·工作研究·

Drupal分类组织机制研究: 一种复合信息组织模式

范　炜

(中国科学院研究生院　中国科学院国家科学图书馆　北京 100049)

文　摘　明确自由标签法与受控词表相结合的发展思路, 剖析了开源内容管理系统 Drupal中的分类组织机制, 提出一种基于 Drupal的复合信息组织模式, 具体思路可供信息资源库建设参考。

关键词　Drupal　Taxonomy　标签法　信息组织　内容管理

A Study on Drupal's Taxonomy Module: A Hybrid Pattern of Information Organization

Fan Wei (Graduate School of Chinese Academy of Sciences, Beijing 100049)

Abstract: Based on the combination of free tagging and controlled vocabulary, the paper analyzes Drupal's taxonomy module, provides a hybrid pattern of information organization for information resource development.

Key words: Drupal; Taxonomy; Tagging; Information organization; Content management

1　明确自然语言与人工语言结合的发展思路

对于信息检索系统中使用自然语言还是人工语言 (情报检索语言)的问题一直以来都是研究者热衷于讨论的问题。自然检索语言与情报检索语言的关系问题, 大多数研究者认为, 自然语言与人工语言各有优缺点, 主张两者并行发展和结合使用[1]。

标签法 (Tagging)是近几年随着 Web2.0发展浪潮而进入大众视野的, 作为一种简单易用的新型信息组织模式被广泛应用于网络资源组织、分享与交流过程之中。一个资源实体根据认识的角度不同包含着不同的分面特征, 如果按照传统单一的线性分类必然会丢失其他分面特征。标签法是一种非等级式、非排他性的组织模式, 它的优势是赋予用户使用自然语言自由标引资源的权利, 允许用户使用多个标签对资源从不同的角度去描述。描述资源的标签处在同一个平面内, 可以看作是资源各个分面的特征描述, 资源通过任一个描述它的标签被查找到。

资源描述的标签词汇越多则为用户的检索提供更多入口, 但同时处在一个平面内的大量标签也使得查询越来越困难, 资源组织体系越来越混乱。另外标签集合中还存在无组织意义的标签和垃圾标签 (恶意标注), 由于这些标签意义所指不明确或错误定义, 因而严重干扰了资源组织的效果。John C. Dvorak对标签的使用提出质疑[2], 他指出标签未来的应用需要加入过滤机制对垃圾标签进行控制, 否则标签的可利用率就大大降低了。标签应用需要采用后控手段的辅助和优化才能充分发挥其作为资源组织工具的效用。

在标签法 (Tagging)和 Folksonomy的应用研究中, 曾就标签法与传统信息组织的关系问题进行过激烈的讨论。客观地讲, 标签法与传统信息组织方法之间应该是相互促进、共同发展的, 并不存在取代关系。Folksonomy的主题概念和语义关系的明晰需要受控词表的后控机制来优化, 反过来, Folksonomy提供的新事物主题和语义关系对传统信息组织工具的扩充和修订起到了积极的促进作用。在具体的信息资源系统中, 一方面允许用户参与资源建设, 使用标签进行描述; 另一方面系统内部采用受控词表来优化 Folksonomy体系, 受控词表与 Folksonomy之间形成互动互补机制。这种设计思想体现了将自然语

言与人工语言结合的一种发展思路,同时也强调了以用户为中心,在资源描述和组织过程中融入用户对内容的理解和认知反馈。本文利用Drupal的分类功能(Taxonomy Module)及相关内容组织模块来探索如何在信息资源建设中实现标签法与受控词表结合的技术方案。

2 Drupal内容管理概述

选择一款内容管理解决方案时,对于内容的描述、分类和组织机制是首先要重点考虑的。如果没有底层坚实的内容组织架构,上层的管理、共享、交流和利用等功能就无法有效开展。Drupal是一款采用PHP技术架构的开源内容管理系统[3](以下简称CMS),与其他CMS如Joomla、Xoops、Plone、Mambo等相比,Drupal引入的后控词表管理机制是其特色之处,也是Drupal开发进化中被重点关注的一个核心要件。以下介绍Drupal中涉及内容描述、分类与组织部分的相关概念和模块[4]。

