轻型标签本体与受控词表的结合研究*

李艳[1]，贾君枝[2]
（1. 西安工业大学图书馆，西安 710021；2. 山西大学经济与管理学院，太原 030006）

摘要：分众分类系统中的标签通过一系列聚类算法可以形成“标签树”，但标签树中的标签间语义关系未能显性化，不能称之为标签本体。另一方面受控词表类目体系或主题词更新缓慢，跟不上网络资源新名词、新主题增长的速度，导致许多资源无法用传统分类法标引。借鉴受控词表现有的语义关系来挖掘标签树的语义关系，形成一个轻型标签本体；另一方面通过标签本体与受控词表的共享词汇，制定筛选规则，将标签本体中符合受控词表选词规则的标签纳入受控词表，使分众分类系统成为受控词表更新源泉之一，使其重新焕发活力。

关键词：分众分类；受控词表；标签本体；标签语义；主题词扩展
中图分类号：G254
DOI：10.3772/j.issn.1673—2286.2014.08.003

1 引言

随着Web 2.0时代到来，网民数量激增，网络资源越来越多，主题越来越多样化，传统分类法在类分网络资源方面越来越捉襟见肘。究其原因主要有以下两个方面：一方面是传统分类法的认知门槛比较高，属于专业类标引工具，对于普通网民而言，要想熟练掌握有一定的难度。另一方面是传统分类法的类目体系或主题词更新缓慢，跟不上网络资源新名词、新主题增长的速度，导致许多资源无法用传统分类法标引。随着Del.icio.us（美味书签）、LibreryThing等分享网站的出现，网络用户开始自己类分自己的资源，以方便自己使用。随着用户的增多，人们发现个体用户的分类法有趋同的趋势，所以将Folk和Taxonomy组合起来，形成了Folksonomy（分众分类法）。虽然该分类法能够弥补传统分类法的不足，但它存在无控性、平面性、分散性、模糊性等缺点[1]。这些缺点无疑将影响到用户检索效率。根本原因在于，标签就像散落在地面的树叶，很难找到它们之间内在的关系。为了解决网络资源的有效组织和网络用户检索的困惑，我们已经通过一系列的聚类算法将散落在网络上的标签构建成“标签树”[2]，我们希望通过挖掘标签语义关系，构建标签本体，将标签内在的语义关系显性化；同时希望将标签作为受控词表主题词的扩展源，使受控词表重新焕发活力。

2 国内外研究现状

调查国内外大量研究文献，研究发现围绕标签语义的研究主要集中在两方面：标签语义抽取、基于标签的受控词表发展研究。

2.1 标签语义抽取研究

标签语义抽取研究主要集中在两个方面：一是标签语义富集实证研究；二是标签与本体的关联研究。

在标签语义富集实证研究方面，Lux和Dsinger首先建立共现标签网络，然后利用标签及共现向量的测度，整合相似标签，尝试抽取标签中的语义关系[3]。国内学者周鑫、王军提出通过界定标签的概念外延来提取标签间的语义关系，并使用Del.icio.us的标签数据进行了验证[4]。朱晓晨等提出在协作标签系统中为资源和标签对应关系找到合适的上下文环境，逐步消除模糊语义[5]。张有志等人提出可以利用Folksonomy系统

* 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。

中包含的社会网络关系来提取标签语义，进而构建本体[6]。唐晓波等提出利用标签系统中的三元组（用户、标签、资源）构建三部图模型以挖掘标签概念间的语义关系，进而构建本体[7]。张云中等提出利用形式化背景和概念格这两种数据结构来构建标签本体[8]。

