支持 Web 2.0 标签层次体系构建的关系识别及层次组合方法研究

高克宁[1],张 引[2],张 斌[2],张聿博[2]
(1.东北大学计算中心,辽宁沈阳 110004;2.东北大学信息科学与工程学院, 辽宁沈阳 110004)

摘 要: 从标签系统中生成层次体系可以支持多种类型的应用,具备重要的意义.当前的研究主要集中于发现标签间的关系,但对如何利用这些关系形成高质量的层次体系却关注不足.针对这一现状,研究了支持 Web 2.0 标签层次体系构建的关系识别及层次组合方法,通过分析并识别已发现的标签间关系所具有的不同类型提升了标签间关系的质量,并提出基于语义流动分析的层次组合方法实现了更高质量的层次体系构建.应用多种评估指标的实验结果表明应用关系识别及使用语义流分析方法可以获得相比评估方法更高质量的标签层次体系.

关键词: Web 2.0; 社会标注; 标签层次体系; 关系识别; 层次组合
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2014)01-0058-04
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.01.009

Research on Relation Recognition and Hierarchy Composition for Web 2.0 Tag Hierarchy Construction

GAO Ke-ning[1], ZHANG Yin[2], ZHANG Bin[2], ZHANG Yu-bo[2]
(1. *Computing Center, Northeastern University, Shenyang, Liaoning* 110004, *China*; 2. *College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning* 110004, *China*)

Abstract: Tag hierarchies generated from social tagging systems can be used in various fields. Current research works mainly focus on the detection of tag relations, but pay much less attention on how to generate high quality tag hierarchy with these relations. This paper studies relation recognition and hierarchy composition methods for Web 2.0 tag hierarchy construction. By analyzing and recognizing the types of the tag relation detected by available methods, quality of tag relation is improved. Moreover, a semantic flow analysis based hierarchy composition method is proposed to acquire tag hierarchy with higher quality. Experimental results measured by various evaluation metrics have shown that by applying relation recognition and semantic flow analysis based hierarchy composition, hierarchies with higher quality with respect to other evaluated methods can be yielded.

Key words: Web 2.0; social tagging; tag hierarchy; relation recognition; hierarchy composition

1 引言

从标签系统构建的层次结构可用于支持多种任务[1~3].当前的标签层次体系构建方法包含三个步骤:建模标签语义;发现标签间关系;形成体系结构.Helic 等[4]使用标签所标注的资源、资源覆盖的相似度及标签的"中心度"实施上述步骤,Eda 等[5]则使用 PLSI 潜在主题、主题分布熵的差值及图分配来形成层次结构.

现有标签层次体系生成方法的效果仍非常有限,这一方面受限于标签自身的质量[6],另一方面也由于现有研究集中于发现标签的语义及标签间关系[7~9],但对如何形成层次体系关注不足.现有研究通常直接组合标签间关系形成层次结构,其存在着两方面不足:首先,标签间关系被作为形成上下位关系的唯一依据.然而,标签间通常存在着多种类型的关系,对这些复杂关系的过分抽象必然影响着体系构建的质量.其次,当前的层次组合方法仅考察成对标签间关系,无法利用更多标签所形成的标签组中蕴含的关于如何形成层次结构的信息.这两方面的问题影响着利用标签关系组合而成的标签层次体系结构的质量.

本文研究 Web 2.0 标签层次体系构建中关系识别及层次组合方法.针对当前方法仅关注特定类型偏序

收稿日期:2012-11-15;修回日期:2013-04-27;责任编辑:李勇锋
基金项目:国家自然科学基金(No.61073062);辽宁省科学技术基金(No.20102060);沈阳市科学技术计划(No.F11-264-1-33)

关系的问题,本文识别标签间关系所对应的多种语义关联,提出了避免形成错误上下位关系的关系识别方法;针对当前方法仅关注成对标签间关系的问题,本文考虑了在局部范围内一组标签之间的语义流动情况,提出基于语义流的层次组合方法.实验表明,本文提出的方法可以有效的改善标签层次体系构建的质量.

2 相关研究

很多研究关注利用标签系统构建层次体系[1,2],其包括三个主要的步骤:建模标签语义;发现标签间关系;层次组合形成体系结构.

语义建模是标签层次体系构建的基础.标签可以建模为对应资源的集合[10],这一方法被很多研究采用[2,8,11],其优点在于可以直接应用 Latent Semantic Indexing(LSI)、Latent Dirichlet Allication(LDA)等进一步提取标签语义信息. García-Castro 等[12]则提出了 HyperTag 模型考察更多类型的对象来建模标签的语义信息.

