

# 学习资源动态语义关联的设计与实现\*

杨现民<sup>1</sup>, 余胜泉<sup>2</sup>, 张芳<sup>3</sup>

(1.江苏师范大学 教育科学学院, 江苏 徐州 221116; 2.北京师范大学 现代教育技术研究所, 北京 100875; 3.南京师范大学 现代教育技术中心, 江苏 南京 210097)

**摘要:** 学习资源是数字化学习生态系统的核心要素, 虽然各种媒介形态的资源不断涌现, 资源数量持续、快速增长, 但是资源之间普遍缺乏关联性。动态建立、发展、挖掘资源之间的各种语义关联, 是实现资源关联进化亟需解决的重要问题。该文基于学习元平台 (Learning Cell System, LCS), 提出一种综合应用语义基因、基于规则的推理、关联规则挖掘等技术实现资源动态语义关联的方法。实验结果表明, 该方法在实现资源动态语义关联上能取得理想的结果, 具有较高的关联准确性。

**关键词:** 学习资源; 动态语义关联; 规则推理; 语义基因; 关联规则

**中图分类号:** G434 **文献标识码:** A

## 一、引言

学习资源间丰富的语义关联, 一方面可以增强资源个体之间的联通, 提高各自被浏览或内容编辑的概率, 促进资源的快速进化; 另一方面还可以为学习资源动态聚合成更大粒度、具有内在逻辑联系的资源群提供数据基础。当前的e-Learning资源普遍缺乏关联性, 因为资源之间的联系是通过一般的超链接形成的人为关联, 基于HTML的数据组织不能体现数据内在的语义联系<sup>[1]</sup>。各种开放知识社区 (Wikipedia、Google Knol、Cohere、Cloudworks、Freebase等) 中, 知识单元间也主要通过多媒体编辑器、关系编辑器或属性编辑页面等以人工方式建立关联。学习资源的进化包括两种模式, 分别是内容进化和关联进化<sup>[2]</sup>。关联进化是指学习资源在生长的过程中不断与其他资源实体建立语义关系的过程, 是资源外部结构的持续发展和完善。如何动态建立、发展、挖掘资源之间的各种语义关联, 是实现资源关联进化亟需解决的重要问题。

学习对象 (Learning Object, LO) 是当前e-Learning领域非常重要的一种资源形态, 国内外已有学者对学习对象的关联技术进行了研究, 主要集中在关系元数据设计<sup>[3][4]</sup>、关联展现方式设计<sup>[5][6]</sup>、相似度量<sup>[7]</sup>、关联路径搜索<sup>[8]</sup>、自动化组装<sup>[9]</sup>等方面。当前学习对象的关联技术研究存在两个方面的

不足: 一方面, 在关联关系的表示上多采用静态的元数据描述技术, 未从语义层面考虑资源之间的关联关系, 缺乏对资源关联的规范化描述; 另一方面, 虽然有些研究开始从资源本体的角度考虑语义关系的计算, 但多限于相似关系的度量, 而忽视了其他资源语义关系 (如前序、后继、相反等) 的动态发现。

学习元平台 (Learning Cell System, LCS) 是为泛在学习环境设计开发的一种新型开放知识社区 (<http://lcell.bnu.edu.cn>), 包括学习元、知识群、知识云、学习社区、个人空间、学习工具等六大功能模块, 其中学习元是LCS中最基础的资源单元, 知识群是多个同主题学习元的集合。本研究基于LCS提出一种综合应用语义基因、基于规则的推理、关联规则挖掘等技术实现资源动态语义关联的方法, 并检验了该方法在LCS中的实际应用效果。

## 二、学习资源间的语义关系设计

Carsten Ullrich<sup>[10]</sup>认为SCORM的CAM中定义的关系元数据仅能描述结构导向的关系, 而无法有效描述语义层面的关系。因此, 很多研究者从语义层面去补充完善SCORM CAM中的关系元数据。Eric Jui-Lin Lu 和 Chin-Ju Hsieh<sup>[11]</sup>在概括分析已有扩展关系元数据研究成果的基础上, 对SCORM CAM中的关系元数据进行了扩展 (见下页表1), 并通过

