分众分类法与受控词表的结合研究进展*

贾君枝

摘　要　分众分类法与受控词表作为网络知识组织系统的重要构成,对于有效地组织与管理网络资源具有重要作用。区别于受控词表自上而下的构建方式,分众分类法的自下而上、用户为中心的构建特色成为个人信息资源组织的方法之一。受控词表具有词汇规范性、严格性以及语义关系丰富等优点,分众分类法具有丰富的标签资源及用户数据,但也有标签自由性、随意性、不准确性等不足。如何利用二者的优势及时更新、修改、完善受控词表,最终达到二者优化的目标,是研究人员关注的问题。本文在调查国内外大量研究文献的基础上,对目前围绕受控词表与分众分类法的结合研究现状作了全面综述,预测其发展趋势,旨在推动国内在该方面的研究与应用,以综合发挥传统信息组织方法与现代新型信息组织方法的优点,实现对现有中文网络资源的有效管理。参考文献25。

关键词　分众分类法　受控词表　网络知识组织系统

分类号　G254

ABSTRACT　As parts of network knowledge organization system, folksonomy and controlled vocabulary play an important role in the organization and management of network resources. Folksonomy is constructed by the user-centered ways of bottom-up to organize the personal information resources while controlled vocabulary is designed by top-down ways. The controlled vocabulary is normative and strict and has rich semantic relations to make up the weakness of Folksonomy whose tags are used freely, randomn, and inaccurate. How to take advantage of rich label resources and user data to complete the update, modification, improvement of controlled vocabulary will cause the attention of researchers in order to optimize them. The paper investigates a large number of rpapers both home and abroad and then conducts a comprehensive review to sum up the characteristics of current research to forecast its trend of development, aiming to promote the development of domestic research and application, integrating the advantages of traditional information organization methods with the modern methods and improving the efficiency of management of Chinese network resources. 25 refs.

KEYWORDS　Folksonomy.　Controlled vocabulary.　Network knowledge organization system.

CLASS NUMBER　G254

1　导言

随着互联网的应用普及、网络用户数激增,有效地组织网络资源,提高网络资源的管理效率成为必要。网络知识组织系统(Network Knowledge Organization System,NKOS)作为对知识结构进行系统化描述和说明、支持网络信息的表示与检索等活动的知识组织系统[1],搭建了用户需求与网络资源之间的桥梁,在用户不需要知道相关知识的前提下,帮助用户浏览或检索到感兴趣的内容。网络知识组织系统既包括分类表、主题词表、规范档(人名、地名)等传统信息组织工具,又包括语义网、本体、分众分

*　本文系山西省人文社科基地项目"山西省农民信息服务研究"(项目编号:20093004)研究成果之一。

类法等新型网络组织工具[2]。当前服务于 Web 2.0 环境下的个人信息资源组织的分众分类法(Folksonomy)工具的兴起与发展,进一步推动了网络组织系统的发展。用户能够对其所发布或使用的资源以标签的形式进行分类标注,以实现对社区内所标注资源的共享及其存取。这种用户自由定义的非正式的社会分类行为不同于原有的受控词表,其采用自下而上的构建方式,灵活随意性大,用户不需要任何学习就可完成分类任务,定义标签简单、直接、迅速。但分众分类法的优点在一定程度上又可能导致一系列问题的产生,如用户选择的标签不能明确地表示所指事物、标签的含义不确定(存在一词多义现象)、标签的语种差异、标签词汇的无层次性等,这无疑将影响到用户检索效率,随着用户人群的增多,更增加了网络资源管理的难度。受控词表是人工从自然语言中选择具有代表性的、有明显含义的词汇建立起来的关于词汇及词汇之间关系的词表,包括分类表、叙词表、术语表等各种类型的词典资源。受控词表在传统的信息组织领域发挥着重要作用,是提供主题存取及检索的主要方法。如果我们应用受控词表的词汇规范性、严格性及其丰富的语义关系等优点,弥补分众分类法的不足,同时挖掘分众分类法中丰富的标签资源及用户数据,以达到及时更新、修改、完善受控词表的目的,使其更符合用户使用的习惯,这样就会达到二者优化的目标。

基于此,我们在调查国内外大量研究文献的基础上,对目前围绕受控词表与分众分类法的结合研究现状作全面综述,旨在推动国内该方面的研究与应用,以综合发挥传统信息组织方法与现代新型信息组织方法的优点,实现对现有中文网络资源的有效管理。

