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，通过相应的数据统计和分析调查，分别从资源、用户及标签三者之间的关系探讨了 social tagging 的普遍机制，再从标签规范性的角度探讨了 social tagging 在科研领域的特点，最后提出了 Connotea 中值得改进的地方。

【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] The article studies on the mechanism of social tagging in Connotea in terms of the relationships among resources, users and tags, discusses the normative of tags in the scientific and medical fields, and gives some suggestions on the improvement of Connotea.

[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、资源及标签是 social tagging 中最基本的三个因素。本文分别从它们两两之间的关系入手。

2．1 标签与资源的对应情况

标签是用户在描述资源时自由选用的词汇，而 social tagging 正是通过标签对资源的聚合以及对相关标签的聚合来不断扩充资源间的联系，从而达到提高用户检全率的目的。在 Connotea 中，用户可以检索同一个标签所标引的所有资源，该项功能反映了资源之间存在可能的相关性，这也从侧面反映出此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，即通过选定某资源的 URL，来显示标注过该资源的所有用户 ID 及其采用的标签，此功能则反映了不同用户对同一资源的不同理解，这将有助于扩展人们从不同角度加深对该资源的认识。此外，我们还可以通过认识某一标签的共现标签，扩大自身的研究角度[7]。

为了更加准确地了解 Connotea 中标签与资源的对应情况，本研究对三个特定标签（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约为 330 条，这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约有 460 个，这说明不同用户对同一资源的理解具有很大的差异，它在保证人们能从不同角度检索到该资源的同时，也增加了检准率的难度。尽管 Connotea 支持对多个标签的联合检索，这可以帮助用户提高检索精度，但由于每条资源只有大约不到 3 个的标签，因此，上述联合检索功能受到极大的限制，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 用户与资源的对应情况

Connotea 提供按用户名聚合资源的功能，以方便人们浏览某一用户所有的标引活动，而从中又可以反映出该用户对 Connotea 的使用率；同时，用户还可以按资源的 URL 实现聚合，这样人们便可以在标引过同一资源的不同用户中，判断出可能与之具有相同或相似兴趣的人，之后再追踪他们的标引过程，从而发现更多有用的资源。

为保证调查数据的可靠性，本项调查采用随机取样的方法，从 Connotea 提供的“Recent Activity”中选取最前面的 100 名用户统计其标引资源总量，结果为 27579 条；随后再选取前 100 条资源统计其用户总量，结果为 107 位。

从用户的平均标引量来看，每个用户约拥有 270 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 1.07 位用户。这两项数据表明尽管 Connotea 的使用率较高，但用户大都各自为战，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，他们更多的是依靠其他方法从 Connotea 系统外部进入内部，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，因此用户间的交互非常有限，这严重阻碍了 social tagging 群体互动优势的发挥。

2．3 标签与用户的对应情况

在 Connotea 中，每个用户的标引记录按照时间先后顺序排列，这一功能将有助于我们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”中选取最前面的 3 位用户，利用辅助工具 Connotea

Explorer[8]，判断出在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、图 2、图 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。

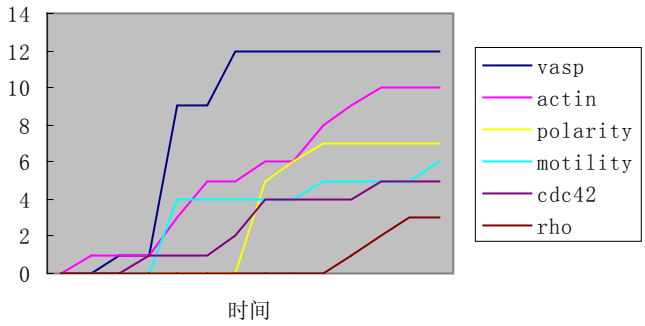

图 1 用户 1（derektwong）的标签增长情况

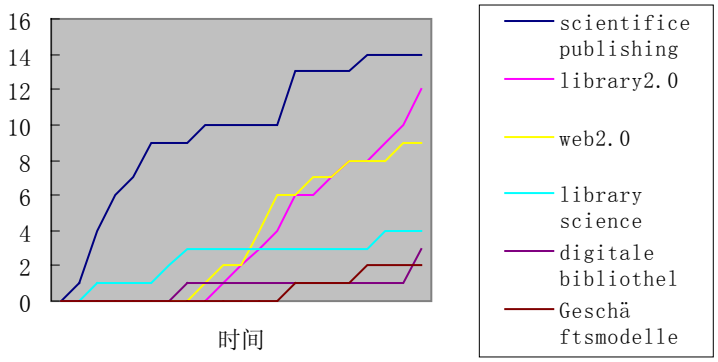

图 2 用户 2（kontext）的标签增长情况

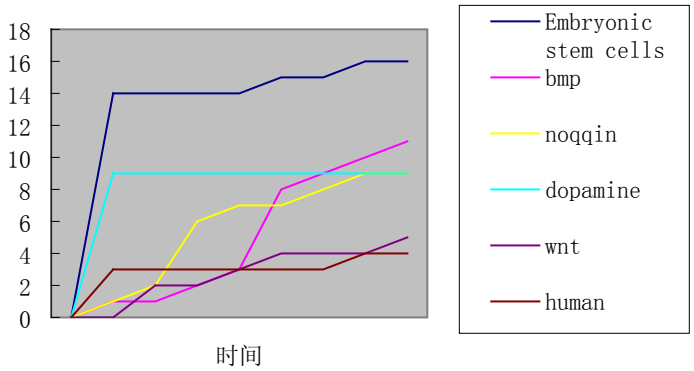

图 3 用户 3（theand）的标签增长情况

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，即线段的斜率越大，标签在这段时间内的增长速度越快，而与横坐标平行的线段表示标签在这段时间内的使用率没有发生变化，亦即在该时间段内用户没有使用此标签。

通过对上面三图的比较研究，我们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与横坐标平行的线段，且这些平行线

段在整条曲线中所占比例最大，且大都与斜线段交替出现，而斜率较大线段的出现频率又较为集中。在图中我们甚至还可以发现某些标签，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。上述这些现象都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总是在不断地转移。若从特定标签的角度来看，则说明用户对它的使用率总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；其次，若选取相同时间段来观察不同标签的斜率，我们又可以发现，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，同时也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

由于 Connotea 的用户以科研人员和医务工作者为主，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”中选取最前面的 150 个标签，同时以两个权威的主题词表作为判断标准：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）[10]，之后再依次对每个标签进行判断，检查其是否属于上述词表的规范词。统计结果如下表所示：

表 2 标签规范性统计情况①

	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，其主要类型有以下几种：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

① 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。

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，这将极大地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[11]，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。我们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，仍需要花一定的时间和精力去寻找对自己有价值的那部分内容。这时，我们应重新定义信息组织的基本单元，使之由过去的网页进一步细分为更小的信息单元，具体方法可考虑借鉴 XML 体系中相关技术的思路，如用 Xpointer 和 Xpath 对文档中的具体内容进行定位，细化标引单位，以便用户能快速地查找到有用的信息。

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

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

[2] Flickr. [Retrieved 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