基于 Folksonomy 的本体构建综述

薛 涵[1,2],秦 兵[1],刘 挺[1]
(1.哈尔滨工业大学计算机科学与技术学院,黑龙江哈尔滨 150006;2.哈尔滨工程大学图书馆,黑龙江哈尔滨 150001)

摘 要: Folksonomy 是随着网络信息海量增长而迅速兴起的新型网络信息组织方式,与传统的信息组织方式本体结合的研究和应用价值正在逐渐受到人们的重视.在概述的基础上,基于 Folksonomy 对本体构建的核心问题展开论述,重点对比主流构建方法,归纳评价方法,并预测应用和研究趋势,以期对后续研究有所助益.

关键词: 本体; Folksonomy; 本体构建; 信息抽取; 评价
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2014)04-0791-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.04.026

A Survey of Ontology Construction Based on Folksonomy

XUE Han[1,2], QIN Bing[1], LIU Ting[1]
(1. *School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang* 150006, *China*;
2. *Harbin Engineering University Library, Harbin, Heilongjiang* 150001, *China*)

Abstract: Folksonomy is a new way of online information organizations. The combination of folksonomy and traditional ontology attracts more and more attention due to its research value and extensive applications. After the overview, we mainly survey the key tasks of ontology construction based on Folksonomy. We focus on the contrastive analysis of prevalent methods about construction, and then propose the evaluation metrics, and finally predict the important applications and future research, which is expected to be helpful.

Key words: ontology; Folksonomy; ontology construction; information extraction; evaluation

1 引言

目前,互联网上各种用户参与的网页、论坛、社交网络等数据,无意间构成大量具有研究价值的自然标注资源[1],标签就是其中一种典型的显式自然标注.通过用户自发为感兴趣的资源赋予标签,并与他人分享的过程,汇集大众标签形成的 Folksonomy(由"folk 或者 folks"和"taxonomy"组合而成,意为民间的分类法[2]),体现了用户的群体智慧,蕴含了丰富的语义知识,被视为社会化本体,受到网络用户和研究者们的青睐.

如果用散落树叶的随意堆积来形容 Folksonomy,那么传统本体就是一棵完整意义上的分类树.本体是得到大多数人认同的、关于概念体系明确的、形式化的规范说明[3],是自然语言处理、Web 信息检索、语义网、数字图书馆等领域的核心问题,是领域内部共同认可的知识概念化和形式化的规范描述,是不同主体(人、机器、代理)系统之间共享和重用的语义基础.

本体构建是本体研究的基础,主要存在以下问题:第一,本体构建通常需要领域专家的参与,并多以工具辅助手工方式来完成[4],耗时费力且不免主观;第二,如果要求推理,本体就要用建立在逻辑基础上的复杂表示语言而不是简单概念图的形式表示,理解和使用本体表示语言[5]对普通用户并非易事,也就无从为本体反馈和更新知识;第三,由于没有统一的评价标准,领域本体无法自适应到其他领域,共享和重用性不高.这些问题使我们思考寻找新的途径来研究解决传统本体构建的难题.

Folksonomy 从大量"草根"用户的标签中发现、获取和组织本体,及时更新和反映网络群体的共识,降低了本体构建和使用的门槛,有效缓解了知识获取的瓶颈,是解决传统本体构建问题最有效的途径.但是,Folksonomy 标签本身的随意性、模糊性和缺乏有效的组织,如果不与规范的传统本体相关联,对语义协助的贡献是有限的.本文旨在通过 Folksonomy 这种新资源研究本体构

收稿日期:2013-03-11;修回日期:2013-11-25;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.61133012,No.61273321);国家 863 高技术研究发展计划(No.2012AA011102)

建的核心问题,以期对传统本体构建中难题的解决和克服 Folksonomy 的缺陷有所助益,并对更多基于此问题的研究和应用提供支持.