2 研究与应用现状

分众分类法中,用户通过标签对资源进行标注后,使标签、资源和用户之间产生一定的联系,借助这些标签,用户会很容易找到自己感兴趣的用户群与资源集。因此分众分类法深受用户的喜爱,近几年得到迅速发展。

从现有的研究文献看,分众分类法的研究也成为网络知识组织的热点问题之一,主要由商业机构、图书馆及一些国际会议的专门工作组共同推动其发展。正如黄国彬在文章中指出,学术界从只是针对大众标注(Folksonomy)本身开始向将大众标注与特定系统、特定领域相结合转变[3];更多的学者逐渐意识到分众分类法的自由性、无控制性需要采用一定的策略加以限制,分众分类法与受控词表的结合研究及其实践应用也就适时地提上了日程。

通过对大量文献的研读,我们发现分众分类法与受控词表结合的研究基本上分为三方面:一是二者结合的必要性问题;二是二者结合的可能性问题;三是二者如何结合的问题。

2.1 分众分类法与受控词表结合的必要性问题研究

2004 年,Thomas 在与 Smith 的讨论中,针对 Furl、Flickr、Del. icio. us 等网站,提出“Folksonomy”一词,认为其是“folk”和“taxonomy”的结合[4],至此分众分类法的名称被广为接受,分众分类法与受控词表结合的必要性研究也拉开了序幕。研究者发现以用户为中心的分众分类系统尽管在个人信息资源组织方面起到了重要作用,但标签词汇的随意性、与主题的无关性、带有明显个人主观性等问题影响了资源的存取效率。针对分众分类法的缺陷,研究人员一致认为有必要引入受控词表,通过推荐受控词汇来帮助用户提高标签的质量。

Mathes 认为分众分类法这种用户参与的信息组织方式是未来的发展方向,但其不受控的标记系统一定程度上是混乱的,易于导致不准确和二义性,这些局限性恰恰是受控词表的优点[5]。Spiteri 指出分众分类法与受控词表并不对立,利用分众分类法的标签资源可以研究用户行为,补充完善受控词表的词汇[6]。Macgregor 指出分众分类法的标签描述的不准确性等问题源于缺乏受控词表的参与,他预测分众分类法和受控词表将共同存在,发挥其各自的优势[7]。Noruzi 提出用户采用标签标注资源时,面

临着单复数形式、多义、同义、具体化程度的描述问题,最终导致用户检索资源的困难,而基于受控词表的分众分类法是必须的,它可以有效地帮助用户正确地选择标签,借助受控词表的语义关系实现词汇的扩检或缩检,提高标注效率及检索的准确率[8]。

相对于国外的发展,国内理论研究稍微滞后一些,目前仅限于提出分众分类法与受控词表的结合需受到关注,但具体为什么结合、如何结合都缺乏深入讨论。国内学者调查了国外网络知识组织的发展现状,并对研究热点进行了预测[9-12]。王军提出了 Web 2.0 环境下个人信息资源组织服务的语义工具[1],如分众分类法将是一个需要关注和深入研究的问题。赖茂生等人提出受控语言与自然语言的融合、网络环境下传统知识组织工具的改造与应用是该领域研究的前沿和重点,分众分类法在促进用户信息交流和知识共享中起着重要的作用,也逐渐进入研究者的视野[13]。一些学者以分众分类法为研究对象,对其呈现出的特征、信息组织机制进行分析[14],徐少同提出建立受控词表是优化分众分类法的手段之一[15];将分众分类法与书目记录的有机结合也是部分学者关注的问题[16-17]。

2.2 分众分类法与受控词表结合的可能性问题研究

受控词表与分众分类法采用的毕竟是两种完全不同的词汇,受控词汇由专业人员定义和选取,多为名词,而实际调查中发现,标签随意性很大,包含形容词、感叹词、动词、介词等非名词。如果二者要实现互操作,则需要对它们之间词汇的重叠性分布状况进行研究,因此研究者从 2007 年后开始着手于分众分类系统与受控词表的对比研究,以发现其共同点及差异性,寻找它们之间有效的结合点。尽管研究者调查结果有差异,认为二者重叠度有大有小,但普遍认为受控词表与分众分类法可以实现有效互补。

