·资源管理·

国外社会化标注系统中标注行为研究现状

Review on the Tagging Behavior of Social Tagging System Abroad

杨青云 裴 雷 吴克文

(湖南湘西民族广播电视大学 吉首 416000) (南京大学信息管理系 南京 210093)

摘 要 通过检索国外研究社会化标注系统中标注行为的相关研究成果,首先论述了社会化标注相关概念,再从标注动机、标注过程、标注结果、垃圾标注等四个方面对已有研究进行归纳分析,探讨社会化标注行为的研究现状与研究趋势,认为现有研究不足之处主要存在于反垃圾标注研究、用户标注过程与标注动机研究等三个方面,对未来研究工作的开展具有导向意义。

关键词 社会化标注 标注行为 社会认同 研究进展

中图分类号 G203 **文献标识码** A **文章编号** 1002-1965(2009)11-0185-04

社会化标注是众多信息用户根据自己的需求,选择合适的网络信息资源,并根据自己的认知水平,确定与之相匹配的社会化标签进行标注的过程。在整个过程中,用户充分体现主动性和个性化,而且也能够参考和借鉴其他用户所选择的资源和标签,体现出一定的交互性和社会性。通过研究用户标注行为,可以了解社会化标注系统用户的标注动机、标注方式、标注影响因素以及信息资源利用等方面情况,因此研究社会化标注系统中的用户标注行为具有重要意义。

1 相关概念定义

1.1 **标签、标注、书签** 标签(Tag)是一种关键字、类目名称或者元数据。标签是用户自由选择的文本关键字,定义了用户思维中的概念与对应网络资源这两者之间的联系[1]。标注(Tagging)是定义标签的过程,是用户为某一资源加入描述性元数据并赋予某项资源关键字来达到划分类目的过程。标注建立起了标签(或标签组)、资源、用户、时间四者之间的联系,这种联系即为书签(Bookmark)。

1.2 **社会化标注** 社会化标注(Social Tagging)也称为协同标注(Collaborative Tagging)、社会化分类(Social Classification)等[2]。社会化标注指的是众多用户所进行的标注行为,用户基于个人的或者社群的目的对资源赋予标签,这种看似私人的行为在众多用户的共同参与时产生了社会性价值:用户对某个特定资源指定了数个标签,标签因此成为了连接用户与资源之间的桥梁。资源之间也可以通过某种方式彼此相互连接(如被同一用户标注过的资源、使用了同一标签的资源),并且用户也可以透过标签或资源的社会网络来相互关联[3]。

1.3 **社会认同理论** 社会认同理论(Social Proof)认为,人们进行是非判断的标准之一就是看别人是怎么想的,尤其是需要决定什么是正确行为的时候[4]。这一理论特别适用于在不确定性的环境下行为人的决策。面对未知的事件,人们很自然地环顾周围了解其他人的反应[4]。

2 标注行为研究进展

2.1 **对标注动机的研究** Shilad Sen归纳出标注行为普遍来说可以支持五种任务:a.自我表达(Self-expression):标签可以帮助用户表达个人意见;b.组织行为(Organizing):标签可以帮助用户组织个人信息资源;c.学习(Learning):帮助用户了解更多关于某个资源相关的知识;d.寻找(Finding):帮助用户找到个人想要的信息资源;e.决策支持(Decision support):帮助用户决定是否使用或浏览某个信息资源[5]。

Golder &Huberman确定了标签的7种功能,其中前五种是用来描述对象资源的属性,如来源、属性、种类、所有人、数量等,这些种类的标签可能产生于组织性动机,社会性动机或者潜在的目标受众。第六种标

收稿日期:2009-05-27 修回日期:2009-08-03

作者简介:**杨青云**(1980-),男,讲师,研究方向电子商务;**裴 雷**(1981-),男,讲师,研究方向为信息资源规划;**吴克文**(1985-),男,博士研究生,研究方向为互联网用户行为。

