网络资源层次组织规律定量研究*

陈　翀　罗鹏程　刘晓兵　陆　炀

摘　要　本文以互联网 FTP 目录树为对象,从大众构造的层次式结构中揭示管理网络资源时的组织规律。通过考察 44,972 个软件资源及其在目录树中的位置,得到如下认识:①对于用户查找、浏览资源来说,资源存放在深度为 2 或 3 层较适合。②当要管理的资源数量在 200 以内时,人们可容忍平均 1 个目录中包含 10 个左右的软件资源;当资源数量更大时,人们可容忍平均 1 个目录中包含 15 个左右的软件资源。③在使用层次式组织结构时,人们会对结构的深度有一个习惯上限,目录结构平均深度一般不多于 6 层。④人们考虑组织结构时,会在"分类体系复杂度"和"分类精细度"之间权衡,在资源量增多时,尽量保证扁形结构。本文揭示了采用层次方式组织资源时的一些共性,有助于设计更贴近大众习惯的网络资源组织与导航系统,并对标签聚类等从扁平型信息组织方式中提取类目层次的研究工作提供依据。图 7。表 1。参考文献 13。

关键词　网络资源　层次组织结构　目录树　大众分类

分类号　G203

The Structure Patterns of Web Resources' Hierarchical Organization

Chen Chong, Luo Pengcheng, Liu Xiaobing & Lu Yang

ABSTRACT　This paper explores the web resource organization patterns from user-generated hierarchy structures. Based on the analysis of 44,972 software resources collected from 1,481 open-access Web FTP sites, we find that: 1) Users tend to arrange the resources to the depth of 2 or 3 in the hierarchy structures. 2) When the amount of the resources is less than 200, on average, up to 10 software resources are put together in one folder; when the amount of resources is equal to or more than 200, 15 resources per folder. 3) Users usually have a upper limit for the depth of the hierarchy structures when organizing resources. Specifically, in our experiment the depth is 6. 4) Users often face the dilemma to choose between the "complexity" and "fineness" of a taxonomy, and with the increasing amount of resources, users prefer to have a flatter structure. In this paper, we reveal the users' common organization behaviors quantitatively. The findings contribute to the design of friendly navigation structure, and provide the empirical parameters to tag clustering methods. 7 figs. 1 tab. 13 refs.

KEY WORDS　Web resource.　Hierarchical organization structure.　Directory tree.　Folksonomy.

文档、软件、游戏、音视频、图形图像等是互联网上备受欢迎的网络资源。它的媒体类型多样、表现力丰富、信息量大,在互联网用户的学习、工作、休闲娱乐中发挥着重要作用。据调查,2007 年网络用户对网络资源的利用首次超过了以往占首位的搜索引擎,成为吸引用户上网的第一驱动力[1]。在 Web 2.0 时代,用户在网络资源的创建和传播上参与度更高,一方面资源量激增,另一方面用户利用资源愈加困难。

究其原因,网络资源通常包括多种媒体类型的文件,其内涵丰富、主题多样、构成形式灵活,严格而专业的分类法(taxonomy)不易概括其

*　本文系国家自然科学基金项目"异构'非网页'资源的组织与融合方法研究"(批准号:70903008)的研究成果之一,并受"中央高校基本科研业务费专项资金"(2011 北京师范大学自主科研项目"基于引用评价的论文学术要点自动提取方法")资助。

通讯作者:陈翀,Email:chenchong@bnu.edu.cn

内容种类;此外,这类资源也不像网页那样通过链接关系表达内容或结构上的关联性,现有多媒体内容分析技术又不足以对其进行大规模的自动处理。

在缺少组织理论和技术手段的情况下,人们采用目录树和标签来组织互联网资源。资源在互联网上主要有三种共享传播渠道:FTP 服务器、P2P 文件共享系统和网站(如优酷、last. fm、flickr 、新浪爱问共享资料等)。在 FTP、P2P 平台上的资源通过人工构造的目录树形成层次状组织,网站发布平台则主要通过用户标签形成扁平状组织。

