基于用户分类标签建立结构性的大众分类法

王 爽 徐 行

【摘 要】为了减少 Folksonomy 中标签语义不明显、标签之间层次和关系不清等问题，文章介绍了一种新的标签分类法——UCTag，诠释了 UCTag 以及如何基于此建立结构性的大众分类法。

【关键词】大众分类法 标签 Web2.0

Abstract: In order to overcome the problems of ambiguity in semantics tags, vague levels and relations among the Tags in folksonomy, this paper introduces a new kind of lable classification– UC Tag, defines UC Tag and build a structured folksonomy based on UC Tags.

Key words: Folksonomy Tag Web2.0

1 引言

大众分类法以其自由灵活的特点及共建共享的特色满足了用户对信息的个性化需求，但由于其自身缺乏层次性及表达概念的模糊性使得 Web2.0 网站上的标签没有明确标签之间相互关系的等级分类。因此，建立标签结构是大众分类法面临的最大的挑战和问题[1]。

2 研究背景

大众分类法是由美国信息构建专家 Thomas Vander Wal 和 Gene Smith 在 2004 年 8 月首先提出来的[2]。基本上，Folksonomy 这个名词是与 Taxonomy（学科分类学或专家分类）所对照发展而来，以显示其自由、草根（Root）的特性。Folksonomy 中的 Folks 本意为一般人、大众、老百姓等，而 sonomy 是由 Taxonomy 一词演化而来，表示一种有系统、专门的学科知识。两者合二为一的意思就是：由大众所产生的一种分类知识。较明确的说法是：“一群人自发性定义的平面非层级式标签分类做法。”所谓的标签（Tag）其实就是关键词（Keyword）或索引词汇（Index Term）。Folksonomy 由 Tag 组成。Tag 在 Folksonomy 中，是微资源指代物[3]。也有人称其为社会化分类法、通俗分类法等[4]。

大众分类法是基于用户标签而创造出来的一组标签信息分类方法，用户可根据个人需求自由地选择信息。由于自由标签使用很简便，大众分类法使得大量网络信息分类变得更加容易[5]。目前，几个 Web 2.0 网站（如 del.icio.us and flicker）就是使用大众分类法来方便地获取和共享信息。用户自定义标签形式的大众分类在现下流行的社会性网络服务中得到了广泛的应用，如：Flickr、Furl、Del.icio.us、Frassle 、Simpy、Spurl 、Technorati、FoToFlix 、OpenBM 等[6]。

自 2005 年起，国外对 Folksonomy 理论方面的研究开始逐渐展开。对于 Folksonomy 在有效性、适用性、功能性、结构性等方面的研究均取得了显著的成果。在关于 Folksonomy 存在价值的研究、基于 tag 的定量分析、基于用户的定量分析、系统的设计及应用研究、Folksonomy 的缺陷解决措施研究、Folksonomy 的检索问题[7] 等方面的研究对于我们了解国外 Folksonomy 的最新研究状况、构建各种模型和研究方法将有积极的促进意义。

相比国外 Folksonomy 研究领域的丰硕成果，国内的研究状况相对比较滞后，大多数文献还主要集中在对 Folksonomy 的介绍和阐述。目前大众分类法主要应用在博客、分享类网站、网摘等狭窄的领域中。美味书签是最早提供共享分类体系的网站，这种基于用户提交关键字建立的一种自由分类体系，在目前的应用比较广泛，有 QQ 书签、google 书签等。目前知道和了解 tag 的网民还仅限于网上博客，通过搜索使用了相同关键字的日志，增进了对资源具有相同感知度和认知度的网民的互相了解和沟通。另外，几乎所有的图片、视频分享类网站都提供 tag 服务，用户可以通过赋予图片、视频关键字，实现与全球网民的共享。

如今 Web 2.0 网站上的标签越来越受欢迎和关注，在使用他们搜寻相关网络信息的同时，也有一定的局限性和缺陷。(1) 正如 Mathes (2004) 所说：标签的最大问题是它们的模糊性和同义词现象[8]。例如 OWL 可以指代猫头鹰 (owl)，也可指网络本体语言 (Web Ontology Language)。其他模糊性现象还出现在像 BC 这样的首字母缩略词中，BC 可指 Before Christ (公元前)，也可指 Boston College (波士顿大学)。(2) 像 automobile 和 car 这样的同义词因为有同样的意思，更容易使人误解。(3) 除了语义上的问题，Golder 和 Huberman (2006) 又指出了标签在认知方面存在的问题。例如，一个人可能会把 JavaScript 定义为 Java，而另一个人则会把它限定为一种编程语言[9]。(4) 由于大众分类法是平面的、非等级的，这是与传统分类法的一个显著不同，因此，系统在整理信息时是在用一个用词构成的平面结构，难以表达复杂的关系。(5) 由于文化背景、语言、思想的不同，在跨国界、跨文化的信息交流中仍然存在很多困难。作为一种由大众产生的分类知识，Folksonomy 仍然存在很多问题。