2 Folksonomy 概述

Folksonomy 的概念由美国著名互联网专家 Thomas Vander Wal 于 2004 年首次正式提出,是基于"开放性元数据标准"和"自然语言社群聚类",采用自底向上的方法,依靠用户贡献和协同分类来实现.网络社区中每个用户都有一个自身对概念知识体系理解的"本体",但整个社会最初并没有定义统一规范的本体,用户之间通过不断的交流,理解对方的"本体",传播自己的"本体",逐渐形成统一的语义本体,常表现为标签云的形式.

图 1[6]形象地描述了这一过程,用户在 Folksonomy 系统中添加标签的行为将用户、标签、资源三者关联起来(实线),同时,也将用户和资源分别隐性关联起来(虚线),这些关联产生了丰富的可供挖掘的潜在语义信息,为本体构建提供可能.许多网站都提供为感兴趣的资源(如对象链接地址、图片、视频、书籍等)自由标注标签的功能.比如,美味书签 Delicious,照片分享 Flickr,目标分享 43things,学术参考资料 CiteULike,视频分享 YouTube,豆瓣,博客,微博等.

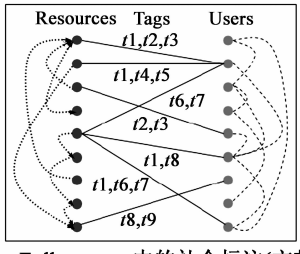


图1 Folksonomy中的社会标注(文献[6])

2.1 Folksonomy 的特点

Folksonomy 作为本体构建的新数据源,与传统本体构建的无结构文本,半结构网页和结构化数据库不同,具有以下优势:首先,实体(标签)已经被用户人工"抽取"出来,这样就避免传统本体构建中实体识别和抽取的一系列自然语言处理任务,如分词、词性标注、句法分析等;其次,发现能精确描述某领域主题的代表性文本语料是很困难的,尤其对于专业性强或者快速变化的领域[7],但是标签却可以更及时、自由地表征领域的热点主题和新词;再次,标注是在亿万网络用户浏览网页或发布信息的过程中轻松完成,虽然标注质量不能与付费的专业标注相比,但及时更新和零成本等优势使这种海量数据具有很大研究价值.

诚然,这种新数据源缺少明显可供挖掘关系的来源和应用的语境(如图 2,"遇龙河是漓江在阳朔境内最长的一条支流,位于桂林市."这样的背景知识和其中所涉及标签之间的关系,是独立于标签和图片之外的).标签词汇本身也存在无统一标准(如图 2,标签"风景"属于通用域的抽象概念;"阳朔"属于特定域的地名)、无层级结构(如图 2,"桂林"属于"中国"这种上下位的层次关系没有标明)等缺陷,给本体构建带来一些新问题.



(标签:中国,桂林,阳朔,漓江,遇龙河,喀斯特,风景)

图2 桂林山水(Flickr网站)

2.2 Folksonomy 的概念模型

根据 Folksonomy 形成过程中的标注行为,国内外学者展开一系列概念模型方面的研究.荷兰科学家 Mika[8]率先提出在概念和实例构成的传统本体中加入用户这一社会维度,形成适合表达 Folksonomy 的三元组本体模型 < actor, concept, instance > (ACI).在此基础上,后续出现一些较典型的模型.

(1)系统协同模型 Gruber[9]的理论模型 Tagging (object, tag, tagger, source, +/-),考虑到标签系统之间的协同标注扩展了 ACI 模型.object 指被标注的内容,tag 指用于标注的标签,tagger 指标注 object 的人或者代理,source 指标注行为所保存的系统,即标签的来源(如 Delicious 或 Flickr 等),+/- 的极性判别参数表示对标注行为是否认可.