近年来,越来越多的研究关注资源标签的挖掘和利用,认为标签能够描述资源的主题、意义、主观感受等,充分体现用户分类与资源描述的智慧。通过层次聚类可以形成反映大众特色的分类体系(folksonomy),以便适应网络资源内容特点,符合用户查找信息的方式[2-3]。其不足在于,标签是扁平的资源组织形式,用词不受控,所聚出的类目名称不规范,而且聚类时还需人为指定类数、层深。鉴于网上的 FTP 和 P2P 站点存在大量以层次体系组织资源的样例,其目录树本质上就体现各个用户群体对资源组织的"大众分类"结构。如果说目录名相当于"标签",那么目录树就类似于标签层次结构。这种层次的合理性来自人的知识,可以想象它要比机器自动聚类的层次更有道理。

扁平模式下,资源可以由其标签集合描述,$r:\{tag_1, tag_2, \cdots, tag_n\}$,$tag_i$ 是用户对资源的标识。层次模式下,资源可表述为其所在目录树从根节点到当前位置的目录名构成的有序 n 元组,$r: < d_1, d_2, \cdots, d_n >$,$d_i$ 是目录名,也就是从根节点逐层对资源的描述。路径 $d_1 - d_2 - \cdots - d_n$ 反映用户对资源组织的分类层次和概括。树中子节点通常是从某个角度对父节点含义的划分或解释,能充分反映用户头脑中的组织体系,而标签之间不存在这种关系。

考察这些目录结构中资源的存取位置,可以发现其中蕴含的大众组织知识。本文研究的主要目的是,找到层次型大众分类在组织资源中的共性规律,尤其是组织结构上的定量特征;这不但能为自动标签聚类方法在参数选择上提供依据,还可帮助设计更加贴近用户习惯的资源导航结构。

1 相关研究

主题词标引在 Web 2.0 中演化为社会标注,以其平坦结构和自由形式降低了资源组织的门槛。网络用户通过关键词标识资源的属性、表达个人看法。但由于标签用词完全不受控,也无约束机制,造成了标注质量不高,反而不利于资源访问。于是,人们想到用机器学习方法对标签按语义自动聚类,形成类目[2-4]或类目体系[5-8]。这一方面减少非受控用词带来的杂乱,同时也兼顾用户对网络资源的组织描述的个性。

为确切了解用户使用标签的情况,研究者对标签用词方式、描述功能展开了研究,并在 del. icio. us 等站点上发现,有一定比例的标签体现了资源与用户活动或任务的关系,如"to read"(待阅)、"irrelevance"(不相干)等在传统分类中很少用的词汇[9-11]。

如果将文件系统中的目录看作资源的容器,其命名体现用户对容器中资源的描述概括,目录树的结构则表达了用户对资源组织体系的认知。我们通过对 FTP 文件目录名的研究,发现用户对目录命名同样具有标签所表现出的个性化和用词不受控等特点[12]。标签不能直接体现用户头脑中的资源组织体系,但文件目录树中却能够反映。因此我们认为:对目录结构的研究能很好地补充人们对大众分类规律的认识。例如知道用户放在同一目录中的平均资源数量有助于聚类算法估计类数,知道资源的平均深度有助于设计符合用户习惯的导航结构。

无论是在 FTP 还是在网站平台上,用户对资源采用的组织描述方式都因其种类不同而存在差异。在对科技文献(connotea)、网页书签(del. icio. us)、照片(flickr)、视频(youtube)四类资源的标签集合分别考察时发现,标签主要描述资源的创建者、类型、日期、来源、制作设备等属性,但因资源种类不同而有所差异,例如科技

文献中几乎不会包含制作设备这类信息,而视频图片类资源则会包括[13]。标签中描述资源内容的相关信息(如语言、内容类别、内容描述)占90%以上,用户还会添加感受、褒贬评价等主观信息。此外还存在的一些与主题无关、无意义的标签,这在 flickr 网站相当常见,由于缺乏为照片逐一加标签的耐心,创建者可能利用网站提供的功能将整理好的照片集打包上传、只加一次描述。因此,资源的组织描述不只与用户的认知和习惯相关,还与所标注的资源种类有关。我们曾对 FTP 平台上的文档、音频、视频、软件类资源进行标定考察,最终选择软件类资源作为研究组织结构的对象。

2 问题定义

2.1 资源及其组织结构

网络资源的组织结构具有多样性,资源自身形态也很复杂。以软件类资源为例,它可以是一个 .exe 的可执行文件,也可以是 .rar 压缩包,还可能如图1中资源 r_2 那样由若干目录和文件组成。为专注于组织结构的研究,我们用人工识别资源,标出它在目录树中的位置及类型(见图1)。在一棵树中,任意资源 r: $< d_1, d_2, \cdots, d_n >$, d_1 是根节点目录名, d_n 是 r 的父节点目录名。树中任意节点的深度 h 为根节点到当前节点的边数,因此,图1中 r_4 可表示为 r_4: $< d_1, d_2, d_3 >$, $h_{r4} = 3$。