签类型(自引 self－reference),反映的是向外界受众传达这种所有权的可能性意图,或者被用做个人信息管理。最后一种类型,任务组织型的标签,表明了用作个人信息组织[6]。经过整理,社会化标注的功能和动机如表 1 所示。标签功能的具体含义分别为:a. 确定事物或人:标签主要确认书签术语的主题,这些术语是人们在讨论信息的内容时,所用到的特性层次的一些普通名词,以及一些专有名词;b. 确定特征的类型:标签能够根据确定的人或物,来确认书签属于的类型;c. 确定拥有者:一些书签能够根据书签内容的拥有着或创造者被标注;d. 提炼分类:一些标签不是孤立存在的,它们要么本身就是分类,要么就是有执行分类的功能;e. 确定信息质量和特征:根据标签者对内容的意见来对书签标注适当的形容词;f. 自引:开头具有"我的"前缀的标签确定了内容与标注者自己的术语相关;g. 任务组织:当收集与任务执行有关的信息时,该信息也许会根据该任务被标注,这是为了组织该信息。

表 1 社会化标注的功能及动机

标签功能	可能的动机	举例
确定事物或人	组织性、引起注意	Javascript
确定类型	组织性	Article、blog
确定拥有者	组织性、贡献和共享	Bill Gates
提炼分类	组织性、互动和竞争	25100
确定质量和特征	组织性、表达观点	Funny、stupid
自引	组织性、自我表达	Mystuff、mycomment
任务组织	组织性	Toread, jobsearch

2.2 对标注过程的研究 标注与索引概念相似。主题索引包含两个主要步骤:概念分析(Conceptual Analysis)和转译(Translation)。概念分析涉及到决定目标资源与什么有关以及与什么相关,其分析结果常常依赖于用户在对目标资源进行标记时的需求与兴趣,不同的人可能对资源的不同地方有兴趣,而导致概念分析结果不一。转译的目的是找到一组适当的索引词汇的过程,而这些索引词汇可以表征概念分析的结果。研究显示,在不同的索引者之间要达到高度一致是很困难的,并且受到很多因素的影响,其中一个因素是所使用的受控词汇表的区别。同义词以及多义词现象是转译过程中经常会出现的问题[2]。

Sinha 认为标注是一种分析过程[7], Benjamin Szekely 甚至认为标记是一种不需要经过太多思考而将关键词归属到某个目标对象的过程[8]。标注是一个相当简单的过程,当用户浏览过需要标注的目标资源后,经过比较、衡量目标资源和候选概念之间的相似性,然后记录下标签。标注行为在这个阶段中并不牵涉过滤或筛选行为,可以为目标资源标注任何数量的联想词汇[7]。

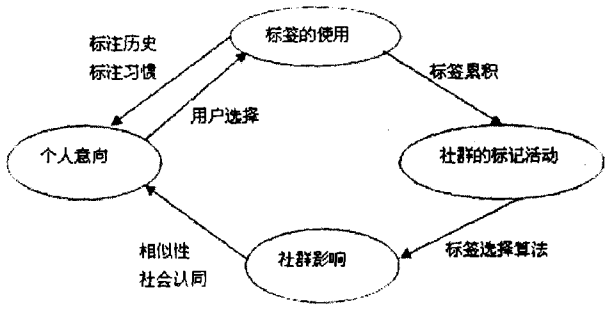

图 1 社群影响和个人意向之间的关系[5]

图 1 显示了可能会影响用户标注的三个因素,包括应用标签的三个因素,包括:a. 用户会基于过去的标注行为来使用标签,即个人意向(Personal Tendency);b. 用户的标注行为会受到其他用户的影响,即社群影响(Community Influence);c. 系统内置的标签选择算法(Tag Selection Algorithm)[5]。

标签的产生和标注习惯的累积会影响用户后继的标注行为,用户所使用的标签,是用户已有知识体系中存在的一种知识,通过后继的行为去改变这种知识的代价是很高的。用户也会受到习惯支配,倾向于重复他们过去经常表现的行为。Shilad Sen 认为,人们未来倾向使用的标签与他们过去使用过的标签词汇是差不多的[5]。

