基于教育主题词表的Delicious中英文标签语义关系抽取*

贾君枝，段佳慧
（山西大学经济与管理学院，太原 030006）

摘要：分众分类标签的语义模糊性和标签无序性使得检索效率越来越低，为准确定位标签语义，旨在研究一种新的方法，从主题词表的词间关系判断标签间的语义关系。以《中国分类主题词表》教育类主题词搜索Delicious网站得到的相关标签作为数据源，分析标签与主题词的重叠度，选择《中国分类主题词表》和ERIC在线教育词表作为标签的语义关系抽取工具，认为相关标签集中语义关系较为丰富。

关键词：标签；主题词表；语义关系
中图分类号：G254
DOI：10.3772/j.issn.1673—2286.2014.08.001

1 导言

标签是伴随分众分类法出现的最为广泛的应用之一，用户可以为每篇文章或每张图片随意添加一个或多个标签，从而根据这些标签把文章或图片进行分类。用户可以在系统中创建自己的标签，也可以对文章或链接使用标签来标记。然而标签也存在一些缺点，主要表现在两个方面：一是由于标签的用户随意性造成了标签中存在很多噪声数据，错拼、歧义以及无实际意义的标签充斥其中；二是由于用户在标注同一信息时使用了不同的标签，但系统无法创建这些标签之间的联系，无法识别各个标签之间的层级关系，造成了标签的无序性。

标签形式的社会性语义越来越占主导地位，Spiteril等人提出分众分类法与受控词表间并不对立，利用分众分类系统中的标签资源可以补充和完善受控词表的词汇[1]。Rolla等人比较了Library Thing网站的用户标签与LCSH（美国国会图书馆标题表）[2]。国内贾君枝教授对分众分类法与受控词表的结合研究进展进行分析，提出了利用标签系统中的标签来解决国内受控词表老化的问题[3]。王东元等分析了Delicious系统中的中文标签特征，并将Delicious的中文标签与《汉语主题词表》的主题词进行比较，发现有近四分之一的中文标签可以在汉表中找到[4]。张宁等对Delicious教育类中文标签与《中国分类主题词表》中教育类主题词进行比较分析，探讨标签和主题词融合的可能性和技术实现方法，提出了相互借鉴融合的理论模型和方法[5]。这些研究仅仅对分众分类法与受控词表之间的关系进行了研究，缺乏对标签语义关系的深入探讨。

Delicious系统中，其通过一定算法所建立的相关标签集（Related Tags），已经过滤了大部分的噪声数据，大部分标签具有实际意义，一定程度上减少了标签的无序性，实现标签聚类。本文以此作为实验数据来源，研究相关标签集的语义关系，参照《中国分类主题词表》、ERIC在线词表的语义关系，抽取出语义关系，再进行人工明确，旨在通过标签与资源之间的关系来细化语义关系，以达到及时完善受控词表语义关系的目的。

2 教育领域标签与主题词的数据采集

2.1 主题词的采集与处理

《中国分类主题词表》（以下简称中分表），是在《中国图书馆图书分类法》和《汉语主题词表》两种检索语言基础上编制而成的、我国第一部分类主题一体化的

* 本研究得到国家社会科学基金青年项目“基于框架网络本体的标签系统语义分析研究”（编号：13CTQ030）资助。

大型综合性词表。目前，中分表作为主题标引和检索的重要工具，已经被我国众多大中型图书情报机构所使用。中分表中主题词间的组织关系和其他主题词表相同，存在三种词间关系：等同、层级和相关。等同关系反映词汇间的同义/近义关系，将非优选词（非规范用词，不能用于标引）与优选词（标准的、规范用词，用于标引的主题词）联接起来。层级关系是主题词间的上下位关系。相关关系用来概括除了等同和层级之外的其他关系。

我们选择《中国分类主题词表》中教育类（G4）对应的类目和所有主题词作为标签检索的入口词。但需要对主题词和类目中的一些非中文字符进行处理。类名和主题词中涉及顿号（、）、圆括号（()）、反斜杠（\）、方括号（[]）、双竖线（||）这五种符号，人工切分出来视为独立主题词。去重处理后得到1059个一般主题词。本文选择这1059个教育类主题词作为检索入口词以获取Delicious系统的相关标签。