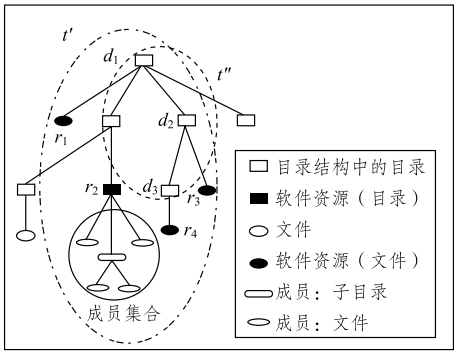

图1 单棵目录树、资源及其路径

为便于描述资源的组织结构,对原始目录树 t 定义了两种剪枝策略,分别得到剪枝树 t'、t'':

1) t':对给定 t,按一个目录及其子孙节点是否包含软件类资源进行剪枝。

2) t'':对给定 t',将资源本身剪去所剩的目录结构。

t'、t'' 分别对应图1中点划线和虚线所围区域。第一次剪枝是为了专注于软件类资源的组织情况;第二次剪枝是为了便于研究用户组织资源时的层次结构特征。

2.2 特征描述

设资源集合 R, $|R|$ 为资源总量,目录树集合为 T, $|T|$ 为目录树总量。$\forall r_i \in R$ 可表示为 $< r_i, t_j, d_{ij} >$,其中 d_{ij} 是资源 r_i 在目录树 t_j 中的路径 n 元组。

对于 $\forall t_j \in T$,其剪枝树 t_j'、t_j''。用 n_{jr} 表示 t_j' 中的资源个数,n_{jd} 表示 t_j'' 中除根节点之外的目录数,n_{jk} 表示 t_j'' 中的叶子数,k_i 为 t_j'' 中任意叶子节点,$i \in [1, n_{jk}]$。

本文主要研究目录树的组织结构规律,如果一棵目录树可理解为一个用户或用户群体对资源的组织行为的反映,通过对目录树的考察,我们希望知道:①用户习惯于在每个目录下放置多少资源;②资源一般会被放置在什么深度。③给定一定数量的资源,用户倾向于建立多大的目录体系;④用户进行层次式的资源组织时,对目录树深度存在何种偏好。对应上述目标,对单个目录树定义如下特征:

目录平均容量 a_j: $a_j = n_{jr} / n_{jd}$

资源平均深度 h_{ar}: $h_{ar} = \sum_{i=1}^{n_{jr}} h_{r_i} / n_{jr}$

结构最大深度 h_{mj}: $h_{mj} = \max_{i=1}^{n_{kj}} h_{ki}$

结构平均深度 h_{aj}: $h_{aj} = \sum_{i=1}^{n_{kj}} h_{k_i} / n_{kj}$

图1所示目录树的各项特征为:$a_j = 4/3 = 1.3$, $h_{ar} = (1+2+3+2)/4 = 2$, $h_{mj} = 2$, $h_{aj} = (1+2)/2 = 1.5$。

通常可以假设:在资源数量不多时,用户不一定要细化分类,而可能直接在原有目录下添加资源,这样会使得 a_j 增加。当资源量上升,用

户会将有共性的资源归到新的目录中，这可能会导致目录数 n_{jd} 增加。我们将 n_{jd} 视为"类目体系复杂度"的一种反映，目录越多，表明这个分类体系越复杂；将 a_j 视为"分类精细度"的体现，a_j 越大，表明分类精细度越低。通常用户希望分类体系的类目体系复杂度低且分类精细度高，但是两者之间是相互制约的。随着资源数量的增加，用户将会面临这两方面的抉择。因此本文也会研究 a_j 与 n_{jd}、n_{jr} 的关系。