Spiteri 对 Delicious、Furl、Technorati 等网站的标签系统进行了小规模调查,发现用户使用的标签词汇基本遵循受控词表的构建准则[18];Heckner 等人将用户指定标签集与书目元数据进行比较,发现 54% 的标签出现在书目元数据中,其中书目元数据的标题项作为标签的频率最高,占到 49%[19];Rolla 将 LibraryThing 网站的用户标签与 LCSH(美国国会标题表)进行比较,认为用户的标签可以提高对图书馆收藏文献的主题存取效率,但不能取代受控词表,因此允许用户使用标签对书目数据进行标注,在一定层面上能够弥补主题词系统的不足[20];Weber 对 LibraryThing 网站的标签与主题词分布频率进行对比,发现它们各自的词汇具有不同的特征,可以起到互补的作用[21];Mikael 让 20 个用户给书目记录做标签,然后与 LCSH 作比较,发现它们之间重合度并不高,只有 9.14% 完全匹配,因此认为两者之间具有互补性,可以利用标签资源帮助用户检索[22];Kwan 研究了 LCSH 与 Deliciou 分众分类系统的词汇匹配问题,在构建 LCSH 树的基础上,调查 299 个用户标签与 LCSH 树中的 291000 个主题词相比较的情况,发现将近 60.9% 的标签与 LCSH 树中的至少一个主题相关联,为 LCSH 进一步在分众分类法中的应用奠定了基础[23]。

2.3 分众分类法与受控词表结合的问题研究

2007 年后研究者同时从理论和实践层面探讨二者有效集成的问题。UKOLN(英国图书馆和信息网络办公室)领导下的 ENTAG 项目,旨在探讨受控词表与分众分类法的有效集成方法[24],使用户输入的标签能够与 DDC(杜威十进分类法)的类名进行匹配,系统可以向用户推荐与该标签相似的上下位类目等,以实现数字资源的发现;Wartena 提出基于共现频率分布的相似度匹配算法来实现标签与叙词的映射[25]。

除了在理论层面的讨论,实践中关于分众分类法与受控词表结合的研究也在开展,主要有两方面:一是基于受控词表的分众分类法系统,如 Librarything(http://www.librarything.com/home/)作为图书类的标签系统,网上用户很容易使用该系统为自己收藏的图书做目录,并可以找到与自己拥有相同图书的人群。用户标注图书目录时,系统会要求输入题名、ISBN、作者

基本信息,Librarything 将此信息与国会图书馆、亚马逊网站、690 个图书馆的信息相比较,返回精确的书目数据,用户借此修改,标出标签信息。二是嵌入分众分类法的图书馆书目服务系统,如桑德贝公共图书馆网站(http://www.tbpl.ca/)能够显示 Delicious 中与该网站所关联的标签云;纳什维尔公共图书馆(http://www.library.nashville.org/)也与 Delicious 相连,建立了标签与书目记录的链接,并推荐一些与这些标签含义相同的网站链接供用户进一步使用;宾夕法尼亚大学图书馆(http://www.library.upenn.edu/)允许用户给书目数据标注标签信息,以帮助用户追踪资源;安娜堡图书馆(http://www.aadl.org/)开发了社会网络工具 SOPAC 与图书馆的目录集成,允许用户评价、浏览、评论相关标签,并通过标签检索相关书目信息等。

3 特点与不足

从分众分类法与受控词表的结合研究中看出,研究者一致认为分众分类法中的标签集反映了用户的兴趣特点,是以专家为核心的传统受控式信息组织方式向以用户为中心的信息组织方式转变的典型代表,如果运用受控词表对其进行适当的规范,分众分类系统将会在网络信息组织领域发挥更大的作用。

3.1 研究呈现的特点

3.1.1 用户标签的量化分析成为结合研究的基础

分众分类法,无论是综合性网站(如 Delicious、Diigo)还是专业类网站(如 Flickr、豆瓣)都已形成了一定规模的用户群。用户通过标签形式对资源进行标注,这种标注行为很大程度上反映了用户的行为特征。从上述二者结合的必要性和可能性研究中,都必不可少地涉及到用户标签使用规律的统计,其目的在于通过分析这些丰富的标签数据,深层次地挖掘用户、标签和资源间的关系,寻找标签词汇与受控词汇之间的关联度,旨在从用户标注行为中发现一些潜在的信息,如用户所使用的自然语言特点、关注的资源类型与内容、用户的兴趣变化、用户群体特征等。而这些研究结果显然会分别促进分众分类法与受控词表性能的优化,如根据用户的兴趣设计基于受控词表的标签推荐系统、资源推荐系统,建立用户自然语言与受控语言的映射,利用标签词汇完善受控词表的词汇等。