2.2 标签的选择与采集

Delicious是国外发展最早、最著名的分众分类系统，创始人Joshua Schachter称其为“a way to remember in public（以公共的方式来记忆）”。目前该标签系统拥有很广泛的用户群，国内外众多研究分众分类法的学者选择该标签作为研究数据。因此，本文也选择Delicious标签系统作为标签数据的来源。

我们于2013年3月将1059个教育类主题词作为标签，逐一输入到Delicious网站的检索框中，直接把检索结果页面右侧的相关标签中的所有标签复制下来。从检索的结果中我们发现有440个主题词没有搜索结果（该主题词没有作为标签进行过标注，并且不存在相关标签），508个搜索结果中该主题词用户作为标签标注资源并且存在相关标签，111个主题词用户没有用此术语标注过资源，但其存在相关标签。究其原因，是因为主题词是经过规范化处理的，以基本概念为基础的表达文献信息主题的词或词组。而Delicious标签是纯粹的自然语言，没有经过任何规范化处理，是用户根据自己的爱好、兴趣和对资源的理解，结合自身的语词功底作出的对资源的标识。这样，就会导致有部分教育类主题词在Delicious中没有检索结果。

2.3 标签数据的处理

将所有复制的标签数据粘贴到excel表格中统一进行处理，按检索主题词以列为单位排列。虽然选用delicious相关标签作为数据集，已经减少了部分噪声数据，但是仍然有一些对于本文无意义的标签存在于我们的实验数据中。标签中包括中文、英文、日文、特殊符号和其他语种等标签，因为实验中只需要中文和英文标签，所以需要去除掉那些非中、英文的标签。首先去除数字、标点、日语、德语、泰文等标签，其次把所有中文繁体字转化为中文简体字，大写英文标签转化为小写英文标签。过滤掉这些无意义的标签后，得到一个15266个中文和英文标签的数据集合。

把这个集合的中、英文标签分离后得到7264个中文标签和8002个英文标签。7264个中文标签去重后得到3928个中文标签，8002个英文标签去重后得到2754个英文标签。

中文标签与中分表教育类主题词进行重合度对比后得到了507个标签的教育类受控标签集，后面我们将重点对这个标签集进行分析。表1是出现频率在5次以上的受控标签。

英文标签与ERIC在线词表进行重合度对比后得到236个英文标签组成的受控标签集，表2为出现频率在5次以上的受控标签。

3 标签语义关系分析

3.1 标签语义关系界定

由于主题词表本身具有清晰的语义结构，所以可以

表1 频率较高的中文受控标签

教育 学习 中国 考试 大学 教育技术 社会 工作 儿童教育 教学 儿童 课程 小学 高考 研究生 日语 学校 中学 大学生 远程教育 教材 基础教育 文化 心理学 教育研究 管理 高等教育 教育改革 初中 中小学 开放教育 学生 幼儿园 应用 博士 中小学教育 教育部 试卷 高中 幼儿 方法 乡村教育 小学生 学习方法 研究 教育学 课件 学历 教学理论 网络教育 通识教育 教师培训 素质教育 双语教学 学校管理 平民教育 试题 法语 考生 学位 教案 电影 玩具 青少年教育 国外 职业教育 教学设计 夜校 应试教育 乡土教育

表2 频率较高的英文受控标签

education study chinese learning community design english business programming research language art music culture japanese reading development health humor politics lifestyle psychology resources technology books travel government teaching training children communication history statistics data information architecture economics food writing mathematics paper games association democracy educationalresources family literature marketing translation journalism languages scholarship medicine performance radio review security youth arts

从中抽取语义关系，但主题词表的语义关系简单，表达过于宽泛和模糊。例如主题词表最常见的等级关系就可能包含以下几种情况：整体和部分关系（如教育心理学和青少年心理学、学生心理学）；实例关系（如教育建筑和宿舍、食堂）等。同样，即使是等同关系也不是完全可以替代的关系，很多情况下是近义关系。至于相关关系，由于没有统一的标准，只是叙词之间除等同关系、等级关系之外语义相关的一种关系，所以基于主题词表的标签语义关系的抽取不能仅根据这三种语义关系来判定，需要对这三种关系进行调整。