此外，我们还将分析 h_{mj}、h_{aj} 与 n_{jd} 的关系，这反映用户组织资源时的结构倾向。

3 实验数据

3.1 基本情况

本文实验所用资源及目录树数据来自中国互联网上（以电信网络为主）公开访问的 FTP 服务器，收集时间为 2010 年 7 月 10 日至 22 日。采集各 FTP 目录树信息，包括目录名、路径、文件名、文件大小，选取目录树信息大于 10K 的站点，以保证所选目录树中确实存有一定量的内容，最终得到样本大小为 1,481 个 FTP 站点集。为了准确识别资源，提高实验数据的可靠性。我们招募 20 名具有信息组织专业知识的志愿者人工标注每棵树中的软件、音频、视频、文档类资源。样本中包含软件、文档、视频、音频的 FTP 站点个数分别为 991、949、478、348。

3.2 资源种类的确定

如果 FTP 使用者为理性用户，会精心管理自己的内容，则其目录分类是有条理有规则的。但实验发现不同 FTP 的资源种类繁多，组织形式个性化明显，有相当一部分是个人学习、工作所用资料，尤其是文档类资源由于其用途差异较大，寻找这类资源的组织共同点意义不大。此外在采样数据集中的音频、视频分布不够广泛，研究其组织结构不具有很好的代表性。而软件类资源的功能、组成都相对标准，用户对其管理中的一致性更强，所以我们选取了软件类资源作为研究对象。后续分析目录树时只考虑软件，并按一个目录及其子孙节点是否包含软件对原始目录树剪枝。经字符转换、去除停用词等方式，剔除一些无意义名（如"17. exe"、"beta 版. rar"）并去掉类别标注错误的资源，得到软件资源 44,972 个。

3.3 资源分布

首先考察软件类资源在互联网 FTP 站点上的存在状况。对于 $\forall t_j \in T$，统计该目录树中包含的软件资源数 n_{jr}，并找到每个唯一的 n_{jr} 值所对应的目录树数量 $f(n_{jr})$。对唯一 n_{jr} 值排序构成横轴，以其对应的 $f(n_{jr})$ 为纵轴，在对数坐标系下得到图 2。考虑到 1,481 个站点中，不含软件的站点数为 490 个，占总数的 33.1%，为此将横坐标取为 $\ln(n_{jr}+1)$。

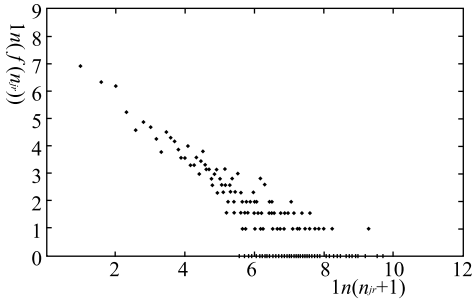

图 2 软件资源在 FTP 中的分布

4 万多个软件资源在 1,481 个 FTP 站点中的分布具有幂律特性。包含软件数大于 100 的目录树为 132 个，但是它们包含的软件资源占软件资源总量的 64.7%。这反映了互联网上某类资源分布的一般状况，即少数站点资源储量大，但更广泛的是储量小的站点。

其次，考察软件类资源不同个体的分布差异。也就是，一种资源在互联网上可能有多个复本。这反映了不同资源的流行程度。由于软件通常有版本的差异，我们将资源名中的版本信息屏蔽，例如"Adobe Reader 9"和"Adobe Reader 7.0"视为同一种资源的两次出现；此外，某软件的"beta 版"、"试用版"、"正式版"也作为同一资源。将 R 中所有资源的名字清理版本信息并归一后，得到唯一资源集合 Q，$Q \subseteq R$，$|Q| = 21,249$。在本文样本中，资源平均复本数为 2.1，最大复本数为 564。

$\forall q \in Q$ 在 R 中对应的复本数就是 q 的频次 $f(q)$。以 $f(q)$ 的降序排序为横轴，记为 $rank(f(q))$，$f(q)$ 为纵轴，在对数坐标系下得到图 3。

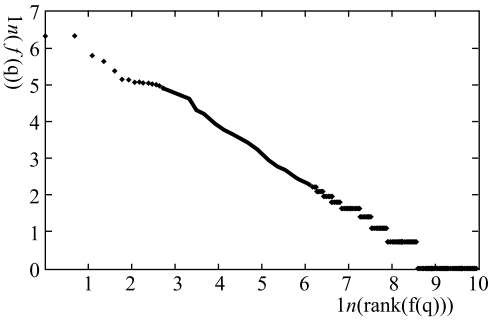

图 3　软件复本分布

出现次数大于 10 的软件数量个数为 394 个，占总数的 1.85%，而它们总的出现频次占所有软件出现总频次的 29.11%，少数软件资源非常流行，出现频次很高，而大多数软件资源出现频次均很低。