• 节点(Node):一个内容对象,每一个节点拥有一个唯一的数字ID。可以赋予节点标题,通过标题查看节点中的内容。这里的节点定义有些抽象,我们可以简单地将节点看作是放置内容的容器,里面可以放一本书、一个博客、一篇文章等等。一个节点里放的具体内容以及内容数量没有预先规定,由内容管理者自行定义。节点是Drupal的核心模块,在Drupal中绝大多数内容都是以节点形式存储的。

• 内容类型(Content Type):定义和描述具体内容,不同的内容拥有不同的类型定义。在一般情况下,内容类型与节点类型是同义词,可交替使用。Drupal内置有Page、Story、Poll、Blog、Book等少数几种内容类型。在具体应用时,这几种简单的内容类型定义远远不能满足我们的需求,这时可以使用CCK[5](Content Construct Kit)模块及配套的相关模块(如针对文件、图片等类型的模块)来自定义内容类型。使用CCK进行内容类型定制的好处在于,内容管理者将注意力集中在如何更好地定义和描述内容主题,而不必过多关心内容的底层技术存储细节。这里Drupal提供的内容类型功能为我们提供了元数据描述的施展平台,对于特定领域已有的元数据描述方案应该在设置内容类型时充分参考。

• 分类(Taxonomy):内容分类和组织的工具。这里的Taxonomy相当于受控词表(Controlled Vocabulary),使用系统后控的手段定义一套分类词表用来对内容管理系统中的内容进行分类和组织。Drupal的Taxonomy功能比较丰富,可以满足当前多样的组织需求,如为不同内容类型定义不同的词汇表(Vocabulary),使用预设主题词描述和用户自定义标签两种方法,设置主题词权重等。Taxonomy是Drupal的基础核心模块,还存在许多对其功能进行扩展与丰富的其他模块,如Tagadelic[6](实现标签云图)、Taxonomy Manager[7](增强Taxonomy的管理和维护功能)等。

节点、内容类型、分类三者组成了Drupal内容组织功能的核心部分。根据Drupal的内容组织流程,我们可以这样形象地理解内容、内容类型、Taxonomy与节点之间的关联:内容通过内容类型描述,Taxonomy归类,最后放入节点这个容器中。节点和内容类型是组织的基础,Taxonomy是组织的核心。接下来,对Drupal的Taxonomy分类组织机制进行分析。

3 Drupal分类组织机制

从信息组织的角度看,Drupal的Taxonomy模块是一种组织技术实现手段,它可以实现以下几种分类组织结构。

3.1 等价关系

将一组同义术语连接在一起,提高搜索引擎对同一主题内容的查全率,降低漏检率。具体设置方法是在术语页面中高级选项的Synonyms部分输入对应的同义术语。需要注意的是,在Synonyms里设置的同义术语与Parents部分设置的上位类原则上不能相同,否则就破坏了分类体系的一致性。

3.2 等级体系

正如Taxonomy一词的含义一样,体现单一线性的等级划分体系结构,即上下位类关系,通过术语编页面中高级选项的Parents部分来设置,点选术语作为其上位词。这个过程为术语设置上位类的同时也定位了其下位类的位置,即一个单向设置完成层级的双向关系。

3.3 复合等级体系

相对于单一线性的等级划分体系而言,复合式等级体系体现了多重列类原则,即允许一个类目从属于两个或两个以上的类别,从而保证交叉综合性主题在多个类目下可以被查找到。多重列类的方法在网络分类目录中多有采用,一般为其设置一个人为的主目录,在其他目录下采用@链接指向到主目录。在Drupal中可以为一个类目指定多个上位类,这些上位类是平等的,不存在主次之分。具体设置方法是在术语页面高级选项的Parents部分一次选中多个术语即可。需要注意的是,复合等级体系虽

然可以在一定程度上满足多重途径查找的需要, 但并不是交叉目录设置越多越好, 要根据实际需要针对性处理。交叉目录设置过度会引起整体组织体系的混乱。

3.4　相关关系

Drupal的 Taxonomy提供了术语相关关系的设置, 类似于叙词表结构中的参见 (See Also)关系, 具体设置方法是在术语页面高级选项的 Related Tems部分选择相关术语, 这里可以进行一次选中多个术语的操作。