我国学者魏来指出现有的研究还没有提出有效的标签同已有概念体系之间的映射规则、语义关联识别规则，并将标签之间的关系界定为同义、包含、通用、同源、实例、非相关6种关系[6]。本文在参考了一些关于标签语义关系界定的研究后将标签之间的语义关系调整为同义关系、近义关系、整部关系、实例关系、因果关系、成员关系、学科-研究对象关系、并列关系、主体-行为关系、行为-目的关系、主体-目的关系、相关关系。其中，叙词表中等同关系调整为同义关系、近义关系两种，等级关系调整为整部关系、实例关系两种，相关关系调整为因果关系、成员关系、学科-研究对象关系、并列关系、主体-行为关系、行为-目的关系、主体-目的关系、相关关系八种。

（1）同义关系：指同一概念的不同名称、全称与简称。例如：试题与试卷。

（2）近义关系：指意义相近但又不是同一概念的两个词之间的关系。例如：班主任与辅导员。

（3）整部关系：指概念之间整体与部分的关系。例如：学生与小学生，book与textbook。

（4）实例关系：指一个概念和概念的实体之间的关系，即个体作为类的成员与类建立联系。例如：课程与音乐课，languages与Japanese。

（5）因果关系：指两个概念之间是原因和结果的关系。例如：nutrition与obesity。

（6）成员关系：指机构、团体、组织或家庭与其组成人员之间的关系。例如：研究所与研究生，family与children。

（7）学科-研究对象关系：指学科或研究领域与研究对象、研究方法之间的关系。如教育学与学生，art与artist。

（8）并列关系：指同一类中性质不相容但又互相联系的关系。例如：英语课与俄语课，children与youth

（9）主体-行为关系：指主体与其行为、过程、动作之间的关系。例如：教师与教学，researchers与research。

（10）行为-目的关系：行为、动作、过程与其施动目的之间的关系。如考试与留学，play与games。

（11）主体-目的关系：主体参与了一系列行为后所要达到目的之间的关系。如教育部与教育改革，web sites与usability。

（12）相关关系：除以上关系之外的概念之间的互相关联。

3.2 中文标签语义关系的分析

把7264个中文标签与中分表教育类主题词进行比较，有1568个标签与主题词重叠，约占中文标签的21.6%，再基于中分表已有的概念之间的语义关系以及标签同主题词重合度的关系，识别标签之间的语义关系。识别步骤为：将标签集同主题词表进行映射，将能够直接进行词形映射的标签归入受控标签集，这个步骤在数据处理步骤已经完成；其次，按中分表的分类号对每个受控标签进行标注，这样方便于关系界定；最后，按照分类号可以界定出等同关系、等级关系，然后再进一步确定以上我们界定的关系。

根据以上步骤从叙词表中界定出词间关系，来构建标签的语义关系。通过对中文标签集的分析，共发现138对语义关系，其中，同义关系4对，近义关系14对，整部关系46对，实例关系8对，因果关系6对，成员关系7对，并列关系15对，主体-行为关系9对，行为目的关系7