4　资源层次组织结构特征

本节分析的对象是软件数量大于 10 的 FTP 站点，这些站点个数为 541，占样本总数的 36.5%。选取这些站点的原因是：当软件数量较多时，人们对其进行组织管理的需求增强，其目录树结构更能反映用户对资源组织的习惯，排除软件数量较少的站点，有助于去掉一些噪音数据。

4.1　目录平均容量与资源总数的关系

由于资源在 T 中的分布是不均匀的，$\forall t_j \in T$，对 t_j 的资源数 n_{jr} 作不等距划分。$n_{jr} \in [11,200]$ 时，以 10 为间距划分得到 19 个区间；$n_{jr} \in [201,300]$ 时，按 50 为间距分为 2 个区间；$n_{jr} \in [301,700]$ 时，以 100 为间距划分 4 个区间；$n_{jr} \geqslant 700$ 为单独一个区间。再对上述每个区间 i，求目录平均容量 a_j 的均值，记为 $\overline{A_i}$。以区间序号为横轴，对应区间的 $\overline{A_i}$ 为纵轴，作散点图，如图 4。图 4 横坐标的前 19 个刻度对应的是一棵 FTP 目录树中资源数量 $n_{jr} \in [11,200]$ 的情况。软件数量为 $[201,+\infty)$ 目录平均资源数与 $[11,200]$ 相比，其波动的位置有所上移。这说明了随着资源数量的增加，人们有更大的可能性在同一个目录下放置更多的资源。

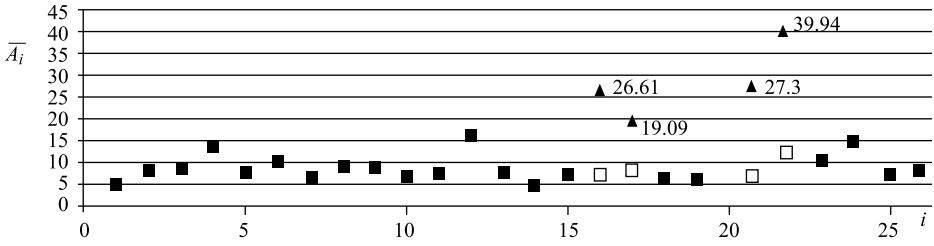

图 4　目录平均容量与资源数关系

在本文采样的数据集中，有 4 个站点，软件数量在 200 左右，但是却只有一两个目录，这种情况显示该 FTP 站点的用户并非为了管理资源，而是出于备份或中转等其他目的。图 4 中黑色三角形表示包含这 4 个站点时的计算值，空心方形“□”表示去掉这 4 个异常值点的计算结果。从图中可以发现，在软件数量为 $[11,200]$ 范围内，每个小区间的平均值基本在 $[5,10]$ 之间波动，在软件数量为 $[201,+\infty)$ 范围内，每个小区间的平均值在 10 附近波动。

由此可见，当要管理的资源量为 200 以内时，人们可容忍平均 1 个目录中包含 10 个左右的软件资源，当资源量更大时，人们可容忍平均 1 个目录中包含 15 个左右的软件资源。如果一个目录中资源数过多，将其分而治之是人的朴素做法。

需要说明的是，通常每个软件的功能明确，不同软件资源之间的关联性较小。由于资源种

类不同，人们对一个目录下可容忍的资源数量会有所差异。比如一个照片目录中的照片数量可能会远远高于软件。

4.2 目录数量与资源数的关系

图4分析了随着资源数量的增加，人们有更大可能容忍在一个目录下放置更多的资源，也从侧面反映了如下现象：目录数的增加会随资源数的增加而变得缓慢。

仍以每棵树的资源数为对象，按照上小节图4的方式划出区间，对每个区间 i，计算其中目录数 n_{jd} 的均值，记为 $\overline{N_{di}}$，为纵坐标；对应区间 i 的目录树的资源数 n_{jr} 的均值，记为 $\overline{N_{ri}}$，为横坐标，得到图5，图中每个点都对应特定区间中各棵树的资源数和目录数平均情况。

图5中去掉了4个异常值点，可以看出，起初随着资源数量的增加，用于管理这些资源的目录数量也快速增加。但是，随着资源数量不断增加，FTP 目录数量增加的幅度明显变得缓慢。