由于显性或隐性的社群的存在,社群会通过改变用户的个人意向来改变标签的选择[5]。Golder&Huberman 研究发现,用户对一个书签所标记的标签,随着时间的推移,各个标签的相对比例呈现一个相对稳定的趋势,由此认为这是用户收到其他先前标注者(社群成员)的标记行为所影响的[6]。Cattuto 也调查了最近被使用的标签是否会影响后继用户对该资源的标签的使用[9]。社会认同理论尤其适合用于解释用户如何去标注陌生的资源,因为"不确定性会促成社会认同的影响[4]",因此,标签推荐功能会给用户在标注时带来更多信息并影响标注结果。如果假设用户遵循最省力原则选取了推荐的标签而不是通过自我思考,根据信息搜寻理论,大量的标注行为会导致信息率(Information Ratio)的提升[10],并且,可以帮助用户构建能够支持其最佳书签组织的结构[11]。Philip J. Binkowski 认为,尤其是用户在标注那些具有复杂内容的网站时,社会认同的效果尤其显著[4]。Cosley 研究发现,推荐系统会引导行为的一致性,不管预测的信息正确与否,用户都会受到系统所显示的评价信息,而影响他们对某部电影的评价[12]。社会化标注系统中的标签推荐就像是其他标注者标记行为的替代品,浏览推荐标签会影响用户的标签选择,最终目的在于促进用户的标签聚合,标签聚合将会使不同用户的标签具

有较佳的一致性，会让标签的分享目的更容易达成[4]。

2.3 **对标注结果的研究**　Golder&Huberman 对 Delicious 从四个角度研究了用户标注结果：a. 观察用户标签集容量的变化。有的用户标签集容量增长迅速，有的则保持稳定；b. 通过考察用户的标签集研究用户通过用户标签容量探查用户标签的使用情况，发现虽然标签集容量在增长，但是标签的使用却不尽相同。有的标签的使用次数稳步增长，有的标签却是在缓慢增长；c. 通过对书签产生趋势的研究发现许多 URL（网络资源）在它们刚被进入 Delicious 时很快达到被标注的高峰期，同时也有很多 URL 只拥有较少的书签数，直到重新被利用而再次到达顶峰；d. 通过对标签的研究发现随着用户标注次数的增加，一些标签所占比例逐渐趋于稳定[6]。

在对 Flickr 的研究中，Cameron Marlow 随机挑选了 10 名用户，对其标签集容量的增长进行分析，同时，也研究了用户之间的标签词汇的重叠问题，结论是，随机抽取的用户在标签词汇上面重叠较少，一般都是常用词汇的重叠，而如果用户之间是朋友关系，则会有较高的标签重叠度[3]。

David R Millen 在研究 Dogear 系统时发现超过 80% 的书签只含有 3 个以下的标签，同时，通过调查访问发现用户对于该系统的基于标签的资源导航功能持正面态度[14]。

Shilad Sen 研究 MovieLens 时着重分析了标注行为的影响因素，发现新 MovieLens 用户比老用户（推出标注服务之前的用户）更愿意共享标注成果，并且标签数量在持续增长[5]。

Connotea 系统与 CiteULike 系统比较类似，都是为科研工作者提供的一种在线社会化书签系统。Ben Lund 研究 Connotea 发现用户的标签集容量分布服从幂律分布[15]。

Umer Farooq 在总结了早前的研究成果后发现，首先，以往的文献在数据量方面都不够显著，而目前的数据收集手段已经可以获得更大容量的数据集，其次，以往的研究并没有提出一套评估标注行为的方法体系，因此提出了一个具有六种衡量指标的体系去描述 CiteULike 系统中的用户标注行为，这六种衡量指标分别为：标签增长，标签重用，标签的显隐性，标签歧视，标签频率和标注方式[16]。其中，衡量标签的显隐性时需要去对目标资源（特指学术文献）进行内容分析，认为如果某个标签在目标资源中没有出现，其就具有更高的表征资源内容的能力，与文本分类中的 TFIDF 原理类似。

2.4 **对垃圾标注的研究**　正如 Email 领域充斥着垃圾邮件一样，社会化标注系统中也不乏垃圾标注。垃圾标注者可能对应的不是真实用户，而是自动程序。RobertWetzker 在研究了 Delicious 系统中最活跃的前 20 位用户后发现其中 19 位用户可能是自动标注程序，因为它们将数万的书签对应的 URL 地址指向了极少数域。作者分析其可能是由于 Delicious 系统开放了其 API 以供外部程序调用所致[17]。RobertWetzker 归纳的垃圾标注者的六种表现特征为：高活跃度、指向域少、单词标签使用量高、单次标签使用量低、大量标注以及综合情况[17]，如表 2 所示。

表 2　垃圾标注者的表现特征[17]