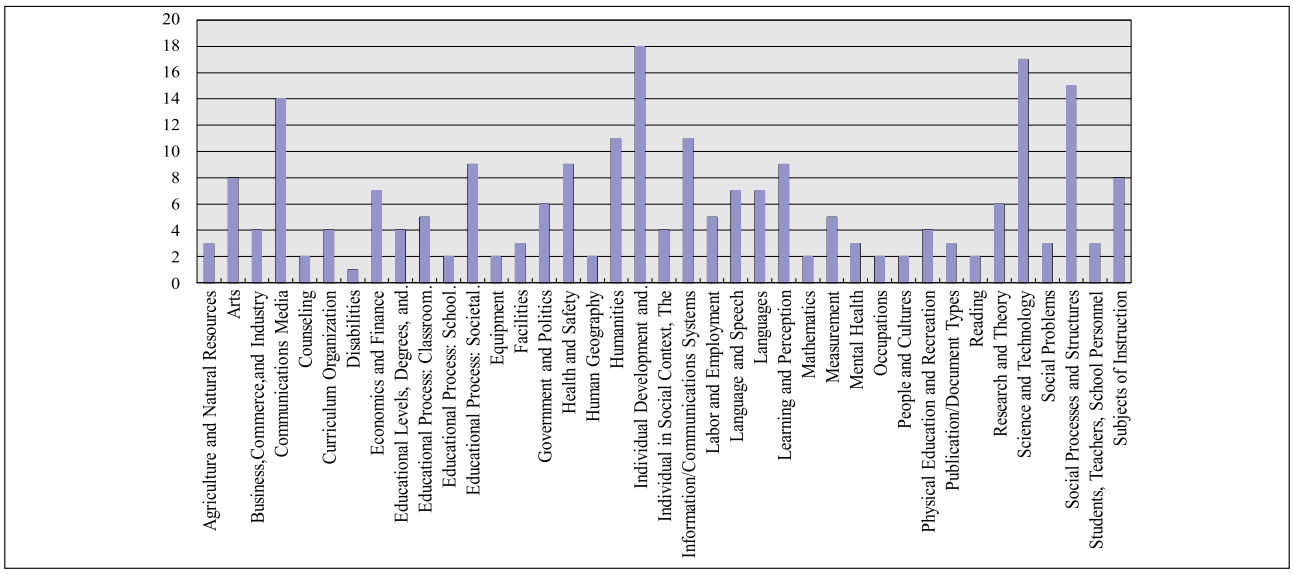

图1 ERIC类目对应的标签数

对，相关关系22对。

3.3 英文标签集的语义关系分析

ERIC（Education Resource Information Center），即美国教育资源信息中心，是美国国家级信息系统之一，包含超过1,300,000条记录和323,000多篇全文文档的链接，时间可追溯至1966年。ERIC主要给教育工作者、研究人员和公众提供教育知识和信息资源，目前是世界上最大的、访问量最大的教育数据库之一，包括几乎所有的教学科目，包括了艺术、计算机科学、数学、语言、健康、科学、物理、信息素养、社会研究、职业教育等多方面内容。

ERIC叙词表是一个与教育相关的主题词列表。主题词用来标引数据库中的文献和期刊论文，便于用户检索资源。ERIC词表将资源按学科分为41个大类，每个大类下分为若干子类，子类下面是具体的概念或术语。ERIC在线词表体现了两种语义关系，没有同义关系和近义关系，采用上位术语（broader terms）、下位术语（narrower terms）和相关术语（related terms）表示等级、相关关系。

本文使用ERIC叙词表作为英文标签的语义提取工具。首先把8002个英文标签与ERIC词表进行比较，有1642个标签与主题词重叠，约占英文标签的20.1%。去重后有236个不重复标签，其中有11个标签是ERIC类名主题词，233个标签分属39个类目（无Bias and Equitty和Tests and Scales这两个类目），图1为每个ERIC类目拥有标签的个数。

英文标签集的语义界定步骤与确定中文标签语义关系的步骤类似，但不需要标记分类号，只要从ERIC在线词表中找到其上位标签、下位标签、相关标签即可对其进行语义界定。整理后发现共有179对语义关系，其中包括44对上、下位类关系，135对相关关系，通过对英文标签集的分析，共发现整部关系40对，实例关系14对，因果关系8对，成员关系1对，学科-研究对象关系6对，并列关系24对，主体-行为关系4对，行为-目的关系5对，主体-目的关系3对，相关关系69对。从数据可以看出，英文标签的相关关系占比较大。

3.4 中英文标签提取关系对比分析

从英文语义关系的界定过程中发现，无法找到同义关系和近义关系，因为ERIC在线词表中只有上下位类关系和相关关系，所以无法提取出同义关系和近义关系。而在中分表中，相同类号的不同主题词可能表达的是同样的含义，所以可以视为同义关系或近义关系。