标签间关系发现利用标签的语义模型发现标签之间的偏序关系.关联规则挖掘是一种基本的标签关系发现方法[10].类似还有同样使用集合关系判断语义距离的方法[11].Rank 方法[2,4,7]及预定义的规则[9]也可以用于发现标签间关系.标签层次关系组合将标签间关系组合为一个完整的体系结构.一类被广泛采用的层次关系组合方法是基于图论的方法[2,11,5].依据标签语义建模及关系发现方法的不同,各个研究也使用不同的层次关系组合方法,如基于聚类的方法[7,3]等.

3 标签间关系分析与识别

3.1 标签间关系的层次语义关系分析

为了分析标签间关系的类型,本文在一个包含 2000 个资源与 3398 个标签的数据集上分别使用资源(TR)、标签共现(TT)、形式概念分析(TC)、LDA 潜在主题(TTP)、正文词汇(TW)五种对象建模标签语义,并使用关联规则挖掘发现标签间关系.通过人工分析发现标签间关系所蕴含的层次语义关系包括:(1)严格上下位关系(>);(2)非严格上下位关系(→):下位概念标签可能有其他的上位概念;(3)自由层次关系(<>):上位概念和下位概念在不同语境下能够自由调换;(4)同层关系(‖);(5)同义关系(=);(6)无明显关系(·).

这些层次语义关系还可以进一步的分为如下两种类型:(1)必须被正确识别的基础层次语义关系,包括严格上下位及同层关系;(2)不会引起严重错误的辅助层次语义关系,包括非严格上下位、自由及同义关系.基于此,有必要研究上下位关系和同层关系的识别方法.

3.2 标签关系的层次语义识别

识别同层关系的一种方法是使用人工构建或借助语义资料自动构建的可查询同层关系的字典.虽然列举出所有同层关系比较困难,但列举出同层的标签却相对较为容易.在一个特定的领域内,这些同层标签比较容易归纳.而借助维基百科等语义资料,这些字典还可以被自动或半自动地构建出来.特别的,由于层次结构中的上层结构在实际应用中更加重要,因此同层关系字典不一定需要覆盖整个领域,而只需纠正关键的错误上下位关系,提升上层层次结构的质量即可.

标签关系的层次语义还可以被视为一个对标签间关系进行分类的过程,并使用分类学习方法进行识别.

4 基于语义流的关系组合方法

基于语义流的关系组合方法的核心是对语义流动的建模.为了便于陈述,这里使用资源建模标签语义,使用下位概念资源的数量占上位概念资源数量的百分比(即分支率)描述语义流动.通过查找最大的语义流动方向,识别标签的下位概念,构建标签层次体系,则选取下位概念的目标为:(1)下位概念对应的资源集合之间的交集尽可能小;(2)标签相对于下位概念的分支率的和尽可能大.

目标(2)表明在选取最相邻下位概念时应优先考虑分支率较大的标签.语义流分析法需要以下两个阈值参数:(1)分支率阈值:选取分支率最大的标签关系时,对分支率的下限进行限制;(2)相似度阈值:近似不相交实例集合间相似度的上限.

在基于资源的标签语义表示中,基于语义流的关系组合从实例集最大的标签开始,选取分支率最高且与已选标签实例集合近似不相交的标签作为最邻近子概念,算法流程如算法 1.

算法 1　基于语义流动分析的标签关系组合方法算法

输入:概念关系集合 R,分支率阈值 α,相似度阈值 β
输出:标签层次关系集合 RT
Begin
将实例最多的标签 $t0$ 加入队列 Q,TA←$\varnothing$
While $Q\neq\varnothing$
　从 Q 的队首取出 ti,TC←$\varnothing$
　从 R 中选取 ti 作为父概念的 rk 构成 Rp
　将 Rp 按照 $BR(ti,tj)$由大到小排序,tj 为子概念
　For each $ri\in Rp$,其中 $ri=<ti,tj>$
　　If $tj\notin TA$ 且 $BR(ti,tj)>\alpha$ 且 $\forall tk\in T$ 满足 $Sim(ti,tk)<\beta$
　　Then 将 ri 加入 RT,将 tj 加入 TC 和 Q 的队尾
　　End If
　End For
End While
End

5 实验

5.1 实验设定

实验使用了 3.1 小节的数据集,并人工对一组关键节点构建了关系标准树.受限于篇幅,标准树及本节所有实验的完整结果可以从作者的主页* 获取.