特征	说明
高活跃度 Very high activity	自动标注程序会比正常使用行为拥有高得多的在线时间
指向域少 Few domains	用户书签所指向的 URL 集中于少数域名
单次标签使用量高 High tagging rate	一些垃圾标注者单次标注时使用极多的标签
单次标签使用量低 Very low tagging rate	一些垃圾标注者看上去不习惯使用标签，持续上传不含任何标签的书签
大量标注 Bulk posts	机器标注会标签为大量的上传记录。但是也有可能是用户使用书签同步工具对系统进行操作。
综合 Combinations	上面因素的综合表现

Robert Wetzker 提出了“注意力扩散”（Diffusion of Attention）的概念，从标签分布的角度而不用借助过滤器就能降低垃圾标注的影响[17]。其分析方法是将在某个时间段内使用了某个标签的用户数定义为该标签的“标签注意力”，将初次使用该标签的用户数定义为该标签的“标签扩散度”，由此衡量某条资源吸引新用户的能力，这样处理反映的是整体用户的使用趋势，限制了单个用户的影响。

Beate Krause 从用户资料、用户地址、用户使用记录、标签语义特征等四个方面，25 个指标建立了针对 BibSonomy 标注系统的垃圾标注识别体系，实验证明使用传统分类技术（SVM、Native Bayes、J48、Logistic Regression）可行但是分类效果没有达到理想水平，基于共现特征（使用了相同标签、标签组或资源的用户）的分类取得了最佳效果[18]。

Grorgia Koutrika 等学者用实验模拟的形式构建了正常标注者与垃圾标注者共存的模拟标注系统，结论认为标注系统中使用的某些保护性机制是一把双刃剑，既可能限制了垃圾标注者的负面影响，也限制了正常用户的行为；复杂的反垃圾机制迟早会被攻破，复杂的机制会产生更复杂的对手；标注系统的繁荣靠的是大量用户的参与，高质量用户参与的越多，系统越能对抗垃圾标注[19]。

Zhichen Xu 认为，一个“特别”的标签可以有效地识别一个资源但是对其他用户发现其他资源便显得无用，相反，一个“普通”的标签适合于发现资源但是对于

精确定位资源不能起很大作用。一个好的标签组同时具有发现和标识两种功能[20]。Zhichen Xu 同时认为好的标签组需要具备如下特征：a. 对于资源的多方面属性具有高的覆盖度。例如，描述一个旅行 URL 的标签组应该含有如下标签：表示分类的“travel”，表示地点的“San Francisco”，表示时间的“2005”，特殊标签“GoldenGate Bridge”，主观性的“cool”等。显然，标签覆盖的角度越多，越有助于帮助用户了解资源的内容；b. 高使用率。如果一组标签被很多用户用来描述一个特定的资源，那么这些标签就不太可能是垃圾标注。该组标签显得比其他标签能更好表示资源内容，并且更可能被后继用户再次使用；c. 省力。用于标识某一资源的标签应该尽可能少，标签组能够标识的资源集合也应该尽可能少，用户只需要几步就能找到目标资源；d. 标准化。因为没有特定的约束规则或者语义集，标签可能具有各种各样的表现形式。不同的人可能在标注相同资源时使用不同的词。这种多样性有两种表现形式，一种是造词法角度的多样，如 blogs，blogging，blog，另一种是同义词，如 cell - phone 和 mobile - phone。这些多样性会增加系统内的噪声。一种可行的解决方案是在外表上尊重用户的标注，在内部处理上建立统一的标准型表示；e. 排除特性类型的标签。例如，用户个人信息组织的标签一般不具有广泛的使用性。因此，需要将这些标签排除于公共用途[20]。

3 结束语

社会化标注系统是 Web2.0 环境下的典型应用，为互联网信息组织与检索的改进、网络用户认知、知识管理等诸多领域提供了新方向。目前，国内相关领域的研究主要集中于概念引入阶段，起步较晚；国外研究虽然较为丰富，但也有不足，归纳如下：a. 反垃圾研究。垃圾标注对系统运行和学术研究的展开均有较大的影响，现有研究局限于从信息计量角度分析垃圾标注特征和从模式识别角度识别垃圾标注，缺乏垃圾标注的成因研究、垃圾标注对于标注质量影响的研究以及有效的垃圾标注解决方案等；b. 用户标注过程研究。现有研究只分析了影响用户标注的因素，并没有较好引入社会认同理论，尤其是用户标注陌生资源时的思维过程。再者，前任用户在标注时附带写入的注释对于后继用户的标注影响研究也未进行，因为注释可以帮助后继用户更好理解该资源并选择高质量标签；c. 用户标注动机研究。现有研究着重从标注者自身信息需求和信息组织出发进行研究，缺乏标注者与其所处的社会化标注环境的交互，如社团贡献等方面研究。以上所述研究不足都可能成为今后情报学、社会学、计算机学等领域研究社会化标注行为的研究重点，有望在今后研究工作中展开。