(2)跨界协同模型 跨界协同模型支持跨系统、用户、资源之间共享标注行为.SCOT[10]是应用最广泛的扩展模型,采用 OWL 语言定义,从结构和语义两方面表达标签数据,并为异构系统数据的共享和重用提供互操作.它同时链接 SIOC, FOAF 和 SKOS 三个本体.SIOC 连接在线社会化服务网站,描述"标签-网站"以及"网站-网站"中的关系;FOAF 描述人或者机器代理,以及他们之间的关系;SKOS 从语义上把标签用属性关联起来.

(3)语义标注模型 语义标注模型,旨在提供机器可读的语义模型.MOAT(Meaning of a Tag)[11]在经典三元组模型的基础上,增加含义类表达习惯性或者用户自定义的标签含义,采用 OWL 语言形式化表示.

(4)数学模拟模型 数学模拟模型通过数学模型模拟用户的标注行为,辅助分析标签的语义.如将主题模型(LDA)的思想引入 Folksonomy 模型,出现了基于标

题词的标签-主题(TT)模型[12],用户-词-主题的(UWT)模型[13],以及用户-概念-实例-主题(ACIT)模型[14]. ACIT 模型融合了 ACI 和 LDA 模型的思想,引入隐含的主题层,避免了 ACI 模型三元超图到二元图的化简;同时,用户-标签-资源之间的三元关系被同步建模,比 TT 和 UWT 模型更有效.

可见,Folksonomy 概念模型基本组成元素已经达成共识,即用户、标签、资源. 基于 Folksonomy 构建本体,建议以数学模型为主,适当融合其他模型的思想.

3　基于 Folksonomy 的本体构建

基于 Folksonomy 构建本体的国内外相关研究可以分为三类:基于聚类的方法、基于已标注语义知识源的方法、基于聚类和已标注语义知识源结合的方法. 实验数据是支持 Folksonomy 的网站标注数据(如 Delicious),并无特殊要求.

基于聚类的方法主要通过分析元素的共现信息,包括标签-标签、标签-资源、标签-用户之间的共现,发现标签之间潜在的层次关系,但无法明确标签的语义和挖掘出标签间隐含的非层次关系. 基于已标注语义知识源的方法,通过标签与语义实体的关联,明确标签的含义和标签之间的关系,但依赖所引入的外部语义知识源的种类、规模、质量、更新和开放程度等因素. 基于聚类和已标注语义知识源结合的方法,通过不同结合方式实现两者优势互补,明确标签的含义和标签间的语义关系.

3.1　基于聚类的方法

根据聚类分析对象的不同,将研究方法分为三种:第一种分析对象仅含标签集合,忽略其他因素的影响;第二种分析对象是用户-标签集合,即考虑用户的影响,区分不同用户给出的标签;第三种分析对象是标签元数据,在语义角度上更细化.

(1)标签集合　Zhou 等人[15]受到物理和化学领域的启发,把 DA 模型用在标签集合的每一次分裂过程中,自动从标签空间浮现出层级语义. 从根节点包含所有标签开始,通过统计共现计算标签之间的语义相似度,循环迭代,不断分裂得到由"核心标签"准确表示出语义主题的一系列聚类. 与其他聚类算法需要指定聚类数目不同,DA 算法能够借助一个参数在分裂过程中自动控制聚类的数目和每个聚类的大小,确保每个节点的语义最终能通过一些标签来识别,缺陷是未能明确标签的语义和关系.

类似的方法还有,Begelman 等人[16]使用谱平分聚类在资源中发现了共享共现模式的标签组,但在每个组中标签的含义或者在不同组中标签之间的关系没有阐述. Schmitz[17]统计词频并基于父类包含子类的思想,建立类的层次体系. 方法基于统计技术,准确性会随数据量的增加而增加,但为了排除重叠关系需要反复遍历节点,效率较低. Heymann 等人的方法[18]是一种自顶向下的只利用标签共现关系的层次聚类方法. 这个方法简单高效,但父标签只是比子标签更宽泛的概念,并非严格的包含关系.