3.1.2 用户高度参与性是结合研究的核心

分众分类系统是用户群体构建的分类系统,用户按照自己的使用习惯来管理个人信息,方便易用性是它优于传统受控式的信息组织方式的最大特点。因此在研究中,围绕用户的参与性,分别从分众分类法和受控词表两个不同角度提出二者结合的解决方案,表现为:一方面研究者更注重于讨论基于受控词表的分众分类系统的优化问题,其前提条件是用户依然可以自由选择标签,受控词表对用户是不可见的,其只起到辅助词表的作用;而另一方面将标签系统的理念引入到图书馆书目系统中,实际上体现传统受控式的信息组织系统开始考虑加入用户的感受及想法,开始注重信息系统的使用尽可能符合用户习惯、反映用户的看法,使用户可以参与到信息管理流程中,达到用户存取资源的便利性目标。

3.1.3 受控词表在网络环境下的发展问题受到关注

受控词表作为专家构建的规范化词表,其在一定程度上减少了一词多义、同义词等自然语言的词义含糊现象,通过有效的词义控制及词间关系构建,提高了主题检索的效率。如何有效地将受控词表的这些优点运用到网络资源组织中,如何利用网络资源进一步完善受控词表,是学者关注的问题,也是受控词表发展的必然。受控词表与分众分类法的结合研究主要涉及两个方面:一是实现受控词表与标签资源的映射,将受控词汇推荐给用户使用,并利用受控词汇扩展检索词汇,提高检索效率;二是运用现有的标签资源实现受控词表的自动构建与更新,并用计算机可理解的形式表达。

3.2 研究中的不足

目前关于分众分类法与受控词表的结合研究还只是处于起步阶段,理论层面还有待进一

步深入:二者如何达到最优化的集成?集成系统与原有的系统相比,性能是否会提高,是否发挥了理想作用?如何实现受控词表、本体、分众分类法之间的集成?如何将二者结合运用到检索实践中?当前以分众分类法为中心的讨论较多,但受控词表如何吸收标签词汇来完善词表将是下一步需要研究的问题。

实践层面,只限于分众分类系统与书目记录的简单结合,而基于受控词表的标签推荐系统还没有成型并推广使用。随着分众分类系统的深入使用,用户规模的增加,标签词汇的无序问题将更显突出,开发基于受控词表的标签系统将会成为必要。另外,书目记录如何利用用户的标签数据进行用户需求分析、构建资源推荐系统也将成为研究重点。

4 发展趋势

随着分众分类法的深入应用,用户数量的稳定上升,一些稳定的网络虚拟社区将会形成,会有不同的用户群体出现,网络服务的功能将根据不同的用户群体更加细化。因此可以预见未来研究与应用的重要方向是:

(1)用户群体细分研究。如何根据标签、资源、用户间的网络关系划分出具有相似性特点的用户,如何针对这些类型的用户拓展延伸专业化网络服务类型。

(2)特定领域的知识表示及组织研究。术语表、受控词表、本体,依然将在信息检索中发挥重要作用。如何有效地从属于特定领域范围的标签、网络资源等这些接近自然语言的语料库中提取相关的概念及知识,以丰富各领域的术语表、受控词表、本体这些组织工具,将具有重要的意义。

(3)针对书目记录的开放性研究与应用。书目记录将不再是封闭在图书馆内部的资源,它将被更多学者关注,这将涉及到网络用户使用的研究、更为有效的知识表示方法与检索方法、如何与其他网络类资源如标签资源建立有效的链接等。

综合以上研究,可以预见当前网络知识组织系统的发展方向:一是用户高效地参与信息组织过程,以使系统更符合用户习惯,易于用户使用;二是信息资源的语义描述,确保计算机从语义层面上理解并解释信息,以提高计算机操作效率,满足用户需求。

参考文献:

[1] 王军,张丽. 网络知识组织系统的研究现状和发展趋势[J]. 中国图书馆学报,2008(1):65-69.

[2] Gail H. Systems of knowledge organization for digital libraries: Beyond traditional authority files [OL]. [2009-11-02]. http://www.diglib.org/pubs/dlf090/dlf090.pdf.