3.5　Folksonomy/Tagging

Drupal在保证系统内部 Taxonomy对内容进行分类组织的基础上, 允许用户对内容使用主题关键词进行描述 (词与词之间用英文半角的逗号分开), 即引入大众标签法。对已有的 Taxonomy词表启用标签方法, 在词表属性页面进行设置。用户贡献的标签将直接成为 Taxonomy词表中的术语, 这些由用户贡献的术语可以看作是原生态的未被加工的词表资源, 它们是处于一个平面之上的标签集合。Drupal在 Taxonomy基础上集成了标签法应用, 这为自然语言与受控语言结合思路提供了一种技术实现手段。

4　复合信息组织模式的技术实现

基于以上对 Drupal分类组织机制的分析, 在信息资源库建设中可以利用这种技术手段实现资源描述与组织过程。内容管理框架如下图所示, 具体流程如下:

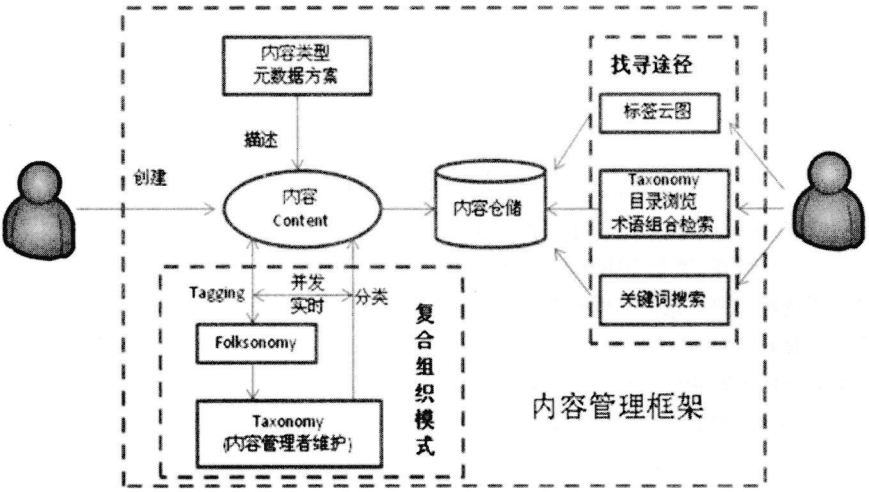

图　内容管理框架示意图

4.1　资源规划阶段

内容管理者首先对信息资源主题的各种内容类型进行定义, 确定内容类型的元数据描述方案, 参考已有的综合或专业词表工具, 初步设计出受控词表结构并与内容类型进行关联。这个阶段形成的具体成果包括内容类型、元数据描述方案、Taxonomy分类词表。

4.2　资源建设阶段

在内容创建过程中, 用户选择特定的内容类型, 按照对应的元数据描述进行内容描述, 然后要求用户为每一条内容记录进行归类。这里提供两种途径, 一种是从该内容类型已有的 Taxonomy分类词表中选择术语, 另一种是允许用户自主定义标签。这两种方法通过 AJAX异步实时响应的交互方法有机地集成起来, 用户在输入语词的同时, Drupal会基于已有的 Taxonomy词汇给出推荐, 类似于 Delicious的标签语词推荐功能。用户自定义的标签词汇进入该内容类型所属的 Taxonomy词表中。这里的用户包括内容管理者和普通用户两类, 他们都可以直接向资源系统贡献内容。

4.3　Taxonomy管理

内容是不断变化的, 对于内容分类组织的工具也应该能够及时响应内容的变化并做出调整, Taxonomy不是一成不变的, 需要在内容建设过程中进行有效维护和管理。我们在资源规划阶段初步设计出一个 Taxonomy词表结构, 在资源建设过程中要依据新增内容与用户贡献的标签词汇对这个词表进行修订。由于用户贡献的标签术语是随新增内容直接进入对应 Taxonomy词表中去, 并未做任何处理。内容管理者需要定期对这些标签集合根据所描述的内

容进行归类和关系描述,使得有用的标签词汇融入到 Taxonomy整体结构中去,同时剔除一些恶意、错误的标注,并对原有词表中的词汇及关系进行调整。这种做法既保证了用户的参与又维持了 Taxonomy结构的一致性,在一定程度上能够保证内容的良好组织。这种方式实现了在一个信息资源系统中自然语言与人工语言的有机结合,形成了内容建设和组织工具两者之间相互促进的良性循环,使得管理内容的 Taxonomy总是处在变化之中,而新的内容又有地方可安置。