(2)用户-标签集合　这类方法区分不同用户提交的标签,辅助分析标签的含义和关系. 比如,Hamasaki 等人[19]将与用户直接相连邻居的标签作为用户的间接标签,克服某些用户标注信息缺失的问题. 假设一个标签被不同组的用户(或邻居)用于标注不同的资源,标签就可能有不同的含义,否则,标签仅有一个(或非常相似的)含义. 方法受到所选择用户组及其标注信息的影响.

(3)标签元数据　在聚类方法中,只有 Kennedy 等人[20]在标签的元数据信息上识别与位置和事件相关的标签. 该方法利用地理标注资源的纬度和经度以及时间戳信息,得到与每个标签相关的空间和时间的分布. 为了识别位置和事件标签,聚类算法应用到空间分布发现共享空间模式的标签组,应用到时间分布发现共享时间模式的标签组. 但是,方法仍然没有明确标签的含义.

3.2　基于已标注语义知识源的方法

按照是否需要用户的参与,分为自动方法和半自动方法.

(1)自动方法　通过标签与 WordNet、Wikipedia 等已标注语义知识源中的概念自动关联,用正式的语义丰富标签的语义.

Angeletou 等人[21]首先从 Flickr 数据集中过滤掉数字、特殊符号和非英文标签. 过滤后每个标签所有可能的单词描述(单数、复数或者复合标签的各种分解情况)作为标签的背景知识与标签一起映射到 WordNet. 如果一个标签对应 WordNet 中多个概念(即一词多义),从 WordNet 中抽取出每个概念的上下位词汇与该标签背景知识中的所有词汇进行相似度计算,相似度最高的概念作为标签的含义. 进而,将此概念在 WordNet 中的同义词和上下位关系词抽取出来,补充到标签的背景知识集合中. 然后,通过本体映射把扩展后的标签集合与语义网搜索引擎 Waston 中的语义实体关联,推理发现标签对之间的关系,构建 FLOR 标签本体. 实验证明,Folksonomy 通过已标注语义知识源能自动获取标签之间的语义关系并对标签消歧,精确率 93%,召回率 49%. 但是,许多标签无法映射到已标注语义知识源中的概念. 研究者继而考虑 Wikipedia 及相关的 DBpedia、YAGO 等开放联合构建的本体,克服由相对封闭成员组

维护的本体更新缓慢的问题. Garcia-Silva 等人[22]使用含有大量标签元数据的关联数据服务站点 TSR 检索与有歧义标签相关的 Wikipedia 候选页面的集合. 根据页面中出现最频繁的术语确定与标签最相似的 Wikipedia 页面. 由于 DBpedia 中每个概念都有一个页面对象属性,连接 DBpedia 概念和对应的 Wikipedia 页面,因此通过最相似的页面找到对应的 DBpedia 概念. 方法使用术语在对应 Wikipedia 页面出现的频率作为术语的权重,其实可以通过探索目前信息检索方法中更加复杂的术语权重如 TFIDF,或者关键词密度值来改善结果.

(2)半自动方法 当标签与已标注语义知识源的映射无法自动完成时,就需要用户参与. Passant[23]提出在标签创建时就应由标签的使用者赋予其语义. 假设标签有歧义,系统就会要求用户从多个概念中选择,如果没有合适的,用户可以向管理员提出增加新的概念,由管理员将此概念添加到已有本体中. 虽然让用户参与到整个标注过程是避免标签歧义最直接的方式,但方法只在公司内部测试,对于大量用户是否认可这种控制他们标注的行为还有待商榷.

3.3 基于聚类和已标注语义知识源结合的方法

这类方法的主要思想是融合基于聚类和已标注语义知识源两种方法,但是融合方式各有不同. 有的方法采取顺序结合的方式,即先聚类标签后关联语义知识源,或者先关联语义知识源后聚类标签;有的方法采取嵌入结合两者的方式,即在聚类方法中嵌入语义知识源的影响. 两种结合方式的出发点都是认为仅从标签本身不足以分析构建出令人满意的本体,需要进一步扩展标签的语境信息,深入分析语义.