\*本文系高等学校博士学科点专项科研基金博导类资助课题“泛在学习环境下的学习资源进化研究” (课题编号: 20110003110029) 研究成果。

调查验证了这些关系元数据的有效性。

表1 扩展的15种关系元数据

Law	Theorem	Process	Procedure
Guideline	Introduction	Remark	Conclusion
Definition	Illustration	Counterexample	Example
Demonstration	Proof	Evidence	

SCORM CAM中的关系元数据定义得比较简单实用，表1扩展的15种关系元数据虽然弥补了语义层面关系定义的不足，但在中文环境下表述容易产生混淆，比如Process、Law等作为关系属性显得有些牵强。综合考虑，本研究将以SCORM CAM关系元数据为核心（去除hasformat、isformatof、isversionof、hasversion等），适当吸收Eric Jui-Lin Lu和Chin-Ju Hsieh定义的扩展关系元数据，作为知识本体中的初始关系属性集（见表2）。

表2 知识本体中的初始关系属性

ispartof	haspart	isversionof	hasversion
isformatof	hasformat	references	isReferencedBy
isbasedon	isbasisfor	requires	isRequiredBy
similarTo	relateTo	oppositeOf	equivalentWith
supplement	isSupplementedBy	isExampleOf	isCounterExampleOf
isUpperConceptOf	isSubConceptOf	isSubsequentOf	isPreviousOf
remark	isRemarkBy	guide	isGuidedBy
demonstrate	isDemonstratedBy	cause	isCausedBy

需要说明的是，表2中的32种资源语义关系仅作为LCS中的初始关系集合，为了支持语义关系的可扩展性，LCS设计了一种语义关系协同创建的机制，允许普通用户在使用过程中不断补充关系属性，通过系统审核后的新关系属性将自动补充到语义关系库中（语义关系库是知识本体库的一个子集）。系统管理员可以在LCS中的资源语义关系管理页面中增加、删除、修改、审核、查询各种资源关系。

### 三、学习资源动态语义关联技术实现

学习资源间的关联主要包括两种类型，一种是显性关联，另一种是隐性关联。显性关联是从语义出发基于系统已有的关系类型建立的资源关联，易被用户观察和识别；隐性关联是从语义上难以通过人工发现，但可以通过数据挖掘技术识别出来的潜在的资源关联。图1描述了LCS中学习资源的动态语义关联技术。在显性关联的建立上分别采用了基于规则的推理技术和基于语义基因的相似关系计算技术，在隐性关联的建立上主要采用了基于语义约束的关联规则挖掘技术。

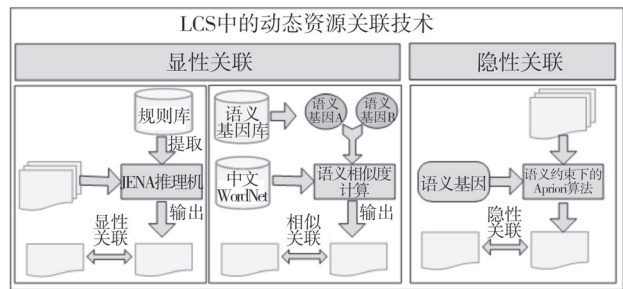


图1 LCS中的学习资源动态语义关联技术

#### （一）基于规则推理的资源显性关联

JENA<sup>[12]</sup>是由HP Lab开发的一款用于Semantic Web应用程序开发的开源框架，除了包含丰富的本体操作API外，还支持基于产生式规则的前向推理。e-Learning领域的研究者们已经开始应用JENA的推理功能实现个性化的学习指导<sup>[13]</sup>、信息检索<sup>[14][15]</sup>、适应性内容推荐<sup>[16]</sup>等。LCS可以应用JENA框架操作本体模型，自定义各种产生式的关联规则，通过JENA推理机实现部分资源显性关联。

基于规则推理实现资源显性关联的基本流程如图2所示：首先编写各种关联推理规则，并存储到推理规则库中；然后，JENA推理机从规则库中提取规则，将规则绑定到本体模型；接着，JENA推理机依据规则对本体模型进行推理；最后，将推理出的显性资源关联集合进行输出。



图2 基于规则推理的资源显性关联流程

应用JENA框架实现基于规则推理的资源显性关联之前，有两项重要工作需要完成。首先，需要将与资源关联相关的数据采用RDF三元组形式存储到JENA支持的本体模型中。其次，要根据JENA推理机定义的规则形式编写各种关联规则，推理机会绑定这些规则并对本体模型进行推理，得到新的推理后的本体模型。针对表2中定义的关系属性，笔者编写了17种关联推理规则（见下页表3）。

