Connotea 中 social tagging 机制研究

张　玫[1,2]　张晓林[1]

[1]（中国科学院国家科学图书馆 北京 100080 ）

[2]（中国科学院研究生院 北京 100049）

【摘要】选取 Connotea 为研究对象，统计分析了标签、被标引资源、标引者两两关联关系，发现标签覆盖资源范围较广，标签共现现象突出，部分用户标引活跃，资源平均标签数较低，用户对内容关注度规律性转移，科研领域的用户比较多地使用规范词。提出增加标引词结构关联和细粒内容定位标引。

【关键词】 Connotea　social tagging　folksonomy 标签　用户行为

【分类号】 G250

A study on the mechanism of social tagging in Connotea

Zhang Mei[1,2]　　**Zhang Xiaolin**[1]

[1]（National Science Library , Chinese Academy of Sciences, Beijing 100080, China）

[2]（the Graduate School of Chinese Academy of Sciences, Beijing 100049, China）

[Abstract] Using Connotea as a social tagging test bed, the relations between tags, relations between tags, resources, and users are analyzed, patterns are discovered to show the broad resource coverage of tags, strong co-occurrence of tags, high tagging activity among active users, low average tags per resource, regular shifts of users' tagging interests, and high usage of thesauri terms by users in scientific fields. Suggestions made to improve structural relations among tags, and to increase tagging granularity.

[Keywords] Connotea　social tagging　folksonomy　tags　users'behavior

1. 引言

Social tagging 又名 social bookmarking，作为一种用户驱动和群体交互式的标引机制，允许用户对资源赋以个性化标签，并可通过标签的聚合和相关度来实现信息组织。实质上，social tagging 利用用户与资源、资源与资源以及用户与用户之间的对应关系，把分散的资源及用户联系起来，帮助用户更好地发现资源以及发现与资源相关联的用户[1]。目前代表性的 social tagging 工具主要包括用于标记图片资源的 Flickr[2]，标记学术资源的 Connotea[3]及 CiteULike[4]，标记网络资源的 del.icio.us[5]等。

Social tagging 作为一种用户驱动的信息组织机制，提供了研究用户认知行为和信息组织行为的良好环境。可以针对用户的标引认知（用户用什么标签引用什么资源）、标引聚合（有多少用户使用相同的标签标引不同的资源）、标引分歧（有多少用户使用不相同的标签标引

相同的资源）、标引共现（有哪些对标签多么经常地被同时用来标引同一个资源）、标引者共现（有哪些对用户多么经常地用同样的标签来标引相同或不同的资源）、标引传递（标签使用按时间在用户间和资源间的传递）等现象进行研究，揭示出用户的信息认知和组织行为。

本文选取 Connotea 系统，具体分析 social tagging 中的用户行为特征，并提出值得改进的地方。Connotea 于 2004 年末由英国《自然》出版集团（Nature Publishing Group）创立，能自动识别并抽取特定学术网站或数据库（如 Nature、Science、D-lib Magazine、PubMed 等）中的书目信息，且能与常见的桌面参考文献管理工具（如 EndNote）进行数据交换[6]。本文依据数据是 2007 年 2 月 5 日到 20 日期间 Connotea 的相关数据。

2．Connotea 中体现的 social tagging 普遍机制

用户、资源、标签是 social tagging 中的要件。本文首先分析验证它们两两之间的关系。

2.1 标签与资源的关联分析

标签是用户在描述资源时自由选用的词汇，而 social tagging 正是通过同一标签对不同资源和同一资源对不同标签的聚合作用来不断扩充主题（标签）和资源间的动态联系。在 Connotea 中，用户可以检索同一标签所标引的所有资源，揭示资源之间存在的内容相关性，反映通过标签发现新资源的能力；同时，Connotea 可以从资源角度聚合用户行为，即通过选定某资源，揭示标注过该资源的所有用户及其采用的标签，反映不同用户对同一资源的不同理解，帮助人们从不同角度加深对该资源的认识。此外，我们还可以通过研究某一标签的共现标签，深化对用户知识认知的认识[7]。

为了准确了解 Connotea 中标签与资源的对应情况，本文对三个特定标签（folksonomy, ajax, diabetes）所标记的资源及共现标签进行统计，调查结果如表 1 所示。

表 1 标签的共现及与资源的对应情况

	出现次数	共现标签数	去重后的共现标签数	标引的资源数量
folksonomy	428	1022	299	195
ajax	458	1242	516	380
diabetes	437	1224	566	424
平均值	441	1163	460	333