标签与本体的关联研究包括：（1）标签同SCOT的关联。Hak提出SCOT是最适合描述Folksonomy模型的本体，在SCOT和MOAT之间建立链接是补充标签意义的有效方式[9]。H. L. Kim等人在int.ere.st系统中使用SCOT本体描述用户的标签云结构，允许用户整合标签云，实现相似标签云的检索[10]。（2）标签同WordNet的关联。David等人将标签聚类为由WordNet驱动的等级结构，TagPlus系统使用WordNet消除标签的歧义，系统为用户返回一个标签在WordNet所有可能的意义供其选择[11]。（3）标签同Wikipedia的关联。Maria等人使用维基百科来处理标签中的一词多义的现象，并利用Wikipedia对标签进行聚类[12]。（4）标签与Google的关联。Qin Jian等人使用Google为标签挖掘语义环境，从而抽取标签的语义关系[13]。（5）标签同三者的关联。L. Specia等人通过Wikipedia、WordNet和Google确定概念标签的含义并识别概念标签之间的关系[14]。Martin等人使用WordNet、Google和Wikipedia联合进行标签过滤，首先通过WordNet过滤概念标签，接着通过Google过滤拼写正确的标签，最后通过Wikipedia过滤缩略语名称[15]。（6）标签同领域词典的关联。Hayman和Lothian介绍了基于传统分类法的分众分类法，利用受控词表来规范和控制标签[16]。魏来提出基于在线词表的标签语义关联识别的总体思路和具体规则，并利用教育类在线词表ERIC作为语义基础进行实证研究[17]。

2.2 基于标签的受控词表扩展研究

Piteri等人首次提出Folksonomy分类法与受控词表之间并不对立，利用Folksonomy系统中的标签资源可以补充和完善受控词表的词汇[18]。Rolla等人比较了LibraryThing网站的用户标签资源与LCSH（美国国会主题词表），认为用户的标签可以提高图书馆馆藏文献的检索效率，但不能取代受控词表，用户使用标签对书目进行标注后，在一定程度上能够弥补受控词表主题词的不足[19]。国内贾君枝教授就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数量众多的中文标签来解决国内受控词表老化的问题[20]。王东元等人分析了Del.icio.us系统中中文标签的特征，并将Del.icio.us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[21]。李婷等人抓取了豆瓣图书文学类中的标签进行特征分析，发现高频标签中有28%可匹配主题词表，45%的标签是作者信息，22%的标签是题名信息[22]。

综上所述，国内外学者基本认为Folksonomy分类法与受控词表之间并不对立，利用Folksonomy系统中的标签资源可以补充和完善受控词表的词汇。国外学者主要关注标签与WordNet、Google和Wikipedia的关联研究，而国内学者研究目前主要集中在标签自身特征的揭示和标签本体的构建上（包括基于受控词表来抽取标签语义），关于标签本体与受控词表的结合研究较少，特别是与传统标引工具《中国分类主题词表》结合研究还有待进一步深入。

3 标签本体与受控词表的结合研究算法

虽然我们已经将标签通过聚类算法形成n棵标签树，但树叶（标签）与树叶（标签）之间的内在关系并未显性化。我们希望借鉴同为树形结构的受控词表，一方面构建标签树的语义关系，形成真正意义上的轻型标签本体；另一方面吸收受控词表中富有活力的主题词进入标签树，增加标签树的用词规范性以及与受控词表的关联性。

有了标签本体，我们可以通过制定一定的筛选规则，将标签本体中的规范用词纳入受控词表，解决受控词表更新慢的问题，尽可能与网络资源新名词、新主题的增长同步，使受控词表重新焕发活力。

3.1 基于受控词表的标签语义关系挖掘

3.1.1 标签本体语义关系的设定

受控词表一定程度上被看作是一个轻本体，其语义关系主要有等同、等级、相关。依照受控词表的语义关系来设定标签本体的语义关系，一方面便于分析总结语义关系挖掘规则，另一方面也便于对受控词表进行扩展。

等同关系分别用大写字母Y、D、T来表示。Y表示

标签B是标签A的概念词，D表示标签A是标签B的概念词。T则表示标签B是标签A的译名。等级关系用大写字母S、F来表示。S表示标签B是标签A的上位类，即标签A继承标签B的一切属性，F表示标签B是标签A的下位类。相关关系用“C”来标示：表示标签A与标签B有一定的关联度，二者之间经常共现（两者标注的资源数都达到一定的阈值，且相似系数也达到一定阈值）。