表1第一、三列分别是图4的横纵坐标，第四、五列是图5的横纵坐标。

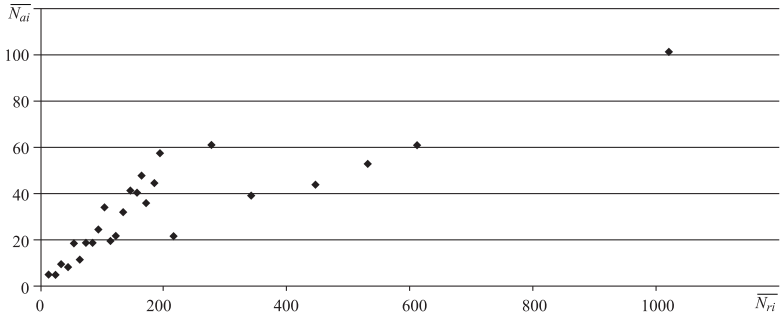

图5 目录总数与资源数关系

表1 资源数区间及各区间的特征均值

区间号 i	n_{jr} 区间	$\overline{A_i}$	$\overline{N_{ri}}$	$\overline{N_{di}}$	区间号 i	n_{jr} 区间	$\overline{A_i}$	$\overline{N_{ri}}$	$\overline{N_{di}}$
1	[11,20]	4.7	14.9	4.7	14	[141,150]	4.5	145.1	40.9
2	[21,30]	8.0	24.8	4.9	15	[151,160]	6.9	154.9	40.3
3	[31,40]	8.2	35.4	9.4	16	[161,170]	6.8	165.8	47.6
4	[41,50]	13.3	45.9	8.1	17	[171,180]	7.9	174.0	35.8
5	[51,60]	7.2	55.3	18.3	18	[181,190]	6.3	186.1	44.6
6	[61,70]	10.1	65.0	11.4	19	[191,200]	5.8	194.9	57.6
7	[71,80]	6.3	74.3	18.7	20	[201,250]	14.0	217.2	21.4
8	[81,90]	8.8	86.2	18.8	21	[251,300]	6.8	278.9	61.0
9	[91,100]	8.7	95.5	24.6	22	[301,400]	12.4	342.2	39.2
10	[101,110]	6.5	105.5	33.8	23	[401,500]	10.4	444.8	43.8
11	[111,120]	7.1	115.1	19.6	24	[501,600]	14.2	532.7	53.0
12	[121,130]	16.3	123.0	21.7	25	[601,700]	6.8	613.7	60.7
13	[131,140]	7.9	134.5	31.8	26	[701, +∞)	7.9	1021.0	101.5

从4.1和4.2小节的分析可以看出，人们在组织资源时，会在分类系统的复杂度和分类精细度之间权衡。随着资源数量的增加，人们有更大的可能性在同一个目录下放置更多的资

源,尽量避免过多目录导致分类系统复杂化。这主要是因为,过于复杂的分类系统不便于维护,用户难以掌握类目体系,从而影响资源的查找和新资源的分类。另一方面,人们也会考虑到目录下资源太多影响分类的精细度,需要增加新的目录以满足需求,从而导致目录较为缓慢的增长。

4.3 资源深度

用户组织资源时习惯于将资源放置在什么样的深度?分别以上述资源数区间和资源平均深度 h_{ar} 为横坐标和纵坐标(见图6)。可以看出无论目录树的具体组织结构如何,平均情况下,用户会将资源放在深度为2到3的目录中。也就是对于用户查找浏览资源来说,资源存放在深度为2或3的层次较适合,但图中也显示资源平均深度有随着资源量增加而缓慢增加的趋势。