从上表中可以看出，每个标签对应的资源总量平均为 330 条，表明标签具有相当的资源发现能力，而每个标签的共现标签平均有 460 个，说明不同用户对同一资源的理解具有很大的差异。这种现象一方面保证人们能从不同角度检索到该资源的同时，一方面也限制了检准率。此外，Connotea 还支持对多个标签的联合检索来提高检索精度，但若用表 1 中的标签数及其标引的资源总量来计算资源拥有标签数的均值，我们可以发现每条资源只有大约不到 5 个的标签，这也使得联合检索功能受到一定限制，而如何在保证检全率的基础上进一步提高检准率将是值得 social tagging 工具长期探讨的问题。

.2.2 用户与资源的关联分析

Connotea 提供按用户名来聚合资源的功能，可以浏览某一用户所有的标引活动，既可以反映该用户对 Connotea 的使用率，又可以按照资源（通过其 URL）来聚合标引了同一资源的用户，继而发现与之具有相同或相似兴趣的人，并可通过追踪他们对其他资源的标引过程来发现新的可能关联的兴趣及相应资源。

鉴于 Connotea 提供的“Recent Activity”功能按标引时间先后排列其所有的标引资源，本研究从中选取最近进行了标引活动的前 100 名用户统计其标引资源总量，随后在最近的标引活动中，选取其标引的前 100 条资源统计其用户总量；此外，为了更准确地衡量标引活动随时间推移的变化情况，并从一定程度上消除仅选取最近发生的标引活动来衡量所有标引活动可能存在的片面性，本研究又选取最早进行标引活动的前 100 名用户统计其标引资源总量，最后选取最早被标引的前 100 条资源统计其用户总量，其统计结果如表 2、表 3 所示。

表 2 用户对应的资源数量

	最近的标引	最早的标引	平均值
100 名用户标引总量	27579	7400	14160
用户平均标引量	270	74	142

表 3 资源拥有的用户数量

	最近的标引	最早的标引	平均值
100 条资源对应的用户总量	107	505	306
每条资源对应的用户数量	1.07	5.05	3.06

首先，若去掉时间因素，仅从用户的平均标引量来看，每个用户约拥有 142 条标引记录；从资源拥有的平均用户量来看，每条资源大约只能聚合 3.06 位用户。这两项数据表明，目前 Connotea 用户对资源的标引率较高，但他们大都各自为战，很少利用 Connotea 提供的资源发现手段，如标签、用户及资源的聚集等，Connotea 对于他们的价值主要体现在管理自身资源链接而非发现更多资源，用户间的交互比较有限，social tagging 群体互动优势的还不是很明显。

其次，若从时间推移角度考察用户的平均标引量和每条资源的平均用户量，则可以发现：用户的平均标引量随时间的推移有明显的提高，若对这两个时间段的具体数据作进一步分析，则又可以发现，在最早进行标引的 100 名用户中，有 68％的用户标引量都低于 10 条；在最近进行标引的 100 名用户中，仅有 30％的用户标引量低于 10 条，低于均值（270 条）的用户比例已超过 90％，上述数据说明，在 Connotea 投入使用的早期，用户的标引活动多属临时行为，主要体现在标引少数几条资源后便很少继续使用；但随着时间推移，大部分用户对 Connotea 的使用逐渐转化为持续行为；上述数据还反映出大部分用户标引量都集中在少数用户上，即这小部分用户的标引频率相当高，且该现象与时间段无密切联系，而这些用户也是今后研究标引行为的过程中值得重点关注的对象。此外，从表 3 中还可以看出，较早标引的资源较最近标引的资源聚集了更多的用户，这说明目前较低的资源平均用户量与人们对 connotea 资源的标引速度有直接关系。

2.3 标签与用户的关联分析

在 Connotea 中，每个用户的标引记录按照时间先后顺序排列，这有助于了解用户标签随时间推移的分布情况。和前面 2.2 的方法类似，本研究在“Recent Activity”中选取最近进行标引活动的 3 位用户，利用辅助工具 Connotea Explorer[8]，列出该用户的所有标签中使用率最高的前 6 个标签，然后借助 Excel 统计出这些标签的使用率随时间的增长情况，具体结果如图 1－3，其中横坐标代表时间，纵坐标代表该标签的使用率，而不同的标签则用不同颜色表示。