参 考 文 献

[1] Marieke Guy. Folksonomies Tidying up Tags? [EB/OL]. http://www.dlib.org/dlib/january06/guy/01guy.html,2009 - 4 - 20

[2] Jakob Voss. Tagging, Folksonomy & Co - Renaissance of Manual Indexing? [EB/OL]. http://arxiv.org/pdf/cs/0701072v2, 2009 - 4 - 2

[3] Cameron Marlow. Position Paper, Tagging, Taxonomy, Flickr, Article, Toread [EB/OL]. http://citeseerx.ist.psu.edu/viewdoc/summary? doi = 10.1.1.74.8883,2009 - 04 - 20

[4] Philip J. Binkowski. The Effect of Social Proof on Tag Selection in Social Bookmarking Applications [EB/OL]. http://etd.ils.unc.edu:8080/dspace/bitstream/1901/358/1/philipbinkowski.pdf,2009 - 5 - 27

[5] Shilad Sen. Tagging, Communities, Vocabulary, Evolution [EB/OL]. http://portal.acm.org/citation.cfm? doid = 1180875.1180904,2009 - 4 - 27

[6] Scott A. Golder. Usage Patterns of Collaborative Tagging Systems [J]. Journal of Information Science,2006,32(2):198 - 208

[7] Sinha . A Cognitive Analysis of Tagging [EB/OL]. http://rashmisinha.com/2005/09/27/a - cognitive - analysis - of - tagging/,2009 - 4 - 27

[8] Benjamin Szekely. Ranking Bookmarks and Bistros: Intelligent Community and Folksonomy Development [EB/OL]. http://labs.rightnow.com/colloquium/papers/tagrank.pdf,2009 - 4 - 2

[9] Ciro Cattuto. Collaborative Tagging and Semiotic Dynamics [EB/OL]. http://arxiv.org/pdf/cs/0605015v1,2009 - 4 - 27

[10] Gattis L F. Planning and Information Foraging Theories and Their Value to the Novice Technical Communicator [A]. Proceedings of the 20th Annual International Conference on Computer Documentation [C]. New York: ACM Press,2002:39 - 43

[11] Abrams D R M. Information Archiving with Bookmarks: Personal web Space Construction and Organization [A]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems [C]. New York: ACM Press,1998:44

[12] Dan Cosley. Is Seeing Believing? How Recommender Interfaces Affect Users' Opinions [A]. In: CHI 2003 [C]. USA: ACM,2003: 585 - 592

[13] C Cattuto. Semiotic Dynamics in Online Social Communities [EB/OL]. http://www.springerlink.com/index/T964J630305 07341.pdf,2009 - 4 - 27

[14] David R Millen. Dogear: Social Bookmarking in the Enterprise [A]. In: CHI2006 [C]. Canada: ACM,2006:111 - 120

[15] Ben Lund. Social Bookmarking Tools (II) A Case Study [EB/OL]. http://dlib.org/dlib/april05/lund/04lund.html,2009 - 4 - 27

[16] Umer Farooq. Evaluating Tagging Behavior in Social Bookmarking

(下转第 184 页)

或方法，可能存在好几种方法都是可行的。本体的构建要面向特定的应用目的、基于一定的专业领域、学科背景或研究课题。实际上，几乎每一个系统的开发都会导致一些不同的本体构建方案产生。目前的本体构建方法论还未能像软件工程那样成为"科学"或"工程过程"的完整方法论。因此只有综合借鉴已有方法，结合具体应用，才能构建出较好的实用本体。