从数据中可以看到，一部分中文标签的语义关系和其对应的英文标签的语义关系是一致的，例如，中文标签教育和远程教育、英文标签education和distance education都属于整部关系。也存在一些不一致的关系，如中分表中教育与学习是等级关系，而在ERIC在线词表中education和learning是相关关系，因此我们进一

步将它界定为主体-目的关系。

本文没有对英文标签的中文释义对照中分表进行语义关系界定，进一步的工作需要对翻译后的英文标签对照中分表分析语义关系。对照《英汉教育词典》把236个不重复的受控英文标签进行翻译[7]。翻译后的中文释义与主题词不能完全一致，所以需要根据中文释义建立受控英文标签与《中国分类主题词表》的映射，找到同义主题词与之对应，如表3。

表3 使用主题词代替中文释义的英文标签

标签	《英汉教育词典》释义	使用主题词代替
textbooks	教科书	教材
Computer Science	计算机科学	计算机应用
curriculum	课程	课程论
courses	科目	课程类型
art education	艺术教育	美育
distance education	广播教育；远程教育	广播教学；远程教学
open education	开放教育	开放式教学
philosophy	哲学	教育哲学
English	英语	英语课
Performance	成绩	学习成绩
ethics	道德观	道德品质
Educational resources	教育资源；教育经济条件	教育资源；教育经济

使用同义主题词代替中文释义可以更好地发现标签之间的语义关系。英文标签的中文释义与教育类主题词和主题词的分词进行对照，有768个英文标签的释义（含一词多义）可以对应教育类主题词，占英文标签的9.5%。1151个英文标签的释义（含一词多义）可以对应教育类主题词的分词，占英文标签的14.4%。

4 结语

Delicious系统中相关标签集分析可以看出，这些标签之间具有明显的语义关系，主题词表的语义关系的明确也可以参考标签之间的关系。两者具有一定的互参照性。本文研究过程中，由于数据量处理较大，只选择存在于词表中的标签进行了语义抽取，忽略掉了其他的标签，从而造成一定语义关系的丢失。同时缺乏对中英文标签关系的统一处理，这些问题将作为进一步研究的重点。

参考文献

[1] SPITERI L. Controlled vocabularies and folksonomies [EB/OL]. [2014-03-10]. http://www.Collections canada.ca/obj/014005 /f2 /014005-05209-ee.pdf.

[2] PETER R J. User Tags versus Subject Headings Can User Supplied Data Improve Subject Access to Library Collections [J]. LRTS, 2009, 53(3): 171-184.

[3] 贾君枝.分众分类法与受控词表的结合研究进展[J].中国图书馆学报,2010(9):96-101.

[4] 贾君枝,王东元,王永芳.基于Delicious中文标签特征分析[J].情报科学,2010(10):1555-1567.

[5] 张宁.分众分类系统的用户行为特征分析[D].太原:山西大学,2013.

[6] 魏来.基于概念空间模型的folksonomy标签聚类方法研究[J].情报杂志,2010(4):137-142.

[7] 周南照,张学忠.英汉教育词典[M].北京:科学出版社,2003.

作者简介

贾君枝，女，1972年生，博士，山西大学经济与管理学院教授，研究方向：知识组织，E-mail：junzhij@163.com。
段佳慧，女，1988年生，山西大学经济与管理学院研究生，研究方向：知识组织，E-mail：752321101@qq.com。

Abstracting of Relationships of Semantic of Tags in Delicious Based on Education Thesaurus

JIA JunZhi, DUAN JiaHui
(School of Business and Management, Shanxi University, Taiyuan 030006, China)

Abstract: Semantic ambiguity and disorder of tags in the folksonomy have made retrieval efficiency low. The paper studies the new way of identifying the semantic relationship of tags according to relationships of terms in the therausus. The paper chooses the Chinese classfication therausus and ERIC online dictionalry as abstracting of relationships of semantic and analyses the overlap between tags and terms by searching realated tags in the Delicious. It shows abountiful relationships in the ralated tags.

Keywords: Tags; Theraurus; Semantic relationship

（收稿日期：2014-07-04）