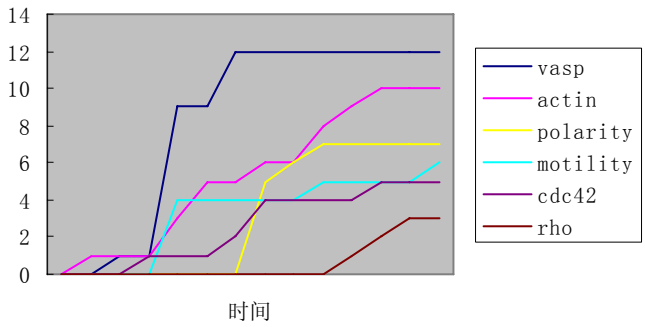

图 1 用户 1（derektwong）的标签增长情况

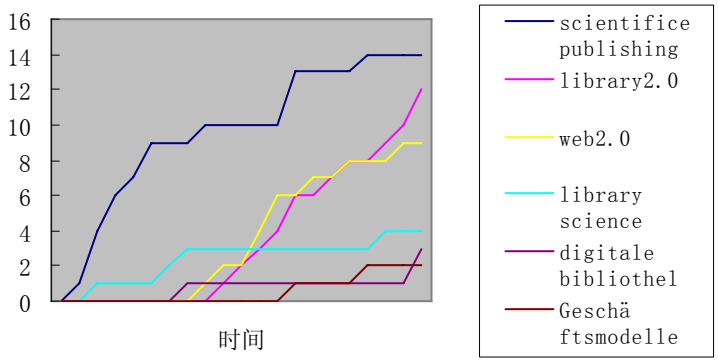

图 2 用户 2（kontext）的标签增长情况

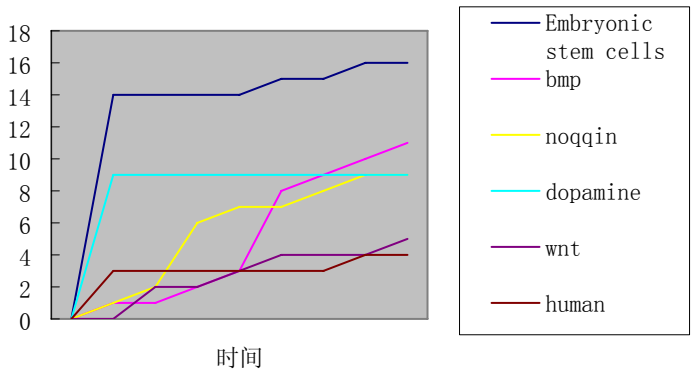

图 3 用户 3（theand）的标签增长情况

在标签增长曲线中，线段的斜率代表标签使用率的增长速度。线段在某段时间内的斜率越大，这个标签的使用率就增长越快；而平行线段则表示这个标签在这段时间内的使用率较低，用户很少使用、甚至可能没有使用此标签。

通过对上面三图的比较研究，可以发现它们具备的共同特征。首先，在每个用户的每个标签增长曲线中都存在一些平行线段，且这些平行线段所占比例最大；第二，许多曲线在开始出现斜率较大线段后就进入一个较长时间的平行线段，典型如图 1 中的 polarity 和 rho 及图 3 中的 dopamine 等标签，它们的增长曲线在经过连续的一次增长后便持续处于与横坐标平行的状态。这些现象表明了标签的生命周期具有阶段性，即用户研究问题的视角可能是在不断地转移。若从特定标签的角度来看，则说明用户对它的使用率可能是集中在某一个或几个时间段内，在其余大部分时间内用户对该标签所代表的问题关注度非常低，而对某些标签来说，用户对它们的关注则属于一次性的短期行为。若选取相同时间段来观察不同标签的斜率，我们发现，增长趋势越接近的标签，其相关性也越高，如图 2 中的标签 library2.0 和标签 web2.0，图 3 中的 Embryonic stem cells 和 dopamine，该现象可以从一定程度上反映出这些标签的共现频率较高的事实，有助于人们判断用户研究问题的视角。

3．Connotea 中体现的 social tagging 在科研领域的具体特性

根据 Connotea 的网站介绍，其用户以科研人员和医务工作者为主[10]，且该群体在日常工作中与规范词汇的接触机会较其他群体要高出很多，因此本研究拟从标签的规范性入手，调查他们在标引资源时是否相对采用了较为规范的词语。

本研究在“Recent Activity”中选取最新标引的 150 个标签，同时利用美国国会图书馆主题词表（Library of Congress Subject Headings, LCSH）[9]和医学主题词表（Medical Subject Headings，MeSH）[11]来判断每个标签是否属于上述词表中的规范词，若某标签同时属于 LCSH 和 MeSH 两个词表，则只把它算在 MeSH 的范围内。统计结果如下表所示：