需要说明的是，上述规则不是固定不变的，随着本体模型中属性的逐渐丰富，将会产生更多有意义的规则，只需将规则按照JENA规定的格式存入规则库，就可以用于资源显性关联的推理发现。这里以规则6定义为例解释JENA规则的代码表示：

```
String rule6="[(?x lc: supplement ?y) -> (?x lc: isSupplementedBy ?z)];"
```

表3 基于JENA的关联推理规则

编号	规则描述
规则1	(x lc:requires y)->(y lc:isRequiredBy x)
规则2	(x lc:isUpperConceptOf y)->(y lc:isSubConceptOf x)
规则3	(x lc:remark y)->(y lc:isRemarkBy x)
规则4	(x lc:guide y)->(y lc:isGuidedBy x)
规则5	(x lc:demonstrate y)->(y lc:isDemonstratedBy x)
规则6	(x lc:supplement y)->(y lc:isSupplementedBy x)
规则7	(x lc:similarTo y)->(y lc:similarTo x)
规则8	(x lc:oppositeOf y)->(y lc:oppositeOf x)
规则9	(x lc:relateTo y)->(y lc:relateTo x)
规则10	(x lc:oppositeOf y)(y lc:oppositeOf z)->(x lc:similarTo z)
规则11	(x lc:equivalentWith y)->(y lc:equivalentWith x)
规则12	(x lc:isExampleOf y)(z lc:isCounterExampleOf y)->(x lc:oppositeOf z)
规则13	(x lc:cause y)->(y lc:isCausedBy x)
规则14	(x lc:isPreviousOf y)->(y lc:isSubsequentOf x)
规则15	(x lc:isPreviousOf y)->(x lc:isbasedfor y)
规则16	(x lc:references y)->(y lc:isreferencedby x)
规则17	(x lc:ispartof y)->(y lc:haspart x)

JENA的每条规则都采用产生式表示，“->”左侧的部分表示推理的条件，“->”右侧的部分表示推理的结果，条件项和结果项都采用RDF三元组(Subject, Predicate, Object)的形式描述。规则6比较简单，条件项和结果项各包含一个三元组，实际上复杂规则的条件项和结果项可以包含多个三元组。系统管理员可以在LCS中的基于规则推理的关联规则管理页面中增加、删除、修改、禁用、查询各种关联推理规则。

(二) 基于语义基因的资源相似关系计算

语义基因是指能够反映资源内容所要表达含义的基本信息单元，形式上表现为基于本体描述的带有权重的概念集合以及概念间的语义关系。区别于文本相似度比较中的文档特征项，语义基因不是简单的关键词集合，而是资源背后所隐藏的语义概念网络。

语义基因在形式上表现为基于本体描述的带有权重的概念集合（包括核心概念以及概念间的关系）。语义基因可以被形式化地表示为有序三元组（见图3），即 $SG = \langle CS, WS, RS \rangle$ ，其中CS是核心概念集合， $CS = \{C_1, C_2, C_3, \dots, C_n\}$ ；WS是概念项的权重集合， $WS = \{W_1, W_2, W_3, \dots, W_n\}$ ，其中 $W_i$ 为 $C_i$ 的权重， $\sum_{i=1}^n W_i = 1$ ；RS为核心概念间的关系

集， $RS = \{R_1, R_2, R_3, \dots, R_n\}$ ，每个关系采用领域本体中的RDF三元组 $\langle Subject, Predicate, Object \rangle$ 表示， $R_1 = \langle Concept1, Relationship, Concept2 \rangle$ ，这里的Concept1和Concept2不一定包含在CS中，可以是领域本体库的其他概念，Relationship是从领域本体库中提取的概念关系。