3.1.2 标签与主题词匹配

将标签树上的标签与受控词表中的主题词进行完全匹配，得到起始受控标签集A（A_1，A_2，A_3，…，A_n）和非受控标签集B（B_1，B_2，B_3…，B_m）。标签集A为能够与受控词表直接映射的标签集合，即标签集A中的标签都存在于受控词表中。标签集B为虽然不能够与受控词表建立直接映射，但是是标签集A中的标签具有一定关联关系的标签集合，有作为新词添加到受控词表的可能性。

标签的非受控性导致受控标签集A存在数据稀疏的现象，我们利用受控标签从主题词集合Z（Z_1，Z_2，Z_3，…，Z_n）中抽取与该标签有语义关系的主题词，将其纳入受控标签集A中。即如果A_i在受控词表中与Z_j存在语义关系，则将Z_j纳入受控标签集A集合中，即$Z_j \in A$。如此重复，直到没有新的主题词Z_j出现为止。

3.1.3 识别受控标签间语义关系

受控标签的语义识别会出现以下两种情况：已知语义关系和未知语义关系。

（1）已知的语义关系

标签A_i可以直接映射到受控词表中的主题词Z_j，即$A_i=Z_j$；标签A_h也能直接映射到受控词表中的主题词Z_k，即$A_h=Z_k$；且Z_j与Z_k在受控词表中存在已知的语义关系（根据主题词间的关系或中图类号）。这时要分情况讨论：

如果Z_j与Z_k在受控词表中存在直接的语义关系，那么它们的语义关系就被赋予标签A_i和标签A_h。例如标签“教育→教学”，主题词“教育”对应的中图分类号为G4，主题词“教学”对应的中图分类号为G42，这两个主题词的语义关系为“F”，则标签“教育→教学”的语义关系标为“F”。

如果Z_j与Z_k在受控词表中存在间接的语义关系，则将受控词表中Z_j与Z_k之间的主题词和关系纳入受控标签集中。例如标签“教育→教育研究”，主题词“教育”的中图分类号为G4，主题词“教育研究”的中图分类号为G40-03，从分类号来看，两者是属分关系，但中间有一个主题词G40教育学，则我们将教育学纳入受控标签集中，将原有的标签“教育→教育研究”改造为“教育→教育学”和“教育学→教育研究”两对标签，两对标签间的关系均为“F”。

（2）未知语义关系

标签A_i可以直接映射到受控词表中的主题词Z_j，即$A_i=Z_j$；标签A_h也能直接映射到受控词表中的主题词Z_k，即$A_h=Z_k$；且Z_j与Z_k在受控词表中不存在已知的语义关系。但是A_i和A_h的相似系数$\lambda \geq 0.6$[2]（由标签聚类得出），这时可以从分类号和标注资源数两个方面来判断两者之间的关系：

分类号：Z_j和Z_k在受控词表中同属一个大类，则表示A_i和A_h不仅存在普遍的共现关系，而且从概念的角度来看也有一定的联系，则将该对标签的关系标识为“C”。例如标签“教学→小学”，该标签间的相似系数$\lambda \geq 0.8$，且教学的中图分类号为G42，小学的中图分类号为C62，则将该对标签的关系标志为相关关系“C”；

标注资源数：虽然Z_j和Z_k从中图分类号来看，属于不同的大类下，但A_i和A_h共同标注的资源数超过了200个，说明用户常常将两者标注到同一资源，且两者组配标注的资源是用户关注的热点，从而将该对标签的关系标注为相关关系“C”。例如标签“学习→英语”，“学习”标注了357个资源，“英语”标注了418个资源，而两者的相似系数为$\lambda \geq 0.8$，故而将该对标签的关系标注为相关关系“C”。

3.1.4 挖掘标签语义抽取规则

分析受控标签间已识别出的语义关系的特征，然后通过这些特征归纳总结基于受控词表抽取标签语义关系的规则，最后利用这些抽取规则，来判断识别出标签间的关系。

3.2 基于标签本体扩展受控词表

（1）标签与主题词的匹配

首先将构成标签本体的标签集T（T_1，T_2，T_3，…，T_n）与受控词表的主题词集Z（Z_1，Z_2，Z_3，…，Z_n）进行匹配，得到受控标签集A（A_1，A_2，A_3，…，A_n）；