3 涉及本项研究的标签分类方法

国外关于大众分类法中标签含义的界定已进行了大量的研究工作，并取得了一定的成果。

3.1 DogmaBank——社会性书签系统提供的标签分类方法

该系统可以对那些在使用书签的过程中收集到的信息进行共享、组织和再利用。当用户输入一个关键词并使用词网本体 (WordNet Ontology) 概念法对网页进行分类时，DogmaBank 就会列出与该键入词相关的一系列潜在概念，且附带有各个概念的解释性描述。用户可以选择一个特定的概念并用其作为该网页的标签，如果用户找不到相关的概念，可以按照个人的定义添加新的标签。

3.2 Semkey——语义协同标签系统提供的标签分类方法

该系统为资源描述提供了一种新的方法，该方法称为语义标签。这种系统可使用户从 WordNet 和 Wikipedia 这两大本体中选择他们的标签定义。这两大本体中有足够的语义可供标签选择，WordNet 中几乎涵盖了所有的词汇，包括许多词与词之间的语义关系；Wikipedia 是随着新概念的引进而不断丰富的，其内容由来自世界各地的编者们不断更新。

3.3 帕桑特 (Passant) ——公司网志系统提供的标签分类方法

该系统通过领域本体技术使每个标签与其本体概念相连，消除了标签的模糊性[10]。使用领域本体技术的原因在于领域本体在一个公开环境中不易定义，比如在网页上；而在一个界定的环境中则很容易定义，比如一个公司内部。当用户键入与某一特定领域术语相关的关键词时，通过检索领域本体就很容易找到合适的概念，用户只需使用该系统对与网志相关的概念进行核实就可以了。

3.4 SweetWiki—Wiki 文献进行分类的系统[11] 提供的标签分类方法

该系统为大众分类法的结构化提出了一种混合型的信息分类方法。Folksonomy (大众分类法) 上的标签是由用户键入关键词对 Wiki 文献进行分类时自行添加的，Folksonomy 的结构是由社区专家们定义的，这些专家可能会对用户标签对应概念进行重新安排。在用户和领域专家们的合作之下，Folksonomy 上的标签内容更加丰富，各标签之间也更加有条理性。

3.5 用户分类标签结构性的大众分类法

以上大众分类法的标签分类方法各自也存在其问题。第一，与键入一个标签相比，将一个标签本体化会浪费用户大量的时间。第二，由于各本体的结构是由领域专家们定义的，所以用户的共识可能不会在本体结构中体现出来。第三，本体结构中可能不会囊括与用户标签相关的所有概念，因此用户会要求系统管理员将新的标签添加到其本体结构中。换句话说，用户所提出的新概念不会在本体结构中立即体现出来。

随着一种叫做用户分类标签 (User-Categorized 简称 UCTag) 的新型标签法的提出，它减少了以上这些问题的出现。此方法的目的是处理不同的自由标签分类相关的问题，我们所提出的方法具有双重优势。首先，用户可以很容易地输入标签及其类目，不用考虑其本体意义或任何预先定义的结构 (如果有的话)。其次，由用户输入自动生成的结构性 Folksonomy 中的等级关系可以用来检索用户感兴趣的文档。结构性 Folksonomy 通常反映用户对每个标签及其相应类目之间关系的一致性。

UCTag 包含一个标签以及与之相应的标签分类。当用户使用 UCTag 时，他们不仅可以自由定义该标签，同时可以按照个人需求对其进行分类自定义。基于 UCTag，一种涉及标签之间等级分类关系的结构性大众分类法便自动生成了。接下来将会介绍如何使用 UCTag 标签方法以及如何自动建立一个结构化的 Folksonomy。

4 建立结构化的 Folksonomy

采用 UCTag 方法进行标签的整个过程如图 1 所示，共包括 4 个步骤:

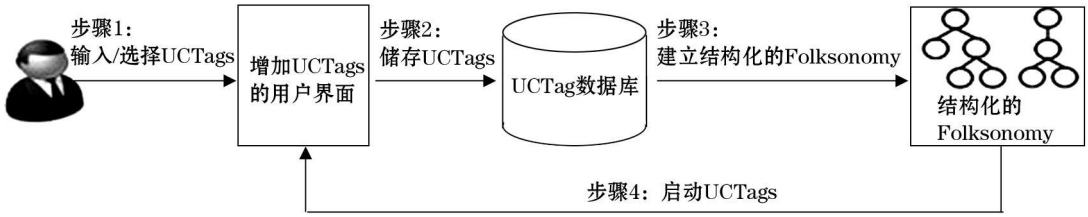

图 1 建立结构化的 Folksonomy 的全过程

步骤 1: 用户输入一个新的 UCTag 或选择特定的 UCTag 对资源进行分类。用户可随意键入 UCTags，每一个 UCTag 都包括一个标签（tag）和一个相应的类目（category）。如果没有预定义的分层概念结构，用户可直接对标签类目定义。UCTag 可表示如下: 标签 < 类目（Tag < Category）。标签具有相应的类目，其含义就会变得更加清楚。因为标签类目为标签提供了背景，特别是使用同音异义词、多义词以及首字母缩略词作为标签时。这种 UCTags 的例子有: 同音异义词（比如 OWL < Bird，OWL < Ontology）、多义词（比如 Bank < Financial Institute，Bank < Building）、首字母缩略词（比如 BC < Before Christ，BC < Boston College）、数字（比如 2009 < year，182 < cm）。

步骤 2: 按照相应的储存规则将 UCTags 存储到 UCTag 数据库中。当用户输入一个新的 UCTag 或者选择一个特定的 UCTag 对资源进行分类时，按照储存规则，该 UCTage 就会被存储到 UCTag 数据库中。具体过程为:

(1) 当记录一个新的 UCTag 时，该 UCTag 就会有一个新的 ID。

(2) 将 UCTag 按 tag 和类目 Category 进行分类后，该 tag 就会被储存在 UCTag 数据库中的 Level _ 1 _ Tag，而 Category 则会被储存在 Level _ 2 _ Tag。

(3) 如果输入一个预先储存过的 UCTag，计数（Count）就会加上 1。

(4) 当用户输入多个 UCTags 时，储存规则也会对各个 UCTags 之间的 IS- A 关系进行检查。

基于储存规则，UCTags 被实时地储存到 UCTag 数据库中，数据库中的 Count 显示的是所记录 UCTags 的总数。也就是说，计数越高表明用户对标签层次的统一认识越高。

步骤 3: 根据创建规则，使用 UCTags 自动建立一个结构性的 Folksonomy。Folksonomy 中的每个标签都具有等级关系。该结构化的 Folksonomy 代表各标签之间的分层关系。创建规则如下所示:

对一个两层的 UCTag，每个标签创建一个节点，用箭头将其连接起来，由高层标签指向低层标签，且按计数给箭头做上标记。

对一个三层的 UCTag，也为每个标签创建一个节点，用箭头将其连接起来，由高层标签指向中层标签，再由中层标签指向低层标签。且按计数给每个箭头做上标识。此外，将代表高层标签的节点同代表低层标签的节点用一个虚线箭头连接起来。

如果一个新的两层标签所对应的节点出现在图表中，只需通过增添标识上新的标签计数对相应的标识进行调整。对于重复出现的三层标签也做同样处理。

如果生成图中存在标签循环问题，将带有最小号码的箭头去除，作为构成循环的箭头中的标识。图表中循环的存在表明由各标签所指代的某些概念关系是错误的，需要将其去除。

UCTag 数据库

图 2 举例说明了从 UCTag 数据库创建一个结构性 Folksonomy 的途径。图中阶段 10 代表在 UCTag 数据库中输入 10 后的处理结果。需要注意的是，对 UCTag 数据库中的第四次输入，由于第一层关系 OWL< Semantic _ Web 已经在图表中展现了出来，只需将其计数加到相应的标识上即可。我们对第二层关系也做了同样的处理。还需要注意的是，所生成的结构性 Folksonomy 图表有两个循环，第一个是在 Web、Semantic-Web 和 RDF 节点中，第二个是在 Bird 和 Eagle 节点中。因此，我们将作为标识的带有最小号码的箭头去除，也就是去除第一循环中的由 RDF 指向 Web 的箭头以及第二循环中由 Eagle 指向 Bird 的箭头。

步骤 4: 帮助用户定义其 UCTags，当用户键入 UCTag 时，结构性 Folksonomy 上的一系列潜在的 UCTags 就会启动。用户可以自由键入 UCTags 的同时，也可以在系统自动启动的潜在 UCTags 中选择一个标签。他们是按