实验采用的第一组指标借鉴 F1 指标提出带权平均 F1(WA-F1)指标,其先将树中每一个节点作为单独的评估对象,计算其 F1 值.整个结果树的 F1 值则为各个节点 F1 值的加权平均值,即:

$$\text{WA-F1} = \frac{\sum \text{F1}(n) \times \text{Weight}(n)}{\sum \text{Weight}(n)}$$

$$\text{Weight}(n) = 2^{-\text{Level}(n)}$$

其中 F1(n)表示 n 的 F1 值,Level(n)表示节点所在层次.WA-F1 不仅要求上下位概念关系正确,而且需要每个节点的下位概念尽可能全面,因此能够准确的反映构建结果树的整体质量.第二组评估指标采用了 Solskinnsbakk 等[13]提出的四个层次体系质量评估方法.

实验进一步进行分层的质量评估,即对于一个深度为 N 的层次体系,实验针对从 1 到 N 的每一个 k,分别计算深度为 k 的子树的 WA-F1 及第 k 层的 Solskinnsbakk 指标.

标签语义表示以及标签间关系发现数据也使用了 3.1 小结中的得到的结果,通过使用不同类型的语义实例以及层次组合方法,实验评估应用本文提出的关系识别方法前后得到的标签层次体系质量.为了评估本文提出的层次组合方法,实验对比了刘等[7]提出的标签层次体系构建方法.

5.2 标签关系识别效果分析

实验首先对比应用关系识别前后标签层次体系的质量.除本文提出的层次组合方法外,实验还对比了基于最小生成树及关系组合的层次组合方法.其中关系组合方法的构建过程为:对于每个标签,选取潜在父概念集合;从父概念集合中选取实例集合最小的标签作为父节点.

图 1 给出了部分实验结果.实验表明,除极少数情况外,应用标签关系识别可以帮助提升各个方法所得到的标签层次体系的质量,且这种提升在不同深度的子树下均有体现.

表 1 给出了应用 Solskinnsbakk 等的指标的部分实验结果.针对 Solskinnsbakk 等的理论 1,标签关系识别在最小生成树的 *TR* 与 *TC* 设定上表现出了对层次体系质量的改进效果,且在这两者上理论 1 与 WA-F1 指标给出了不同的结果,说明两种方法在不同侧面上给出了关于层次体系质量的描述.

表 1 关系识别前后最小生成树的 Solskinnsbakk 理论 1 指标

语义实例	关系识别	第 2 层		第 3 层	
		父子	兄弟	父子	兄弟
TR	前	0.0034	0.0058	0	0.0466
	后	0.5413	0.1216	0.8165	0.2881
TC	前	0.0014	0.1178	0.0018	0.1677
	后	0.6502	0.3202	0.9011	0.4857

针对 Solskinnsbakk 等的理论 3,实验表明有 12 组结果表现出了更好的质量,而 10 组结果则在应用关系识别后出现了质量的下降,另有 8 组结果在应用关系识别前后其值均为 0.上述结果表明关系识别可以显著的改善 WA-F1 指标,且在 Solskinnsbakk 等的评估指标上也表现出了一定的改善效果.

5.3 标签层次体系构建效果分析

实验接下来评估本文提出的基于标签关系识别以及语义流动分析的标签层次体系构建方法,并对比了刘等[7]提出的方法如图 2 所示.

实验结果表明,结合标签关系识别以及语义流动分析方法后得到的标签层次体系结构取得了优于刘等提出的方法的质量.以资源语义实例为例,其平均质量提升接近一倍.

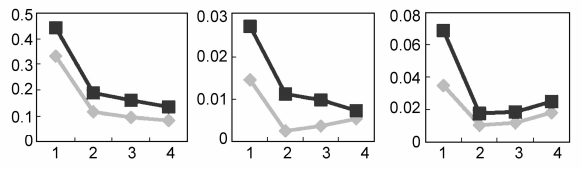

图1 TT语义表示下应用关系识别前后WA-F1指标变化。菱形和方形分别对应应用前后，X轴为子树深度，Y轴为WA-F1，从左起分别为语义流、最小生成树及关系组合方法

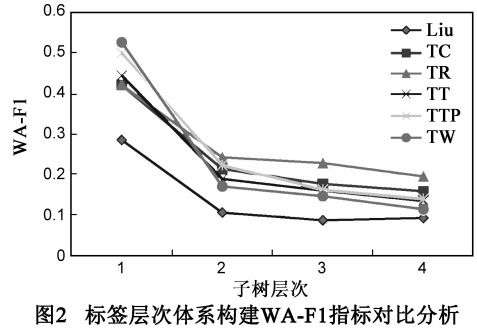

图2 标签层次体系构建WA-F1指标对比分析

在刘等提出的方法上应用 Solskinnsbakk 等提出的评估指标的结果如表 2 所示.由于采用了不同的层次体系构建方法,这里仅考虑理论 1 的结果.实验结果表明应用刘等提出的方法所得的层次体系中仅第 1 层符合

* http://faculty.neu.edu.cn/ise/zhangyin

的理论 1,而应用语义流动分析方法得到的结果均复合这一理论.