（2）抽取待扩展标签

如A_x为受控标签，则将标签本体中与A_x存在语义关系的标签抽取出来，形成数据集A_x（T_y，R_{xy}），其中T_y指与A_x存在语义关系的标签，R_{xy}指A_x与T_y的语义关系。如$T_y \notin A$，则将T_y归为待扩展标签集中DT（DT_1，DT_2，DT_3，…DT_m）。例如标签“学校—双语学校”，“学校”是受控标签，“双语学校”是非受控标签，则“双语学校”归入待扩展标签集中。

（3）扩展标签筛选

我们认为能进入受控词表的标签应该是高质量的标签，既符合受控词表选词规范又属于热门标签，所以从词形、语法、使用频次等方面来对其进行考察。

词形筛选。标签的词形考察主要分为以下几部分：拼写规范，一般在构建标签本体的初期就通过电子词典和专用词汇表（如人名、地名及其他领域的专有名词表）等电子资源过滤掉不符合拼写规范的标签；字符数，以《中国分类主题词表》为例，我们对教育类主题词的字符数统计发现，其中2-6字词占到96%，所以扩展标签主要是2-6字词。因此将字符数作为标签的一个属性：DT_x（k），其中k为标签字符数。将不符合受控词表词形规则的标签剔除掉。以《中国分类主题词表》为例，标签“天主教教育工作者”字符数为9，则将其剔出待扩展标签集。

语法筛选。利用分词软件对剩余标签进行词性标注和分词处理，得到各个标签的分词结构和词性。受控词表的主题词主要由名词、动词、形容词以及它们的组合形成，故而我们只选择名词、动词、形容词以及它们组合的标签作为扩展词。将词性和分词结构作为标签的属性：DT_x（K，C，F），其中C为词性，F为分词结构。将不符合该字词词性或主要切分结构的标签剔出待扩展标签集。

流行度筛选。高频标签也叫热门标签，反映了用户对某一事物和概念的认同，也在一定程度上表明该类网络资源较丰富。所以通过设定一阈值，来筛选出热门标签作为扩展词。将标注资源数作为标签属性：DT_x（K，C，F，R），其中R表示标签标注资源数；随后人工设置标注资源数阈值R_λ，当$R < R_\lambda$，则将其剔出扩展标签集。例如我们设置标注资源数大于10的标签为热门标签，而“大学预科”只标注了2个资源，则说明网络上该主题的资源较少，则将其剔出待扩展标签集。

（4）扩展受控词表

根据受控词表的语义关系规则有选择地将扩展标签纳入受控词表。例如《中国分类主题词表》是树状结构，主题词至多有1个上位类。标签“计算机教育—虚拟大学”与词表中存在的“虚拟大学—学校”会导致出现“虚拟大学”存在两个上位类的情况，这在《中国分类主题词表》中是不允许的。

4 试验

4.1 试验对象的选取

我们通过聚类构建了教育领域的标签树，选取《中国分类主题词表》教育类主题词和ERIC叙词表（Education Resource Information Center）两大词表，ERIC叙词表是美国教育领域的在线数字图书馆按照学科组织的由与教育相关的词和短语构成的受控词表，包括近10000个教育领域的主题词[23]。我们尝试利用教育领域的英文词表（词汇更新快）来扩展标签本体。试验中根据聚类集构建的“中学”标签树为例，验证标签本体语义关系的挖掘和受控词表的扩展。

4.2 基于《中国分类主题词表》挖掘标签语义关系

4.2.1 标签与主题词匹配

“中学”标签树原有标签35个，通过中英文互译合并后剩下标签29个；通过标签与主题词的完全匹配后，得到起始受控标签18个，占标签总数的62%。见表1。

表1 标签分类表

受控标签	非受控标签
中学；教育；事件；中学生；英语；学习；社区；日语；犯罪；教师；小学；初中；数学；图解；教学；教育研究；中小学教育；学习	孩子；互联网教育；研究；目录；教育观察；生活；陈列；英语教学；门户网站；英语教师；网络；英语能力考试