照标签关系的递减顺序排列的。如果用户选择一个特定的 UCTag，那么该结构性的 Folksonomy 就会进行实时更新。

ID	Level_1_Tag	Level_2_Tag	Level_3_Tag	Count
1	Semantic_Web		5	
2	RDF	Semantic_Web		10
3	OWL	Semantic_Web		15
4	OWL	Semantic_Web	Web	3
5	Eagle	Bird		7
6	Bubo	OWL	Bird	3
7	OWL_DL	OWL	Semantic_Web	5
8	Bird	Eagle		1
9	RDF	Semantic_Web	Web	4
10	Web	RDF		2

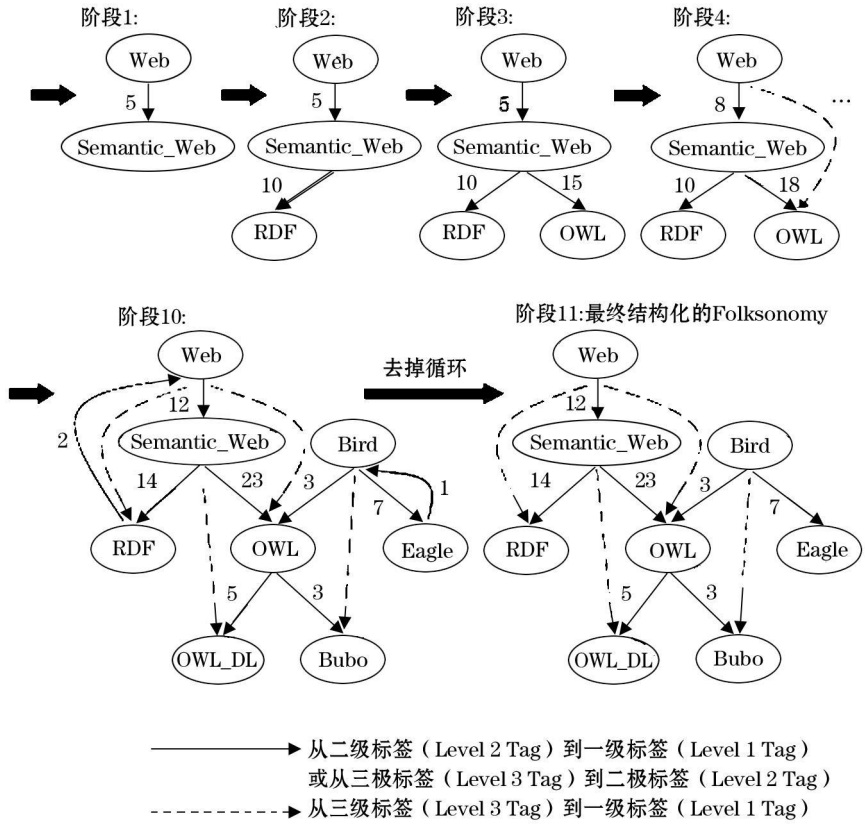

图 2 从 UCTag 数据库创建一个结构性 Folksonomy 的例子

注释

[1] M. Buffa, F. Gandon, G. Ereteo, P. Sander, P. Sander, and C. Faron. SweetWiki: A semantic wiki. Journal of Web Semantic, 2008 (1): 84–97

[2] 刘洋. 大众分类法的应用现状及前景分析. 现代经济信息, 2010 (5): 205–206

[3] 陈丽冰. 大众分类法及其在图书馆中的应用. 中共四川省委省级机关党校学报, 2010 (2): 93–96

[4] http://en.wikipedia.org/wiki/Folksonomy

[5] A. Mathes, "Folksonomies: Cooperative Classification and Communication through Shared Metadata," December 2004. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

[6] 余金香. Folksonomy 及其国外研究现状. 图书情报工作, 2007 (7): 38–40

[7] LambiotteR, AusloosM. Collaborative tagging as tripartite network. http://arxiv.org/abs/cs.DS/0512090, 2006–08–15

[8] A. Mathes, "Folksonomies: Cooperative Classification and Communication through Shared Metadata," December 2004. http://www.adammathes.com/academic/computer-Mediated-communication/folksonomies.htm

[9] S. Golder and B. A. Huberman. Usage Patterns of collaborative Tagging Systems. Journal of Information Science, 2006 (2): 198–208

[10] A. Passant, "Using ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs," International Conference on Weblogs and Social Media (ICWSM 2007), March, 2007, Colorado, USA

[11] M. Buffa, F. Gandon, G. Ereteo, P. Sander, and C. Faron. SweetWiki: A semantic wiki. Journal of Web Semantics, 2008 (1): 84–97

王 爽 徐 行 吉林大学管理学院研究生。