(1)顺序结合 Specia 等人[24]以与其他标签的共现统计作为该标签的背景知识向量,对比寻找相关的标签. 对于相关标签的每个聚类,用户通过语义网搜索引擎(如 Swoogle)查找包含聚类中标签对的本体,并用找到的语义信息建立标签间的关系. 方法在训练时没有预先定义关系发现或者映射的类型,因而是无监督的. 但需要人工分析从语义网搜索引擎检索到的本体. Liu 等人[7]通过语义知识库为每个关键词推断概念集合并从搜索引擎获得该关键词的上下文信息. 然后,从知识库和上下文信息两方面丰富关键词的信息. 最后,基于贝叶斯的层级聚类方法自动归纳分类体系.

(2)嵌入结合 Giannakidou 等人[25]基于标签共现和语义相似度混合的策略,对标签和资源聚类,发现标签的语义. 其中,标签和资源之间的相似性是标签与标注该资源的所有标签之间的最大相似性. 两个标签之间的相似性是它们社会相似性(共现)和语义相似性(把标签映射到语义资源 WordNet)的加权之和. 资源和标签关联的二元图随后用光谱图聚类算法得到彼此不相交的聚类. 如果一个标签有多个含义,这个方法只能根据标签共现的模式识别出标签最常用的含义. Tomuro 等人[26]使用 DSCBC 聚类算法,将 Wikipedia 作为标签的特征先验,自动识别标签的 Wikipedia 领域,并把标签聚类到不同的含义,用层级聚合聚类自动构建标签本体. 方法考虑到标签消歧,但仅用到 Wikipedia 中文章的主题信息,如能结合利用资源结构抽取 Wikipedia 的相关研究,效果会更好.

综上,由于没有统一的评价标准,很难简单地说三类方法哪个更好,但根据目前聚类技术和已标注语义知识源的发展状况来看,两者结合的方法应该是研究趋势.

4 基于 Folksonomy 的本体评价

本体评价是运用科学的方法,遵循统一的标准和特定的指标体系,对本体的建设理念、应用需求、概念组织、功能设计和实际运行状况等进行综合测评的过程和方法[27]. 缺少统一的测试平台和标准是传统本体和基于 Folksonomy 构建本体的共同难题. 基于 Folksonomy 的本体评价方法可归纳为以下四类.

(1)人工评价 根据人工预先设定的标准对 Folksonomy 学习产生的本体打分,通过协调不同评价人员的结果得到最终的评分. 如 Tang 等人[12]就要求社会标签的提供者直接评价系统最终产生的标签本体,对是否满意标签之间的关系打分,精确率为满意数与反馈总数的比率.

(2)语义评价 将 Folksonomy 学习产生的本体与参考本体使用词召回率(术语或者概念分别在两者对应目录中出现的数目)和分类法重叠率(父亲-孩子对处于正确顺序的数目),还有增强的精确率和召回率以及 OntoRand 进行对比[28,29]. 参考本体可以是普遍认可的标准本体,如 ODP、LACC、DC、Wikipedia、WordNet、中图分类法等;也可以是数据来源本体,如 Ali 等人[14]就是将基于 Folksonomy 学习产生的本体与来源本体对比,评价提出的本体构建方法;还可以是较先进技术生成的本体.

(3)实用评价 将 Folksonomy 学习产生的本体应用到具体任务中,通过对任务支持效果的评价,间接评价生成本体的质量. 如 Helic 等人[30]把 Folksonomy 算法和网络中的分散搜索关联,提出一个新颖的框架,从实用的角度评价不同算法得到的 Folksonomy 本体,核心思想是把标注系统的探索导航任务构建成标签网络的分散搜索,从 Folksonomy 中学习产生的本体作为搜索时的背景知识,通过对探索导航任务的支持效果间接评价学习本体的质量.