为解决挖掘标签语义关系中数据稀疏的问题，最初的18个受控标签通过受控词表的语义关系从受控词表中抽取出172个受控标签，受控标签扩大为190个。

4.2.2 识别受控标签间语义关系

已知语义关系的识别过程中：直接语义关系共识别出195个语义关系，其中“S”关系5个，“F”关系135个，“D”关系30个，“C”关系24个；间接语义共识别出6个语义关系，其中“S”关系1个，“F”关系5个。

未知语义关系的识别过程中，共识别出“C”关系18个。

4.2.3 挖掘标签语义抽取规则

从词形来看，“F”(“S”)关系标签对有如下特征：一是标签A_i是标签A_h的一部分（A_h是比A_i更具体的概念），这样的标签对有87个，占“F”(“S”)关系标签对的62%。例如标签“教育→地方教育”。二是标签A_i和标签A_h有部分重叠，不重叠的部分已知是“F”(“S”)的关系，这样的标签对有23个，占F”(“S”)关系标签对的16%。例如标签“成人学校→成人中专”，其中我们知道标签“学校→中专”是“F”的关系。这两者合起来占到关系为“F”(“S”)标签对的78%，是“F”(“S”)关系标签对的重要特征。其余关系的标签看不出显著的特征。

我们利用上面总结出的“F”(“S”)关系标签对的特征来制定我们的关系抽取规则：

（1）首先我们选取标签间的相似系数$\lambda \geq 0.6$的标签作为待判断标签；

（2）每对标签进行字符匹配，如果标签$A_i(A_h)$是标签$A_h(A_i)$的一部分，则将该对标签的关系标示为“F”(“S”)；

（3）如果标签$A_i(A_h)$是标签$A_h(A_i)$有重叠，且不重叠部分构成的标签间刚好是“F”的关系，则将该对标签的关系标示为“F”(“S”)。

4.3 基于ERIC叙词表扩展标签本体

基于《中国分类主题词表》构建的标签本体与ERIC叙词表进行完全匹配，得到36个英文受控标签；依据这36个受控标签从ERIC叙词表中共计抽取出1134个主题词。以“education”为例，我们在ERIC叙词表中共抽取出“education”150个主题词，而在《中国分类主题词表》中我们只抽取出8个主题词，体现出ERIC叙词表的专业性、细粒度更高。对抽取出的1134个标签进行中文翻译后，我们选取关系更为紧密（除“C”以外的关系）的标签纳入标签本体中，最终我们共得到由478个标签、625个关系构成的标签本体。标签本体中各种关系的分布见图1。

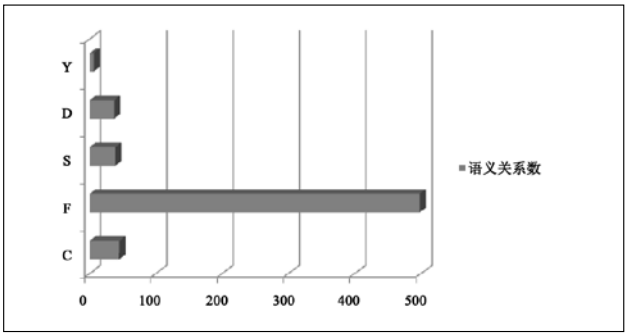

图1 标签本体语义关系统计图

4.4 轻型标签本体构建结果

利用“中学”标签树中的35个标签，基于《中国分类主题词表》和ERIC叙词表构建标签语义关系，最终得到的轻型标签本体包含543个标签，654个关系。基于《中国分类主题词表》和ERIC叙词表挖掘出标签树中36个标签语义关系，占标签树隐含关系总数的46%。最后我们通过本体构建软件protégé将抽取出的36个标签语义关系进行可视化显示。结果见图2。