[3] 黄国彬. 大众标注研究进展[J]. 图书情报工作,2008(1):13-15,55.

[4] Gene S. Folksonomy: Social classification[OL]. [2009-07-07]. http://atomiq.org/archives/2004/08/folksonomy_social_classification.html.

[5] Mathes A. Folksonomies: Cooperative classification and communication through shared metadata [OL]. [2009-10-01]. http://www.adammathes.com/academic/computermediated communication/folksonomies.html.

[6] Spiteri L. Controlled vocabularies and folksonomies[OL]. [2009-11-10]. http://www.collectionscanada.ca/obj/014005/f2/014005-05209-e-e.pdf.

[7] Macgregor G, McCulloch E. Collaborative tagging as a knowledge organisation and resource discovery tool[J]. Library Review, 2006, 55(5):291-300.

[8] Noruzi A. Folksonomies: Why do we need controlled vocabulary[J]. Webology, 2007, 4(2):1-7.

[9] 司莉,舒欣. 国外网络知识组织系统研究现状与发展趋势[J]. 图书情报知识,2008(9):82-85.

[10] 柯平. 数字时代的信息与知识组织研究[J]. 中国图书馆学报, 2008(3):67-68.

[11] 王兰成,敖毅,曾琼. 国外知识组织技术研究的现状、实践与热点[J]. 中国图书馆学报, 2008

(2):93 - 97.

[12] 王曰芬,吴鹏. 国外几种典型的知识组织系统及应用[J]. 情报理论与实践, 2008 (2):193 - 197.

[13] 赖茂生,屈鹏,谢静. 知识组织最新研究与实践进展[J]. 图书情报工作,2009(1):19 - 23.

[14] 黄国彬. Tag 信息组织机制研究:以 de.l icio.us、flickr 系统为例[J]. 图书馆杂志,2008(5):45 - 48.

[15] 徐少同. 网络信息自组织视角下的 Folksonomy 优化[J]. 图书情报工作,2009(5):102 - 105,120.

[16] 王翠英. Folksonomy 与主题标引[J]. 情报理论与实践,2007(5):693 - 697.

[17] 吴江. OPAC 与豆瓣融合改进体现 FRBR 的编目模式研究[J]. 图书情报工作,2009(4):43 - 46,58.

[18] Spiteri L F. The structure and form of Folksonomy Tags: The road to the public library Catalog[J]. Information Technology and Libraries, 2007, 26(3):13 - 24.

[19] Heckner M, Muhlbacher S, Wolff C. Tagging: A classification model for user keywords in Scientific Bibliography Management Systems[C/OL]. Proceedings of the 6th European Networked Knowledge Organization Systems (NKOS) Workshop at the 11th ECDL Conference, Budapest, Hungary. [2009-12-01]. www.comp.glam.ac.uk/pages/.

[20] Peter R J. User Tags versus Subject Headings Can User-Supplied Data Improve Subject Access to Library Collections[J]. LRTS, 2009,53(3):171 - 184.

[21] Weber J. Folksonomy and controlled vocabulary in libraryThing [OL]. [2009-12-15]. http://jonathanweber.info/samples/2452 - Folksonomy.pdf.

[22] Wetterstrom M. The Complementarity of Tags and LCSH: A tagging experiment and investigation into added value in a New Zealand Library Context [J]. The New Zealand Library & Information Management Journal, 2008(4): 296 - 310.

[23] Kwan Yi, Lois Mai Chan. Linking folksonomy to Library of Congress subject headings:an exploratory study[J]. Journal of Documentation, 2009,65(6): 872 - 900.

[24] Golub K, Lykke N M, Moon J, et al. Enhancing social tagging with a knowledge organization system [C/OL]. Emerging trends in technology: libraries between Web 2.0, semantic web and search technology, IFLA 2009 Satellite Meeting, Florenc. [2009-11-15]. http://www.slideshare.net/michaelday/enhancing-social-tagging-with-a-kos.

[25] Christian W, Rogier B. Instanced-Based Mapping between Thesauri and Folksonomies[C], The Semantic Web-ISWC 2008, Springer Berlin.

贾君枝 山西大学管理学院教授。通讯地址:山西省太原市坞城路92号。邮编:030006。

(收稿日期:2010-03-01;修回日期:2010-03-07)