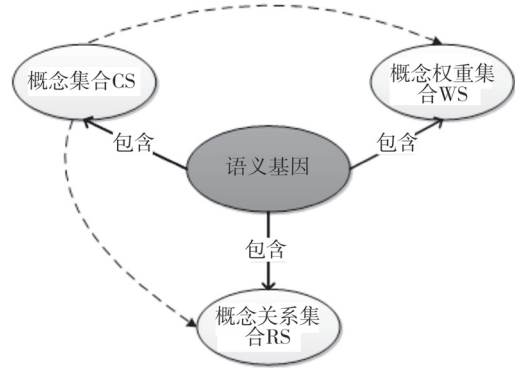


图3 语义基因的结构要素

基于语义基因的相似关系计算的基本思路是：首先，基于通用的语义词典和领域本体计算语义基因中两两概念间的相似度；然后，结合概念在语义基因中的权重值设置相似度的权值；接着，将所有相似度进行加权平均得到两个语义基因的相似度；最后，根据设定的相似度阈值判断两个资源是否具有相似关系（见图4）。

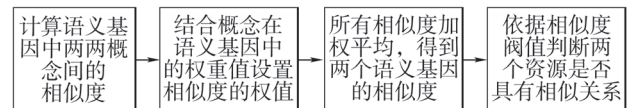


图4 基于语义基因的资源相似关系计算流程

本研究采用吴思颖等<sup>[17]</sup>提出的3d-sim方法计算两个概念间的相似值，进而计算两组语义基因的相似度（见图5）。假如有两组语义基因X和Y， $X = \{(C_{11}, C_{12}, C_{13}, \dots, C_{1n}), (W_{11}, W_{12}, W_{13}, \dots, W_{1n}), (RS_{11}, RS_{12}, \dots, RS_{1t})\}$ ， $Y = \{(C_{21}, C_{22}, C_{23}, \dots, C_{2m}), (W_{21}, W_{22}, W_{23}, \dots, W_{2m}), (RS_{21}, RS_{22}, \dots, RS_{2s})\}$ ，n为X中概念集合的概念数量，t为X中概念关系三元组的数量，m为Y中概念集合的概念数量，s为Y中概念关系三元组的数量。

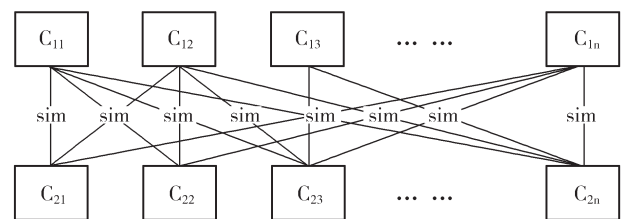


图5 计算两两概念项的相似度



果显示,截止到2012年3月1日,LCS中的资源关联总数AT(Association Total) = 3557,平均资源关联度AAD(Average Association Degree) =  $AT / RT = 0.91$ ,也就是说在LCS中平均每个学习元和其他0.91个学习元之间存在着语义关联。直观地看,0.91的结果说明LCS中资源之间语义关联并不是很丰富。通过进一步对资源关联结果的统计分析发现,共2918个学习元的关联数为0,说明当前LCS中有一大部分资源的内容差异较大,和其他学习元之间不存在关联关系。当前LCS中的资源关联主要分布在857个学习元(约占系统总资源数的22.7%)之间,形成局部的资源关联网,局部平均资源关联度AAD=4.15。从上述统计结果可以发现一个有趣的现象:LCS中的资源关联分布基本呈现“二八原则”,80%以上的关联集中在20%的资源上面。

从关联类型的应用上来看,目前LCS已经实际应用的关联类型有29种(占总关联类型的80.56%),说明LCS中有约五分之一的关联类型暂时处于“休眠”状态。而在29种关联类型中,关联数量排在前五位的关系分别是相似、相关、是基础、前序和后继。尤其是相似和相关的关系数量之和,所占比例超过关联总数的52%。下一步,可以将这些常关联类型按使用频度排在前面,以方便用户手动编辑资源关联。

表5显示了LCS中资源关联方式的分布情况,其中手动建立关联的比例为13.6%,自动建立语义关联的比例为86.4%(语义基因计算:18.7%,基于规则推理:51.6%,关联规则挖掘[收藏事务]:15.8%,关联规则挖掘[订阅事务]:0.3%)。统计结果表明,当前LCS中用户建立的关联数量较少,绝大多数的资源关联是采用动态语义关联方法自动建立的。动态语义关联技术在LCS资源群体的关联进化上发挥了重要作用。

表5 资源关联方式分布情况

	手动建立 关联	语义基因 计算	基于规则 推理	基于收藏 事务	基于订阅 事务
数量	570	783	2156	660	12
百分比	13.6%	18.7%	51.6%	15.8%	0.3%

