网络信息分类法与大众分类法的调查分析比较
——兼谈大众分类法的改进

付玲玲 袁龙 南京农业大学，江苏 南京 210095

摘要：本文主要通过对新浪网和豆瓣网进行调查，分析了其采用的“传统网络信息分类法”与“大众分类法”之间的区别，在此基础上指出了豆瓣网“大众分类法”的不足之处，最后，针对“大众分类法”的改进，提出了自己的建议。

关键词：网络信息分类法；大众分类法；改进

中图分类号：TP393 **文献标识码：**A **文章编号：**1003-9767（2010）05-0081-02

2010年互联网发展报告显示，截止到2009年底，我国互联网网页数量达到336亿个，年增长率超过100%，网民数量达到3.84亿。面对浩如烟海的信息资源，海量的用户，用户如何快速、有效地检索到所需的信息已成为一个亟待解决的问题。在这种情况下，网络分类法应运而生，并且在一定程度上缓解了这个问题。

“Del.icio.us”、“豆瓣”网用户根据自己的习惯和爱好自由标记网络资源，形成了一种新的分类方法——大众分类法，大众分类法就是由网络信息用户自发为某类信息定义一组标签进行描述，并最终根据标签被使用的频次选用高频标签作为该类信息类目的一种为网络信息分类的方法。如在“豆瓣网”中，一个读者把一本JAVA编程方面的书自定义为“计算机语言”，而其他人把这本书定义为“JAVA”，最后系统统计出使用“JAVA”来定义这本书的频率较大，那么“JAVA”就是这本书的大众分类法的表示。

大众分类法是对网络分类的一种新尝试，它突破了传统网络分类法事先规定的类目以及不能由用户自己定义和修改类名的限制。但作为一种新型的、未被广泛采用的网络分类法，它与传统网络分类法相比，究竟有什么优势，存在那些不足，能不能对其进行改进呢？下面笔者就试着分析下它们的区别，并就大众分类法的改进提出一些自己的看法。

1.网络信息分类法与大众分类法的分析比较

1.1 类目等级的不同

网络分类法直接用自然语言的语词组织信息，以事物/主题为中心划分体系，如“新浪网”共分18个基本部类，如新闻、财经、科技等每个基本部类下又分若干一级类目，如新闻下分国内、国际、社会、军事等。

而“豆瓣”网是以词为类的，类目之间是平等的，非等级的，不存在根节点。如豆瓣图书的标签，郭敬明、春上春树、红楼梦等。

1.2 类目产生的途径不同

传统的网络分类法的类目一般是由专业人士事先规定，如搜狐网的娱乐、电影、游戏、图片等十六个大类，都是在网站创建伊始，由专家制定好的。而大众分类法则将“大众”二字发挥的淋漓尽致，它的分类类目是由“大众网络用户”自行制定的，如“豆瓣”网允许用户对该网站提供的图书信息添加标签，对同一本书，不同的用户可能会用不同的标签组进行标记，使用频率高的标签会保留下来作为该书的通用标签，达到一定次数后会显示在标签云图中。

1.3 类目自由度不同

跟传统的网络分类法相比，大众分类法类目自由度较大，它可以由用户自己创建类目，没有约束，用户可以自由地表达自己的思想，有人把这形象地比喻为“像脱缰的野马没有了羁绊，可以自由自在地奔跑”，如豆瓣图书中，有的网友把《红楼梦》定义为“中国古典小说”，有的网友则直接定义为“红楼梦”，自由度很大。

由以上分析可知，大众分类法作为一种新兴的网络分类法跟传统的网络分类法有很大的不同，它更注重广大网友的主动性的发挥，更有利于网友思想的表达、情感的宣泄，但是就目前的情况来看，大众分类法也不可避免地存在着一些不足。

2.大众分类法存在的问题

2.1 用户自行分类时存在错误

大众分类法的类目是由广大没有经过专业训练的网友，自行划分的，抽样显示还是挺容易出错的。例如，在对郭敬明的散文《爱与痛的边缘》进行分类时，豆瓣成员共给它添加了535个标签，其中排在前八位的是“郭敬明(2526)、爱与痛的边缘(1378)、青春(648)、80后(506)、散文(492)、小说(302)、小四(281)、成长(220)”，这里就出现了一个问题，错误的标签“小说”出现的频率较高。

2.2 系统缺乏对标签的控制

在大众分类法中，系统缺乏对标签的控制，虽说一个概念的不同表示方法可以给用户的检索带来方便，但是像“超级女生”与“超女”，“人民大学出版社”与“中国人民大学出版社”，“管理学”和“management”等用来表示同一个意思的词语，从一定程度上增加了网站的负担，使标签数量急剧增加，不利于管理和利用。