其次，要尽可能的复用已有领域资源，这样可以有效解决共享和重用问题。构建本体的主要目的是达到知识共享，因此所提出的本体应该在最大程度上能为别人所接受。这就要求本体必须尽可能地涵盖多数本领域的人们对该本体的理解。生物医学领域的知识组织系统（Knowledge Organization System，KOS）如UMLS、MeSH主题词表等已经在生物医学研究中得到广泛应用，它们所包含的概念及其定义都是领域内所公认的，利用这些公认的概念和关系构建的本体当然也能为其他人所接受，并且许多的知识组织系统都已经提供了概念等价性和概念层次性的描述，为获得本体提供了很方便的转换基础。另外，复用已有领域资源还可以减少领域专家的干预，提高本体构建的效率。目前大家公认在构建领域本体的过程中，需要领域专家的参与和协作[10]。但是人为的干预具有很大的主观性，即使是领域专家，构建出来的本体也有可能是大相径庭的，这样构建的本体就违背了引进本体的初衷。而如果基于生物医学领域公认的词表或本体中的概念和关系来构建本体，这样的问题就不复存在了。因此生物医学领域本体的构建应该考虑充分利用现有的知识组织系统、兼容和复用现有的领域本体。

参 考 文 献

[1] 何 琳，杜慧平，侯汉清. 领域本体的半自动构建方法研究[J]. 图书馆理论与实践，2007，28(5)：26－27

[2] Hadzic M, Chang E. Ontology－based Support for Human Disease Study[A]. IEEE. Proceedings of the 38th Hawaii International Conference on System Sciences[C]. Hawaii: IEEE, 2005: 143－151

[3] Ontologies[EB/OL]. http://www.isi.edu/natural－language/projects/ONTOLOGIES.html, 2009－06－06

[4] Bernaras A, et al. Building and Reusing Ontologies for Electrical Network Applications[A]. Proc of the European Conf on Artificial Intelligence[C]. Budapest, Hungary: John Wiley and Sons, 1996: 298－302

[5] Uschold M, King M. Towards a Methodology for Building Ontology[C]. Workshop on Basic Ontological Issues in Knowledge Sharing: International Joint Conference on Artificial Intelligence. Montreal, Canada, 1995: 373－380

[6] Gruninger M, Fox M S. Methodology for the Design and Evaluation of Ontologies[A]. Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing held in conjunction with IJCAI－95[C]. Montreal, 1995

[7] IDEF5 Method Report[EB/OL]. http://www.idef.com/pdf/idef5.pdf, 2008－10－24

[8] Fernandez M, Gomez－perez A, Juristp N. Methontology: From Ontological Art Towards Ontological Engineering[R]. AAAI－97 Spring Symposium on Ontological Engineering, SS－97－06. Stanford University: AAAI, 1997: 33－40

[9] Natalya F Noy, Deborah L. Mcguinness Ontology Development 101: a Guide to Creating Your First Ontology[EB/OL]. http://protege.stanford.edu/publications/ontology-development/－ontology101.pdf, 2008－10－24

[10] 李 景. 本体理论及在农业文献检索系统中的应用研究—以花卉学本体建模为例[D]. 北京：中国科学院文献情报中心，2004

[11] Gruber TR. Towards Principles for the Design of Ontologies Used for Knowledge Sharing[J]. International Journal of Human－computer Studies, 1995, (43): 907－928

[12] International Classification of Diseases (ICD)[EB/OL]. http://www.who.int/classifications/icd/en/index.html, 2009－06－06

（责编：贺晓利）

（上接第188页）

Systems: Metrics and design heuristics[EB/OL]. http://research.microsoft.com/pubs/78959/p351－farooq.pdf, 2009－4－2

[17] RobertWetzker. Analyzing Social Bookmarking Systems: A del.icio.us Cookbook[EB/OL]. http://www.dai－labor.de/fileadmin/files/publications/wetzker-Delicious-ecai2008-final.pdf, 2009－4－2

[18] Beate Krause. The Anti－Social Tagger － Detecting Spam in Social Bookmarking Systems[EB/OL]. http://airweb.cse.lehigh.edu/2008/submissions/krause-2008-anti-social-tagger.pdf, 2009－5－27

[19] Grorgia Koutrika. Combating Spam in Tagging Systems: An Evaluation[EB/OL]. http://ilpubs.stanford.edu:8090/816/1/2007－30.pdf, 2009－5－27

[20] Zhichen Xu. Towards the Semantic Web: Collaborative Tag Suggestions[EB/OL]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.4194&rep=rep1&type=pdf, 2009－4－2

（责编：白燕琼）