除了从关联数量上说明动态语义关联方法的效果外,还需要进一步检验资源关联的准确性。这里将动态语义关联方法作为一个整体,检验其自动生成资源关联的准确性。

检验方法为:通过程序从LCS中随机选取150条资源关联记录,存储到Excel表中;邀请2名研究生登录LCS,每人负责判断其中的75条资源关联的

正确性,将结果记录到Excel中;汇总结果,统计关联准确率。正式实验前,进行了Kappa一致性系数检验, $Kappa=0.81 > 0.75$ ,表明两位研究生的判断具有较高的一致性。

表6显示了用户判断的正确/错误关联数量。自动建立关联的准确率precision = 71.33%,说明动态语义关联方法具有较高的可靠性,其自动建立的资源之间的关联大部分是准确的。

表6 LCS中资源自动关联的人工判断结果

	用户A	用户B
正确数	58	49
错误数	17	26

此次实验仅对动态语义关联方法的整体准确性进行了初步检验,下一步,随着LCS中资源数量和关联数量的不断增加,将分别对基于语义基因的关联、基于规则推理的关联、基于关联规则挖掘的关联等三种具体资源关联方法的效果进行检验。此外,基于语义基因进行资源相似关系计算存在执行效率低、时间开销大等缺陷,下一步将对算法进行进一步测试、优化。

### 五、结束语

学习资源间语义关联的动态建立和发展是实现资源关联进化的核心。语义Web技术和数据挖掘技术在资源语义关联实现方面具有重要应用价值。本研究提出的学习资源动态语义关联方法,经过初步检验证明是有效的,可以有效促进学习元平台中资源间语义关联的建立。

不足在于,资源关联方法的运行效率偏低,关联算法有待优化。接下来,将从如下两方面开展重点研究:(1)结合中文WordNet、哈工大同义词词林等语义词典,设计资源关系计算规则和算法,通过计算发现资源间更丰富的语义关系;(2)将本研究提出的资源关联方法推广运用到其他开放知识社区,以进一步检验和完善该方法。

### 参考文献:

- [1] 邵国平,余盛爱,郭莉.语义Web对E-Learning中资源管理的促进[J].江苏广播电视大学学报,2008,19(5):23-26.
- [2] 杨现民,余胜泉.泛在学习环境下的学习资源进化模型构建[J].中国电化教育,2011,(9):80-86.
- [3] [10] Carsten Ullrich. The learning-resource-type is dead, long live the learning-resource-type[J]. Learning Objects and Learning Designs, 2001, 1(1): 7-15.
- [4] [11] Eric Jui-Lin Lu, & Chin-Ju Hsieh. A relation metadata extension for SCORM Content Aggregation Model [J]. Computer Standards & Interfaces, 2009,(31): 1028-1035.

- [5] 吕翘楚, 杜辉. 基于知识地图的学习内容管理系统的系统设计[J]. 硅谷, 2010,(8): 57-58.
- [6] 施岳定, 张树有, 项春. 网络课程中知识点的表示与关联技术研究[J]. 浙江大学学报(工学版), 2003, 37(5): 508-511.
- [7] 张骞, 张霞, 刘积仁. SCORM学习资源的语义相似度度量[J]. 华中科技大学学报(自然科学版)增刊, 2003, 31(10):296-298.
- [8] 李艳燕. 基于语义的学习资源管理及利用[D]. 北京: 中国科学院计算技术研究所, 2005.
- [9] Robert G. Farrell, Soyini D. Liburd, & John C. Thomas. Dynamic assembly of learning objects[A]. Proceedings of the 13th international World Wide Web conference on Alternate track papers \& posters (WWW Alt. '04)[C]. Lisbon: ACM press, 2004.162-169.
- [12] Yu, L. Y. Jena: A Framework for Development on the Semantic Web [M]. New York :Springer Berlin Heidelberg, 2011.
- [13] 陈和平, 郭晶晶, 吴怀宇等. 基于Ontology和Jena的个性化E-Learning系统研究[J]. 武汉理工大学学报(交通科学与工程版), 2007, 31(6):1049-1052.
- [14] 耿科明, 袁方. Jena推理机在基于本体的信息检索中的应用[J]. 微型机与应用, 2005, 24(10): 62-64.
- [15] Huang, C., Duan, R., Tang, Y., Zhu, Z., Yan, Y., & Guo, Y. EHS: An Educational Information Intelligent Search Engine Supported by Semantic Services[J]. International Journal of Distance Education Technologies (IJDET),2011, 9(1): 21-43.
- [16] Ion-Mircea Diaconescu, Sergey Lukichev, & Adrian Giurca. Semantic Web and Rule Reasoning inside of E-Learning Systems[J]. Studies in Computational Intelligence, 2008,(78): 251-256.
- [17] 吴思颖, 吴扬扬. 基于中文 WordNet 的中英文词语相似度计算[J]. 郑州大学学报(理学版), 2010, 42(2): 66-69.
- [18] R. Agrawal, & R. Srikant. Fast algorithms for mining association rules in large database[R]. San Jose, CA: Technical Report FJ9839, IBM Almaden Research Center, 1994.
- [19] 毕建欣, 张岐山. 关联规则挖掘算法综述[J]. 中国工程科学, 2005, 7(4): 89-93.
- [20][22] 生佳根, 刘思峰. 一种基于本体的关联规则挖掘方法[J]. 南京理工大学学报(自然科学版), 2008, 32(4): 401-405.
- [21][23] Zhang L., Xia S.X., Zhou Y., & Xia Z. G. Study on association rules mining based on semantic relativity [J]. Journal of Southeast University (English Edition), 2008, 24(3): 358-360.