表 4 标签规范性统计情况

	规范词		非规范词			
	LCSH	MeSH	包含规范词	规范词的一部分	规范词的单复数形式	其他
标签数量（个）	27	45	4	17	4	53
所占比例	18.0％	30.0％	2.7%	11.3%	2.7%	35.3%
合计比例	48％		52％			

从上表我们可以看出，属于规范词的标签约占总数的一半（48％），而在非规范词中，除去与规范词联系紧密的词语（如规范词的单复数形式等）外，真正与规范词完全没有联系的标签仅有三分之一左右（35.3%）。对于这部分词，主要有以下几类：过于宽泛的实词（如 photo）、最近出现的词汇（如 ajax）、专用词汇（如 o’reilly）以及非英文单词（如 rescusicitation），而在其他 social tagging 工具中较为常见的拼写错误在 Connotea 中却极少出现。

此外，值得注意的是，有不少用户把合成词拆分成几个独立的单词作为标签，但仍按

词组顺序排列，如把 web 2.0 拆分成 web 和 2.0 两个标签，这说明了部分用户不太了解 Connotea 中关于把词组作为标签的规则，该现象也削弱了 social tagging 的实际效用，因为很多组成词组的单词与词组的意思差距很大，甚至可能毫无意义，进而影响了利用标签来发现资源的能力。

4．Connotea 中仍待改进的地方

Connotea 作为 social tagging 工具的代表，尽管能提高人们发现资源的能力，但由于 social tagging 机制本身的缺陷，因此它仍有一些不足之处需要改进。

4.1 标签的平面性

在 Connotea 中，标签之间是平等关系，其他分类体系中最基本的词间关系，如上位类、下位类等在 Connotea 中均无法体现，且由于一词多义及同义词现象较为普遍，加上不能很好地在诸多标签中给某特定标签定位，因此无法揭示该标签与其他标签之间复杂的关系，这将容易妨碍人们宏观把握知识的体系结构，从而导致他们失去很多查找新资源的途径。

虽然 Connotea 提供了“相关标签”的功能，从一定程度上缓解了标签平面性所带来的缺陷，但这种方法并不能完全解决上述问题。有学者提出把标签先按人为大类存放的基础上再允许用户对其细分的方法[12]，但划分标签的过程实质是把事先存在的分类体系强加于用户，违背了 social tagging 最基本的原则——从用户自身角度进行知识划分，因此该方法并非上策。可以考虑在用户添加标签后，利用人工智能和 ontology 的方法对该标签进行分析定位，并向用户显示其所处的树状，甚至网状的知识体系结构，从而方便用户从整体上去认识该问题。

4.2 标引对象的局限性

目前，Connotea 可标引的对象仍局限于某个网页或某篇文章，但在科学研究中，有时对人们真正有用的信息只是其中的一部分，一个段落甚至一句话，因此 Connotea 的用户在通过标签找到该资源后，仍需要花一定的精力去寻找对自己有价值的那部分内容。这时，可以考虑让用户定位资源中的“相关内容区域”，例如一个或若干个自然段、一句或若干自然句、图或表等，使资源内的具体内容单元可以被区分和单独标引。具体方法或者可采用 Xpointer 和 Xpath 对文档的内容定位方式，或者对文档内容进行结构化标识封装，以便用户能快速地查找到有用的信息。不过，也要防止过分细分标引单元给标引者带来的负担。

参考文献：

1． Tony Hammond,et.al. Social Bookmarking Tools（I）：A General Review. D-Lib Magazine. 2005, 11(4). [2007-3-3]. http://www.dlib.org/dlib/april05/hammond/04hammond.html.

2． Flickr. [2007-3-3]. http://www.flickr.com/

3．Connotea. [2007-2-20]. http://www.connotea.org.

4．CiteULike.[2007-3-3]: http://www.citeulike.org/

5．del.icio.us. [2007-3-3]: http://del.icio.us/

6．同 3.

7．Ciro Cattuto, et.al. Collaborative tagging and semiotic dynamics. [2007-3-3]. http://arxiv.org/pdf/cs.CY/0605015.

8．Connotea Explorer: Pierre Lindenbaum 2006. Integragen. [2007-3-3]. http://lindenb.integragen.com/connotea.

9．WebDoc LCSH Interfaces. [2007-3-3]. http://fantasia.cse.msstate.edu/lcshdb/index.cgi.

10．About Connotea. [2007-3-3].http://www.connotea.org/about.

11．MeSH Browser. [2007-3-3]. http://www.nlm.nih.gov/mesh/MBrowser.html.

12．马然，陈树年. 网络信息分类组织的新星——Folksonomy. 新世纪图书馆，2006（4）：37－39