(4)分布评价 通过概念的分布效果评价 Folksonomy 学习产生的本体. Tomuro 等人[26]提出本体密度的概

念,认为所有被赋予同一个资源的语义相关的标签应该被置于本体中靠近彼此的位置上.通过测量本体中赋予资源的标签间连接边的数目,估计资源的本体密度,就可以统计出所有资源的平均密度,继而就能评价本体的质量.

总之,研究者可以根据实际选择其中一种或者综合多种方法来评价.人工评价简单常用,但是评价指标的设定和评价过程都依赖人工参与;语义评价关键是参考本体的选择,随着可用参考本体的丰富,是较为通用的评价策略,有助于本体评价的规范化;实用评价多用于应用需求驱动的任务本体,但对于以知识积累为目标的本体,增加了评价的难度;分布评价直观新颖,值得深入探索更多用于描述本体质量的评价指标.

5　基于 Folksonomy 的本体应用

基于 Folksonomy 构建的本体,是传统本体和社会化本体的折中,虽然是初步的、轻量级的本体,但应用也很广泛.

(1)用于揭示网络资源　近几年国外研究人员非常关注 Folksonomy 对网络信息资源组织产生的影响,积极开展实证研究,如 Int.ere.st 项目、FLOR 项目等,通过丰富 Folksonomy 的语义信息,优化网络信息资源组织,便于发现和重用.

(2)用于个性化检索　陈毅恒等人[31]通过本体对标签进行抽取和筛选,实现了对检索结果聚类的优化.张志强等人[32]的研究说明,除了分析搜索日志,还可以结合 Folksonomy 构建的本体实现更好的个性化服务.

(3)用于辅助自动标注　标注数据是许多研究领域的宝贵资源,但人工标注成本高,自动标注准确率低.司献策等人[33]用基于 Folksonomy 构建的本体辅助自动标注,捕获到短文本的言外之意.

可见,基于 Folksonomy 的本体构建研究,在理论和实践上都极具前瞻性的探索空间.

6　结论和展望

Folksonomy 技术兴起于国外,目前已拥有较成熟的概念模型和方法,而且也举办多次以此技术为主题的学术会议,如欧洲机器学习与数据挖掘会议(ECML/PKDD)2008 年垃圾标签用户的识别任务,2009 年标签推荐在线评测等.与国外相比,国内对于 Folksonomy 的研究工作有待深入.多数从概念认识、特征层面研究,逐渐出现社会标签分析与推荐系统方面的研究[33],但与本体构建相关的研究[12,14,32]较少,尤其研究中文本体构建的文献截止到本论文完成还没有调研到.

本文对基于 Folksonomy 构建本体的关键问题展开论述,包括 Folksonomy 的特点和概念模型的概述,本体构建主流方法的对比分析,评价方法的归纳及应用成果的阐述等.从文献调研来看,基于 Folksonomy 构建本体的方法论、表示语言(工具)与传统本体构建并无差别,故不在此赘述.此外,我们认为还有一些具有价值的研究点,列举如下,希望对其他研究者有所启发.

(1)传统本体与 Folksonomy 的融合　传统本体和 Folksonomy 这两种本体既有相似又有不同.如何将两者结合,发挥彼此的优势,相互促进,是本体研究者共同努力的方向.标签语义信息抽取文本和半结构化数据中相对成熟的信息抽取技术,如何调整和应用到 Folksonomy 这个新的数据源是要进一步探讨的.

(2)本体进化　动态更新是 Folksonomy 数据源的一大优势,如何在不重新构建本体的前提下,将局部的更新反映到本体概念和关系的变化中,是值得深入思考的.

(3)跨语言　对跨语言的研究,不仅可以借助国外先进的技术和成果,也有益于解决中文本体构建资源不足的问题,是很有意义的研究方向.