4.4 搜索与浏览

良好的内容组织架构是实现有效找寻的基石。通过系统前端的用户标签描述和系统后端的后控词表控制两种方式的结合,为用户查找内容提供了多种途径:1)标签云图,通过依附于 Taxonomy的 Tagadelic模块可以实现用户标签集合的云图显示,用户通过点击感兴趣的标签多角度选择内容。2)Taxonomy术语组合检索,Drupal提供了一种与大众标签系统类似的组合检索功能,用户可以将不同的术语词汇以布尔逻辑方式在 URL里构造组合检索式。3)分类目录,通过将 Taxonomy组织结构在界面层次上的外显化形成一个导航式(类似 DMOZ)的分类目录浏览。导引式浏览对用户信息需求模糊情况下进行探索性找寻特别有效。4)关键词搜索,Drupal的搜索模块提供了关键词搜索功能。在高级搜索功能中提供了从 Taxonomy词表结构中选择特定目录进行限定搜索。

5 小结

从以上的分析和讨论中可以看出,Drupal作为一项开源的内容管理技术平台实际上提供了一套完整的资源描述、组织和管理解决方案,其中 Taxonomy功能与自由标注相结合的分类组织机制是其整体技术架构中的一个亮点所在。本文的研究仅是针对 Drupal分类组织机制展开分析和讨论,并未涉及到 Drupal的区域设置、视图结构、主题界面、人员管理、系统配置等其他方面。这里提出的基于 Drupal的复合信息组织模式可供信息资源库建设参考。

最后,需要指出的是,技术驱动的信息组织模式创新与变革是当前需要正视的一个问题。在数字化网络环境下,关注技术所能解决的信息组织问题是关键所在,而非单纯研究技术本身。信息组织方法的研究与应用不可避免地要结合技术手段来实现,发掘技术发展所带来的新型组织模式并纳入到信息组织体系和为传统信息组织方法寻找技术支撑实则是同一问题的两个方面。

参考文献

1 张琪玉. 情报语言学基础. 武汉:武汉大学出版社,1997:304
2 John C. Dvorak To Tag or Not to Tag That Is the Question. http://www.pcmag.com/article2/0,2817,1819101,00.asp [2009-3-25]
3 http://www.drupal.org/ [2009-3-25]
4 Drupal Handbook http://drupal.org/handbooks [2009-3-25]
5 http://drupal.org/project/CCK [2009-3-25]
6 http://drupal.org/project/tagadelic [2009-3-25]
7 http://drupal.org/project/taxonomy_manager [2009-3-25]

范 炜 博士研究生。

(收稿日期:2009-03-27 编发:许桂菊)

(上接第30页)

参考文献

1 张继东,予以胜. 利用叙词表构建本体的方法研究. 图书情报知识,2006(6):82-85
2 贾君枝. 简单知识组织系统与汉语主题词表. 中国图书馆学报,2008(1):75-78
3 Maria Lee, Stewart Baillie. TML thesaurus markup language http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.2101 2009.01.01
4 Thesaurus RDF-The RDF Thesaurus descriptor standard http://ceres.ca.gov/thesaurus/RDF.html 2009.01.02
5 Dietrich H. Fischer. Converting a Thesaurus to OWL: Notes on the Paper "The National Cancer Institute's Thesaurus and Ontology". Fraunhofer Institute Integrated Publication and Information Systems, 2004(2)
6 袁梅宇. XML Schema, RDF Schema及 DAML比较. 计算机应用研究,2004(10):67-69
7 杨力. 从 RDF, DAML+OIL到 OWL. 农业图书情报学刊,2005,17(11):108-110
8 刘春艳,陈淑萍,伍玉成. 基于 SKOS的叙词表到本体的转换研究. 现代图书情报技术,2007(5):32-35
9 范 炜. 语义网环境中的叙词表实例研究. 情报科学,2006,24(7):1073-1077

贾君枝 女,教授,山西大学管理学院硕士生导师。
卫荣娟 女,山西大学管理学院在读研究生。

(收稿日期:2009-04-17 编发:许桂菊)