#### 作者简介:

杨现民: 博士, 研究方向为移动与泛在学习、数字化学习资源设计、网络教学平台开发、信息技术教学应用(yangxianmin8888@163.com)。

余胜泉: 博士生导师, 教授, 研究方向为移动与泛在学习、教育信息化、信息技术与课程整合。

张芳: 硕士, 助理实验师, 研究方向为精品课程建设、信息化教学研究、教学资源与软件研发。

收稿日期: 2012年11月4日

责任编辑: 宋灵青

#### 简讯

### 第八届全国教育技术学博士生学术论坛在西南大学举行

教育技术博士生论坛是研究生教育创新工程的重要平台之一, 一直以来都广受关注。2012年12月9日至10日, 由西南大学研究生部主办、西南民族教育与心理研究中心承办的第八届全国教育技术学博士生学术论坛在西南大学举行。论坛的主题为: “数码时代的文化选择”, 来自北京师范大学、东北师范大学、西北师范大学、南京师范大学、首都师范大学等30余所高校、科研院所的200余名硕士、博士生参与了论坛, 《中国电化教育》作为主要合作媒体应邀参加。

西南大学校长张卫国、副校长崔延强出席论坛开幕式, 张校长代表学校致欢迎辞。他指出, 研究生教育不仅是强校之路, 也是立校之本, 希望广大青年学者能够通过此次论坛启发学术思维, 拓宽学术视野, 提升学术能力。西南民族教育与心理研究中心主任张诗亚教授以云时代为背景, 提出了云时代学习的思维和方法, 得到与会代表的一致认同。

开幕式后, 教育部高等学校教育技术学教学指导委员会主任、华南师范大学教育信息技术学院徐福荫教授, 北京师范大学教育学部副部长黄荣怀教授, 南京师范大学教育科学学院李艺教

授, 西北师范大学教育技术与传播学院郭绍青教授等4位专家应邀作了主题报告, 就教育技术学科建设与人才培养、教育信息化领导力、教育技术研究方法及教师培训等方面的内容和与会代表进行了深入交流, 得到了广泛的响应。

开幕式后, 会议分为四个分论坛, 就提交的论文进行了宣读与讨论, 并对部分论文进行了集中研讨与评审。此次会议征文涉及: 教育技术学科建设、人才培养模式研究, 新技术支持的未来教育, 数码时代的视觉文化与媒介素养教育, 虚拟世界中的知识学习与文化传播, 泛在学习的理论与实践, 教育技术促进教师专业发展, 教育技术与民族文化保护, 教育技术与西部地区教育发展等多个方面, 最终共评选出优秀论文7篇。

12月9日晚, 《中国电化教育》《开放教育研究》《远程教育杂志》《现代远程教育》四家刊物的代表分别就本刊的关注重点和选投稿方式与参会的研究生进行了座谈, 进一步拉近了学术期刊与学生的距离。

(本刊记者 马小强)