参考文献

[1] 孙茂松.基于互联网自然标注资源的自然语言处理[J].中文信息学报,2011,25(6):26 – 32.
Sun Maosong. Natural language processing based on naturally annotated web resources[J]. Journal of Chinese Information Processing,2011,25(6):26 – 32.(in Chinese)

[2] Trant J. Studying social tagging and folksonomy: A review and framework[J]. Journal of Digital Information,2009,10(1):1 – 42.

[3] Gruber T R. Toward principles for the design of ontologies used for knowledge sharing[J]. International Journal of Human Computer Studies,1995,43(5):907 – 928.

[4] 邓志鸿,唐世渭,等.Ontology 研究综述[J].北京大学学报(自然科学版),2002,38(5):730 – 738.
Deng Zhihong, Tang Shiwei, et al. Overview of ontology[J]. Acta Scientiarum Naturalium Universitatis Pekinensis,2002,38(5):730 – 738.(in Chinese)

[5] 岳静,张自力.本体表示语言研究综述[J].计算机科学,2006,33(2):158 – 162.

[6] Marlow C, Naaman M, Boyd D, et al. HT06, tagging paper, taxonomy, Flickr, academic article, to read[A]. Proceedings of the Seventeenth Conference on Hypertext and Hypermedia[C]. ACM,2006. 31 – 40.

[7] Liu X, Song Y, Liu S, et al. Automatic taxonomy construction from keywords[A]. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. ACM,2012. 1433 – 1441.