表 2　标签层次体系构建应用 Solskinnsbakk 假设评估的结果

第 1 层		第 2 层		第 3 层	
父子	兄弟	父子	兄弟	父子	兄弟
0.0704	0.0009	0.0603	0.3216	0.0157	0.0993

6　结论

本文研究了用于 Web 2.0 标签层次体系构建的标签间关系识别及基于语义流的层次组合方法.本文认为导致错误层次构建的一个重要原因是将具有同层关系的标签误认为具有上下位关联,并提出了标签间关系识别以避免该问题.另一方面,通过利用标签所形成的标签组中蕴含的关于如何形成层次结构的信息,本文提出了基于语义流的层次组合方法.实验使用了一组评估指标将本文提出的关系识别方法在不同的设定下进行了评估,证明了这一方法可以有效的帮助改善应用不同方法得到的标签层次体系的质量.实验还表明,应用本文提出的基于关系识别与语义流动分析的标签层次体系构建方法可以获得质量显著高于对比方法的层次体系.

参考文献

[1] Y Song, B Qiu, U Farooq. Hierarchical tag visualization and application for tag recommendations[A]. Proceedings of the 20th ACM Conference on Information and Knowledge Management[C]. New York: ACM, 2011. 1331 – 1340.

[2] K S Candan, L D Caro, M L Sapino. Creating tag hierarchies for effective navigation in social media[A]. Proceeding of the 2008 ACM Workshop on Search in Social Media[C]. New York: ACM, 2008. 75 – 82.

[3] M Peter. Ontologies are us: A unified model of social networks and semantics[J]. Journal of Web Semantics, 2007, 5(1): 5 – 15.

[4] D Helic, M Strohmaier. Building directories for social tagging systems[A]. Proceedings of the 20th ACM Conference on Information and Knowledge Management[C]. New York: ACM, 2011. 525 – 534.

[5] T Eda, M Yoshikawa, et al. The effectiveness of latent semantic analysis for building up a bottom-up taxonomy from folksonomy tags[J]. World Wide Web, 2009, 12(4): 421 – 440.

[6] B Zhang, Y Zhang, K N Gao. Modeling consensus semantics in social tagging systems[J]. Journal of Computer Science and Technology, 2011, 26(5): 806 – 815.

[7] 刘凯鹏,方滨兴.基于社会性标注的本体学习方法[J].计算机学报,2010,33(10):1823 – 1834.
Liu Kaipeng, Fang Binxing. Ontology induction based on social annotations[J]. Chinese Journal of Computers, 2010, 33(10): 1823 – 1834. (in Chinese)

[8] C Trattner, C Korner, D Helic. Enhancing the navigability of social tagging systems with tag taxonomies[A]. Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies[C]. New York: ACM, 2011. 18.

[9] E Tsui, W M Wang, C F Cheung, A S M Lau. A concept relationship acquisition and inference approach for hierarchical taxonomy construction from tags[J]. Information Processing and Management, 2010, 46: 44 – 67.

[10] G Solskinnsbakk, J Gulla. A hybrid approach to constructing tag hierarchies[A]. Proceedings of the 2010 Conference of On the Move to Meaningful Internet Systems[C]. Berlin: Springer, 2010. 975 – 982.

[11] P D Meo, G Quattrone, D Ursino. Exploitation of semantic relationships and hierarchical data structures to support a user in his annotation and browsing activities in folksonomies[J]. Information Systems, 2009, 34(6): 511 – 535.

[12] L García-Castro, M Hepp, A García. TagSorting: a tagging environment for collaboratively building ontologies[A]. Proceedings of the 17th International Conference on Knowledge Engineering and Management by the Masses[C]. Berlin: Springer, 2010. 462 – 472.

[13] G Solskinnsbakk, J A Gulla, V Haderlein, P Myrseth, O Cerrato. Quality of hierarchies in ontologies and folksonomies[J]. Data & Knowledge Engineering, 2012, 74: 13 – 25.

作者简介

高克宁　女.1963 年 3 月出生,辽宁沈阳人.2006 年于东北大学获得博士学位.现为东北大学教授,主要研究方向为 Web 信息处理.
E-mail: gkn@cc.neu.edu.cn

张　引　男.1985 年 5 月出生,辽宁沈阳人.2006 年于东北大学获得学士学位.现于东北大学攻读博士学位,从事 Web 信息处理技术方面的有关研究.

张　斌(通讯作者)　男.1964 年 2 月出生,辽宁本溪人.1997 年于东北大学获得博士学位.现为东北大学教授、博导,主要研究方向为服务计算、信息检索以及数据挖掘.
E-mail: zhangbin@ise.neu.edu.cn