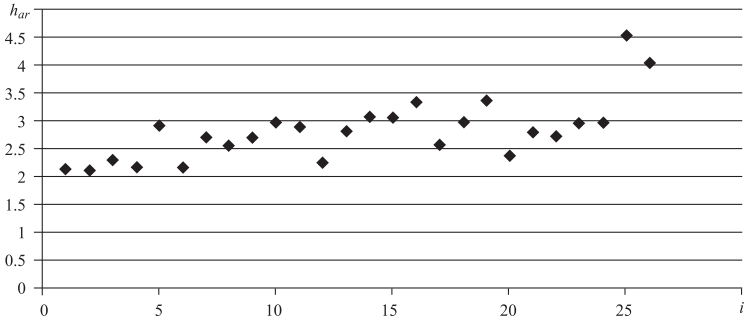

图6 资源平均深度与资源数关系

4.4 目录树深度与目录数量的关系

分别以 h_{mj}、h_{aj} 为纵坐标、以 n_{jd} 为横坐标作图(见图7),其中每个点对应一棵树的结构最大(平均)深度和目录数量。从图7中蓝色散点分布能看出当 n_{jd} 增加时 h_{mj} 的变化情况。例如,当 n_{jd} 在100以内增加时,h_{mj} 绝大部分在1到6之间,很少超过6,h_{mj} 的增加显然出于对资源细化分类的需要。但是当达到一定深度之后,n_{jd} 再增加,h_{mj} 也无明显增加,图中基本在10以下。这反映了人们在组织资源时,会根据实际需要增加目录树的深度,但是人们对目录树的最大深度一般有上限值,不习惯目录树的深度超过这个值。

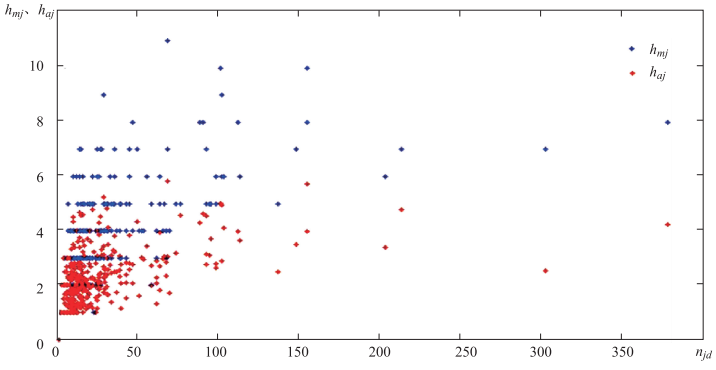

图7 目录树的结构深度与目录数量关系

进一步研究目录树的结构平均深度 h_{aj}，如图 7 红色散点所示。可以看出，h_{aj} 与 h_{mj} 的变化趋势相一致，但前者明显小于后者，通常 $h_{aj} \leqq 6$，而 $h_{mj} \leqq 10$。这说明在平均情况下，人们管理资源时会把组织结构限制在一个更浅的程度。

从目录的结构深度随目录数量的变化情况可以看出，随着 n_{jd} 的增加，人们不会明显增加树的深度，而是尽量使组织结构扁平。

5 结语

目前，越来越多的研究关注如何从标签中构造贴合人们习惯的大众分类体系。这样有助于设计符合人们组织共识的资源导航结构，帮助人们在海量信息环境下更好地利用互联网资源。然而在利用标签自动形成分类体系时，往往是研究者人为设定分类结构的参数，没有实验证明参数的合理性。

与现有以标签为对象的研究不同，本文以互联网上可收集到的真正由大众自发构造的层次组织结构——FTP 目录树为研究对象，从其层次组织中揭示人们管理资源时的结构规律。通过考察 1,481 个目录树中 44,972 个软件资源及其摆放位置，找到层次组织结构中的目录平均容量、资源平均深度、结构深度等特征，刻画这些特征随着资源数量和目录数量的变动情况。本文所选的软件类资源在互联网 FTP 站点上分布广泛，具有代表性，其数量及其复本数在样本站点中均呈幂律分布。

本文的主要结论包括：

(1) 对于用户查找浏览资源来说，资源存放在深度为 2 或 3 的层次较适合。

(2) 当要管理的资源量为 200 以内时，人们可容忍平均 1 个目录中包含 10 个左右的软件资源，当资源量更大时，人们可容忍平均 1 个目录中包含 15 个左右的软件资源。如果一个目录中资源数过多，增加目录数将其分而治之是人的朴素做法。

(3) 人们在考虑组织结构时，会在分类体系复杂度和分类精细度之间权衡，随着资源数量的增加，人们倾向于在同一个目录下放置更多的资源，降低目录数量增加速度，尽量避免过多目录导致分类系统复杂化。

(4) 在使用层次结构管理资源时，人们会对组织结构的深度有一个习惯限制，树的目录结构平均深度一般在 6 层以下。随着分类体系的复杂化，人们不会明显增加它的深度，而是尽量使组织结构扁平。