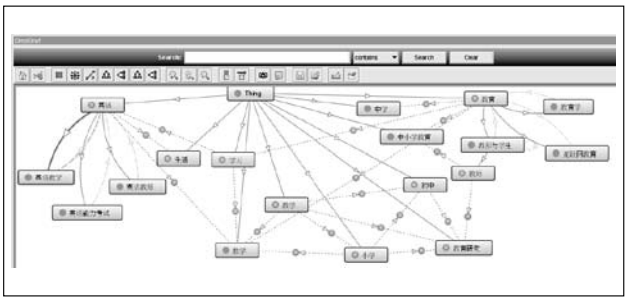

图2 标签本体可视化图

4.5 基于轻型标签本体扩展《中国分类主题词表》

将构建的轻型标签本体与《中国分类主题词表》的

表2 扩展标签集（节选）

标签	字符数	分词	分词结构	资源数
小学校	3	小学校/n	N	552
演示	2	演示/v	V	418
辅导	2	辅导/v	V	280
人群	2	人群/n	N	276
名师	2	名师/n	N	274
文科	2	文科/n	N	163
英语口语	4	英语/n 口语/n	N/N	112

5 结语

为了挖掘标签树的语义关系，我们设计了一套基于在线词表抽取标签语义关系，进而构建轻型标签本体的流程，并利用教育类的标签对该流程进行了验证。借助受控词表（《中国分类主题词表》和ERIC叙词表）中已有的语义关系，制定出了标签语义关系挖掘规则。同时构建出一个包括542个标签、634个关系的轻型标签本体。该轻型标签本体一方面可以为Folksonomy系统用户提供相关主题词推荐，另一方面可以利用标签本体的关系，来提升用户的检索效率。

表3 主题词扩展结果表（节选）

主题词	译名	类别	标签	分词结构	关系
教师	teachers	G451	代课教师	Substitute Teachers	F
学院	Colleges	G64	法学院	Law Schools	F
导师	Tutors	G643	辅导	Tutoring	F
学校	schools	G47	国际学校	International Schools	F
学习	learning	B842.3	合作学习	Cooperative Learning	F
教育	education	G4	互联网教育	e-learning	F
学校	schools	G47	教育机构	Educational Institutions	D

教育类主题词进行完全匹配，得到244个受控标签，通过这244个标签从轻型标签本体中抽取出269个带扩展标签。对待扩展标签集中的标签进行字符数计算，将不满足2-6字符的标签过滤掉；滤掉不符合切分模式或词性的标签；通过在Del.icio.us系统中检索剩余的待扩展标签标注的资源数，其中146个标签在Del.icio.us未标注资源，占总数的68%；有43个标签标注了10个以下的资源，占总数的19%；有28个标签标注了10个以上的资源，占总数的13%。根据二八法则，我们选取标注资源数β=10作为阈值，将$\beta\geq10$的28个标签（见表2）归入扩展标签集中。将扩展标签集中的标签作为受控标签，重复上面的步骤，结果没找到合适的标签进入扩展标签集。

最后我们选择扩展标签集中语义关系为“F”、“D”的标签作为最终的主题词扩展结果，最后共选出25个标签作为扩展主题词，具体见表3。

基于构建的轻型标签本体，我们提出一种基于标签本体扩展受控词表的算法，该算法通过一系列过滤规则，筛选出既在形式上符合《中国分类主题词表》，又在内容上属于热门标签（人们频繁使用的标签）的标签推荐给《中国分类主题词表》作为其扩展主题词。我们通过试验最终得到25个高频标签以及它们与主题词的关系作为《中国分类主题词表》教育类的扩展主题词，从而验证了该算法的有效性。

参考文献

[1] MIKA P. Ontologies are Us: a Unified Model of Social Networks and Semantics [C]// Web Semantics: Science, Services and Agents on the WorldWideWeb, 2007(3): 5-15.

[2] 李艳,贾君枝.基于向量空间模型的标签树构建方法[J].情报学报,2014,33(3):277-283.

[3] LUX M, DOSINGER G. From folksonomies to ontologies: Employing wisdom of the crowds to serve learning purposes [J]. International Journal of Knowledge and Learning, 2007(3): 515-528.

[4] 周鑫,王军.基于概念外延的语义关系挖掘方法[J].现代图书情报技术,2008(10):6-10.