[8] Mika P. Ontologies are us: A unified model of social networks

and semantics[J]. The Semantic Web-ISWC, 2005, LNCS (3729):522 – 536.
[9] Gruber T. Ontology of folksonomy: A mash – up of apples and oranges[J]. International Journal on Semantic Web and Information Systems (IJSWIS), 2007, 3(1):1 – 11.
[10] Kim H L, Breslin J, Yang S K, et al. Social semantic cloud of tag: semantic model for social tagging[J]. Agent and Multi-Agent Systems: Technologies and Applications, 2008, LNCS (4953):83 – 92.
[11] Passant A, Laublet P. Meaning of a tag: a collaborative approach to bridge the gap between tagging and Linked Data [A]. Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008)[C]. Beijing, China, 2008. 1 – 5.
[12] Tang J, Leung H, Luo Q, et al. Towards ontology learning from folksonomies[A]. Proceedings of the 21st International Joint Conference on Artifical Intelligence[C]. Morgan Kaufmann Publishers Inc, 2009. 2089 – 2094.
[13] Bundschus M, Yu S, Tresp V, et al. Hierarchical bayesian models for collaborative tagging systems[A]. Ninth IEEE International Conference on Data Mining[C]. IEEE, 2009. 728 – 733.
[14] Ali D, Juanzi L. Modeling ontology of folksonomy with latent semantics of tags[A]. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology [C]. IEEE Computer Society, 2010. 516 – 522.
[15] Zhou M, Bao S, Wu X, et al. An unsupervised model for exploring hierarchical semantics from social annotations[J]. The Semantic Web, 2007, LNCS (4825):680 – 693.
[16] Begelman G, Keller P, Smadja F. Automated tag clustering: improving search and exploration in the tag space[A]. Collaborative Web Tagging Workshop at WWW2006[C]. Edinburgh, Scotland, 2006. 15 – 33.
[17] Schmitz P. Inducing ontology from flickr tags[A]. Collaborative Web Tagging Workshop at WWW2006[C]. Edinburgh, Scotland, 2006. 1 – 4.
[18] Heymann P, Garcia-Molina H. Collaborative creation of communal hierarchical taxonomies in social tagging systems[R]. Info Lab Technical Report, 2006. 1 – 5.
[19] Hamasaki M, Matsuo Y, Nishimura T, et al. Ontology extraction using social network[A]. International Workshop on Semantic Web for Collaborative Knowledge Acquisition[C]. Hyderabad, India, 2007. 1 – 6.
[20] Kennedy L, Naaman M, Ahern S, et al. How flickr helps us make sense of the world: context and content in community – contributed media collections[A]. Proceedings of the 15th International Conference on Multimedia[C]. ACM, 2007. 631 – 640.
[21] Angeletou S, Sabou M, Specia L, et al. Bridging the gap between folksonomies and the semantic web: an experience report[A]. the 4th European Semantic Web Conference[C]. Innsbruck, Austria, 2007. 2 – 15.
[22] Garcia-Silva A, Szomszor M, Alani H, Corcho O. Preliminary results in tag disambiguation using dbpedia[A]. 1st International Workshop in Collective Knowledge Capturing and Representation (CKCaR09)[C]. California, USA, 2009. 1 – 9.
[23] Passant A. Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs[A]. Proceedings of International Conference on Weblogs and Social Media[C]. Boulder, Colorado, 2007. 1 – 8.
[24] Specia L, Motta E. Integrating folksonomies with the semantic web[J]. The Semantic Web: Research and Applications, 2007, LNCS (4519):624 – 639.
[25] Giannakidou E, Koutsonikola V, Vakali A, et al. Co-clustering tags and social data sources[A]. The Ninth International Conference on Web-Age Information Management[C]. IEEE, 2008. 317 – 324.
[26] Tomuro N, Shepitsen A. Construction of disambiguated folksonomy ontologies using wikipedia[A]. Proceedings of the 2009 Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources[C]. ACL, 2009. 42 – 50.
[27] Brank J, Marko G, Dunja M. A survey of ontology evaluation techniques[A]. Proceedings of the 8th International Multi-Conference on Information Society[C]. Ljuljana, slovenia, 2005. 166 – 169.
[28] Plangprasopchok A, Lerman K, Getoor L. Growing a tree in the forest: constructing folksonomies by integrating structured metadata[A]. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [C]. ACM, 2010. 949 – 958.
[29] Dellschaft K, Staab S. On how to perform a gold standard based evaluation of ontology learning[J]. The Semantic Web-ISWC, 2006, LNCS (4273):228 – 241.
[30] Helic D, Strohmaier M, Trattner C, et al. Pragmatic evaluation of folksonomies[A]. Proceedings of the 20th International Conference on World Wide Web[C]. ACM, 2011. 417 – 426.
[31] 陈毅恒,秦兵,等.基于 ontology 抽取优化初始选择的检索结果聚类[J].电子学报,2008,36(12A):166 – 170.
Chen Yiheng, Qin Bing, et al. Search result clustering based on centroid optimization by ontology extraction[J]. Acta Electronica Sinica, 2008, 36(12A):166 – 170. (in Chinese)
[32] 张志强,孟庆海,谢晓芹.个性化的社会标签查询扩展技术研究[J].计算机科学与探索,2010,4(9):812 – 829.
Zhang Zhiqiang, Meng Qinghai, Xie Xiaoqin. Research on personalized social tag query expansion techniques[J]. Journal of Frontiers of Computer Science and Technology, 2010, 4 (9):812 – 829. (in Chinese)
[33] Xiance Si, Zhiyuan Liu, Maosong Sun. Modeling social anno-

tations via latent reason identification[J]. IEEE Intelligent Systems, 2010, 25(6): 42 – 49.

作者简介



薛　涵　女,1982 年生于河南开封.哈尔滨工业大学计算机科学与技术学院博士研究生.研究方向为信息抽取、社会计算、本体构建等.
E-mail: hxue@ir.hit.edu.cn



秦　兵　女,1968 年生于黑龙江哈尔滨.哈尔滨工业大学计算机科学与技术学院教授,博士生导师.研究方向为自然语言处理、文本挖掘、情感分析等.
E-mail: qinb@ir.hit.edu.cn



刘　挺　男,1972 年生于黑龙江哈尔滨.哈尔滨工业大学计算机科学与技术学院教授,博士生导师.研究方向为自然语言处理、文本挖掘、文本检索等.
E-mail: tliu@ir.hit.edu.cn