需要说明的是结论(1)和(4)的关系，(1)中所述为资源深度特征，(4)中所述目录结构平均深度是刻画目录树的深度特征，它由剪枝树 t'' 的叶子节点深度决定，如果多数资源是放在层深较浅的内节点，是不会影响结构平均深度的。

我们对目录结构之外的研究还在进行，如目录命名所体现的组织意图和功能，及其是否与标签具有相似性。此外，由于本文分析的对象仅限于 FTP 站点中的资源，且选取的资源类型单一，希望以后能够扩大分析的范围以及选择更多的资源类型，从而在更一般的情况下来考虑用户组织资源的规律习惯。

致谢：感谢北京大学深圳互联网工程中心的雷凯老师、叶牡丹同学提供了本文所需的 FTP 数据集。感谢北京师范大学信息管理专业 2008 级的全体同学标注了实验数据。

参考文献：

[1] CNNIC. 第21次中国互联网络发展状况调查统计报告[OL]. [2012-04-02]. http://www.cnnic.cn/research/bgxz/tjbg/. (CNNIC. The 21st development survey report of the China Internet[OL]. [2012-04-02]. http://www.cnnic.cn/research/bgxz/tjbg/.)

[2] Begelmana, Keller P, Smadja F. Automated tag clustering: Improving search and exploration in the tag space [C]//Proc. of 15th International World Wide Web Conference. WWW2006.

[3] H G-M Paul Heymann. Collaborative creation of communal hierarchical taxonomies in social tagging systems[R].

InfoLab, Stanford University, 2006. [2011 - 03 - 01]. http://ilpubs.stanford.edu:80901775/?auth = basic.
[4] Xian W, Lei Z, Yong Y. Exploring social annotations for the semantic web[C]//Proceedings of the 15th international conference on World Wide Web. ACM, 2006: 417 - 426.
[5] Anon Plangprasopchok, Kristina Lerman, Lise Getoor. Constructing folksonomies by integrating structured metadata [C]//In Proc. 19th International World Wide Web Conference. ACM, 2010: 1165 - 1166.
[6] Christopher H B, Nancy M. Improved annotation of the blogosphere via autotagging and hierarchical clustering [C]//In Proceedings of the 15th international conference on World Wide Web. ACM, 2006: 625 - 331.
[7] Kiu Ching-Chieh, Tsui E. Taxonomy-Folksonomy integration from knowledge navigation through unsupervised data mining techniques[C]//In Proceeding of the Workshop on Integrating Taxonomies and Folksonomies from Enhanced Knowledge Navigation at Practical Applications of Knowledge Management. 2008: 19 - 27.
[8] Zhou M, Bao S, Wu X, Yu Y. An unsupervised model for exploring hierarchical semantics from social annotations [J]. In Lecture Notes in Computer Science-The Semantic Web. 2007, 4825: 680 - 693.
[9] Golder S, Huberman B A. Usage patterns of collaborative tagging systems[J]. Journal of Information Science. 2006, 32.
[10] Kipp M E I, Campbell D G. Patterns and inconsistencies in collaborative tagging systems: An Examination of tagging practices[C]//In Proceedings of the 2006 Annual Meeting of the American Society for Information Science and Technology. 2006, 43(1): 1 - 18.
[11] Kipp M E I. @toread and cool: Tagging for time, task and emotion[C]//In Proceedings 8th Information Architecture Summit. Las Vegas, Nevada, USA, 2007.
[12] 陈翀,闫宏飞. 网络资源命名及用户命名行为的分析[J]. 情报学报, 2009, 28(4). (Chen Chong, Yan Hongfei. Web resource naming conventions and user behavior analysis[J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(4).)
[13] Markus H, Tanja N, Christian W. Tree, funny, to_read, google: What are tags supposed to achieve? A comparative analysis of user keywords for different digital resource types[C]//In Proceeding of the 2008 ACM workshop on Search in social media. ACM, 2008.

陈　翀　北京师范大学管理学院信息管理系讲师。
通讯地址:北京市海淀区新街口外大街19号。邮编:100875。
罗鹏程　北京师范大学管理学院信息管理系2011级硕士研究生。通讯地址同上。
刘晓兵　北京大学信息科学技术学院硕士研究生。
通讯地址:北京大学信息科学技术学院网络所。邮编:100871。
陆　炀　北京大学信息科学技术学院硕士研究生。通讯地址同上。

(收稿日期:2012 - 04 - 03)