[5] 朱晓晨,高飞.协作标签系统中的信息检索问题研究[J].电脑知识与技术,2008(5):602-604.

[6] 张有志,王军.基于Folksonomy的本体构建探索[J].图书情报工作,2008(12):122-125.

[7] 唐晓波,全莉莉.基于分众分类的本体构建探索[J].情报理论与实践,2008(6):931-936.

[8] 张云中.一种基于FCA和Folksonomy的本体构建方法[J].现代图书情报技术,2011(11):15-23.

[9] KIM H L. The state of the art in tag ontologies: A semantic model for tagging and folksonomies [EB/OL]. [2009-12-24]. http://scot-project.org/pubs/Kim_TagOnt.pdf.

[10] KIM H L, et al. Tag mediated society with SCOT ontology [EB/OL]. [2009-12-24]. http://www.cs.vu.nl/~pmika/swc-2007/SCOT.pdf.

[11] LANIADO D, et al. Using WordNet to turn a folksonomy into a hierarchy of concepts [EB/OL]. [2009-12-24]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.4393&rep=rep1&type=pdf#page=200.

[12] GRINEVA M. Harnessing Wikipedia for smart tags clustering [EB/OL]. [2009-12-24]. http://wwwnew.ispras.ru/en/modis/downloads/grineva02.pdf.

[13] QIN Jian, et al. Semantic relation extraction from socially generated tags: A methodology for metadata generation [EB/OL]. [2009-12-24]. http://edoc.hu-berlin.de/conferences/dc-2008/chen-miao-117/PDF/chen.pdf.

[14] SPECIA L, MOTTA E. Integrating folksonomies with the semantic web [EB/OL]. [2009-12-24]. http://www.eswc2007.org/pdf/eswc07-specia.pdf.

[15] SZOMSZOR M, ALANI H, CANTADOR I, et al. Semantic modelling of user interests based on cross folksonomy Analysis [EB/OL]. [2009-12-24]. http://arantxa.ii.uam.es/~cantador/doc/2008/iswc08.pdf.

[16] HAYMAN S. Taxonomy directed folksonomies: integrating user tagging and controlled vocabularies for Australian education networks [EB/OL]. [2009-12-24]. http://www.eswc2009.org/pdf/eswc09-specia.pdf.

[17] 魏来.基于在线词表的folksonomy语义关联识别方法研究[J].图书情报工作,2011(3):104-108.

[18] SPITERI L. Controlled vocabularies and folksonomies [EB/OL]. [2009-11-10]. http://www.Collections canada.ca/obj/014005/f2/014005-05209-ee.pdf.

[19] 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.

[20] 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.

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

[22] 贾君枝,李婷.分众分类与书目记录结合研究[J].情报理论与实践,2011(7):38-43.

[23] ERIC叙词表[EB/OL]. [2014-07-03]. http://www.eric.ed.gov/.

作者简介

李艳，女，1981年生，硕士，西安工业大学图书馆，研究方向：计算机信息检索。
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。

The Study of Combination of Light Label Ontology and Thesaurus

LI Yan[1], JIA JunZhi[2]
(1. Library of Xi'an Technological University, Xi'an 710021, China; 2. School of Economics and Management of Shanxi University, Taiyuan 030006, China)

Abstract: Folksonomy system labels can be formed "tag tree" through a series of clustering algorithm,but the semantic relationship between tag in the tree's be missed,"tag tree"can't be called tag ontologies. On the other hand,the category system and subject of thesaurus updated slowly, failed to keep pace with the growth of new network resources,and the new theme. This had led to many resources can't use thesaurus indexing. The paper mined semantic relationships of tag tree based on thesaurus, thus built a lightweight tag ontologies; On the other hand, through a shared vocabulary of tag ontology and thesaurus,making the filtering rules,the labels those matche the selection rules of thesaurus Were chosen to into thesaurus.So folksonomy system has became a source of thesaurus vocabulary update, revitalized it.

Keywords: Folksonomy; Thesaurus; Tag ontology; Semantic tags; Keyword expansion

（收稿日期：2014-07-04）