堂在线中的课程、6 门属于中国大学 MOOC、4 门属于 Coursera、3 门属于 Edx; 而在现有 MOOC 平台上以同样的检索条件检索的结果分别为: 学堂在线为 23 门、中国大学 MOOC 为 0 门、Coursera 为 609 门、Edx 为 0 门. 现有 MOOC 平台不能实现语义检索,如学堂在线平台仅抓取其中的“java”关键字进行检索,忽略了“男”、“教授”等限制条件; 中国大学 MOOC 平台由于仅根据输入的整个语句作为关键词进行检索,不具有语义检索,因此未检索到相关内容; Coursera 平台检索结果虽然较多,但大多是不相关课程; Edx 平台未检索到相关内容,同样也是由于其仅根据输入的条件作为关键词进行检索. 因此,构建的 MOOC 统一检索平台检索出的结果均符合学习者的查询需求,而现有 MOOC 平台未能有效检索出符合条件的结果,MOOC 统一检索平台的查准率和查全率比现有 MOOC 平台更高,并且该平台在 MOOC 资源数据不断剧增的情形下仍能实现高效查询.

现有的 MOOC 平台仅仅是根据输入的条件作为整个关键词或抓取其中某个关键词进行匹配,难以检索出学习者想要的课程,其查准率和查全率较低,从而影响学习者的学习体验,而构建的 MOOC 统一检索平台不仅实现了跨 MOOC 平台的高效检索,而且检索到的课程符合学习者的查询要求,所以其查准率和查全率大幅提高.

3 结语

本文构建的 MOOC 统一检索平台不仅能够检索出各个信息源 MOOC 平台上的课程,实现语义检索,而且随着 MOOC 学习资源的爆发式增长,学习者仍能够在该平台上快速地寻得符合查询需求的课程资源.这有效解决了课程资源繁多对其造成的困扰,在一定程度上可提高学习者的学习效率和学习兴趣,并为进一步实现 MOOC 个性化推荐奠定坚实的基础.

MOOC 统一检索平台还存在一些有待完善之处,如 MOOC 本体构建有待完善、语义检索模型有待改进等.下一步将获取更多的 MOOC 平台数据以丰富 MOOC 本体,采用图像处理和语音识别技术分析教学视频以抓取教师授课时的脸部、眼睛、手势、说话的音速和音调等授课行为进而分析出该教师的教学风格,从而优化 MOOC 统一检索平台的语义检索模型,并获取学习者学习特征进而分析学习者的个性化特征,为其推荐符合其个性化的学习资源,最终在该 MOOC 统一检索平台上构建一个多维度的 MOOC 个性化推荐系统.

4 参考文献

[1] 吴文涛,张舒予.语义慕课:语义网环境下 MOOC 的发展愿景[J].中国电化教育,2016(9):51-58.

[2] 吴鹏飞,余胜泉.语义网教育应用研究新进展:关联数据视角[J].电化教育研究,2015,36(7):66-72.

[3] Dietze Stefan,et al. Linked education:interlinking educational resources and the Web of data[EB/OL].[2018-09-17].https://dlacm.org/citation.cfm?id=2245347.

[4] 刘小乐,马捷.语义网环境下基于本体的知识集成研究进展[J].现代情报,2015,35(1):159-163,169.

[5] 位传海,范太华.基于本体的学习资源语义检索系统研究与设计[J].电化教育研究,2012(2):70-74.

[6] 丁国柱,余胜泉.基于本体学习算法的学科本体辅助构建研究:以学习元平台语文学科知识本体的构建为例[J].中国电化教育,2015(3):81-89.

[7] 吴鹏飞,余胜泉.学习资源语义关联关系及其可视化研究[J].中国电化教育,2015(12):97-104.

[8] 王向前,张宝隆,李慧宗.本体研究综述[J].情报杂志,2016,35(6):163-170.

[9] Zhang Chen,Sterck H D. Supporting multi-row distributed transactions with global snapshot isolation using bare-bones HBase[EB/OL].[2018-09-20].https://ieeexplore.org/document/5697970.

[10] 陆婷,房俊,乔彦克.基于 HBase 的交通流数据实时存储系统[J].计算机应用,2015(1):103-107.

[11] Chebotko A,Abraham J,Brazier P,et al. Storing,Indexing and querying large provenance data sets as RDF graphs in apache HBase[EB/OL].[2018-10-11].http://ieeexlore.org/lpdocs/epico3/wraper.htm?arnumber=6655668.

[12] Sun Jianling,Jin Qiang. Scalable rdf store based on hbase and mapreduce[EB/OL].[2018-10-09].https://dl.acm.org/citation.cfm?id=1786977.

[13] Papailiou N,Konstantinou i,Tsoumakos D,et al. H2RDF:Adaptive query processing on RDF data in the cloud[C].New York:ACM,2012:397-400.

[14] 汪璟玢,方知立,张燕琴.面向分布式的 SPARQL 查询优化算法[J].计算机科学,2014,41(7):227-231.

[15] 郭朝敏,姜丽红,蔡鸿明.一种关系数据库到本体的自动构建方法[J].计算机工程与应用,2012,48(7):115-120,248.

[16] Khadilkar V,Kantarcioglu M,Castagna P,et al. Jena-HBase:a distributed,scalable and efficient RDF triple store[EB/OL].[2018-10-19].https://dl.acm.org/citation.cfm?id=2887401.

[17] 吴熠潇.中文分词相关算法研究[J].科技经济导刊,2018(2):122-123.

The Automatic Construction and Application of MOOC Unified Search Platform

XIAO Qingquan,LI Yunqing*

(School of Computer and Information Engineering,Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract: With the increase of MOOC platforms and the proliferation of learning resources under the same platform,how to achieve efficient cross-platform semantic retrieval has become one of the issues that MOOC need to solve urgently. Learning resource data of multiple well-known MOOC platforms is obtained through web crawlers,relevant data is preprocessed and stored into Mysql database,MOOC ontology is automatically constructed according to the mapping relationship between database and ontology,Jena is used to parse the MOOC ontology into RDF triples and the RDF triples are stored to the HBase database,and finally a MOOC unified search platform is built to recommend learning resources for the learners to meet their search requirements. The experimental results show that the constructed semantic MOOC platform can effectively improve the search accuracy and recall rate.

Key words: unified search platform; automatic construction; ontology; HBase database; MOOC

(责任编辑:冉小晓)