2.3 类目的平面结构、数量过多易隐藏信息

类目的平面非等级显示会隐藏重要信息。自由分类不具有等级结构，不存在根结点，标识信息的标签或者是字顺显示，或者是随机罗列在页面上，尽管重要的、点击频次高的标签通过特殊颜色或字体等被突出显示，但是由于标签数量众多，仍难免被浩如烟海的信息所淹没。

2.4 标签的意义可能混淆

不同的网络用户可能会把相同的标签用在不同的地方，从而容易产生混淆。如法国著名短篇小说作家莫泊桑的《项链》，它的标签中出现了“项链”，而如果又有一本是介绍项链的样式、材质等的时尚杂志，网友也为它添加了“项链”这个标签。很明显，前面的“项链”指的是一本小说，而后面的“项链”则更倾向于一件装饰品，这是截然不同的两种事物。而这两种不同的概念硬是被“项链”这个标签混淆到了一起。

2.5 类名（标签）专指度不够

如在“豆瓣”网中，标签的专指度不够的现象俯拾即是。如英国著名作家狄更斯的《双城记》，豆瓣成员给的标签依次是“狄更斯(1003)、外国文学(621)、双城记(477)、小说(370)、英国(301)、外国名著(226)、名著(199)、经典(193)”，很显然，使用频率较高的标签“外国文学”和“小说”都不够专指，最好能使用它们的下位概念，如“英国小说”来提高专指度。

3.改进大众分类法的建议

“非控词表与生俱来的问题导致了大众分类的局限与弱点。”为了使大众标注走向更健康的发展道路，需要采取一定的措施引导。豆瓣网不能仅局限于做表面的统计工作，而且要整合这些标签，针对“豆瓣”标签出现的一些问题，笔者建议应该加强系统对标签的管理机制建设。

(下转第83页)

阻止计算机病毒对磁盘的操作，在网络环境下，计算机病毒传播扩散快，随着计算机技术的发展，计算机病毒变得越来越复杂和高级，因此在网络环境下，防范病毒问题显得尤其重要。它主要包括终端用户防毒、服务器防毒、系统防毒和蠕虫病毒、木马程序、恶意代码、网页病毒、邮件病毒的防范。

（3）主动防御技术

对待网络安全问题，传统的做法是被动防御，而实践表明传统的被动防御已不能解决现今日益凸显的网络安全问题。主动防御具有双重保护与多重检测响应功能以及对攻击行为进行取证等优点。

目前取证技术已成为人们研究与关注的焦点，入侵取证已成为国际上网络安全最重要的课题之一。并且在检测到非法入侵或恶意行为时，利用IDS收集电子证据，是IDS新的应用方向，也是IDS领域的研究热点之一。将计算机取证结合到入侵检测等网络安全工具和网络体系结构中进行动态取证，可使整个取证过程更加系统并具有智能性和实时性。动态取证是计算机取证的发展趋势，依靠IDS，Honeypot，Honeynet的紧密结合，实时获取数据并进行分析，从而获取攻击者的犯罪证据，进而以法律来威慑攻击行为。

3.4数据备份

信息资源管理具体包括对存储器上的数据建立有效的级别、权限，有必要时对数据进行加密；在网络系统中，最贵重的是计算机存储的宝贵数据。建立网络最根本的用途是要更加方便地传递与使用数据，为防止数据丢失，对存放重要信息的存储器备份两份并分两处保管非常有必要。

3.5完善网络安全立法和执法

关于计算机安全的立法，我国先后出台《中华人民共和国计算机信息系统安全保护条例》和《计算机信息网络国际联网安全保护管理办法》等法律法规，但目前，相关的立法还远不能适应形势发展的需要。在技术上我们可以借鉴国外的先进技术，在法律上我们同样可以参照国外的经验，从而确保我国计算机网络健康有序的发展。

做到了立法，强有力的执法也必不可少。对于有关执法人员需要在安全意识、职业道德和事业心、责任心等方面进行相关培训和教育，做到执法程序标准化、规范化和科学化。

4.结束语

我们可以看出，有效地保护我国计算机网络的安全迫在眉捷，并且单纯的网络安全技术和网络产品无法解决网络安全的全部问题，网络安全建设是一个集政策、技术、法律等作用于一体的、长期的、复杂的系统工程。我们不仅要提高网络安全意识，而且要采取必要的安全技术来抵御各种攻击，同时要通过必要的安全技术来对攻击进行监控和分析，进而进行打击。相信在全民的共同努力下，我们的网络环境会变得更好。

参考文献：

[1]朱江.网络安全与防范措施[J].信息技术，2003

[2]肖薇.网络安全现状分析与对策[J].湖北警官学院学报，2003

[3]龙冬阳.网络安全技术与应用[M].华南理工大学出版社,2006.

[4]王建军,李世英.计算机网络安全问题的分析与探讨[J].赤峰学院学报,2009

[5]白兆辉.浅析计算机网络安全防范的几种关键技术[J].科技信息,2009

[6]胡道元、闵京华.网络安全[M].清华大学出版社，2004

[7]周明天、汪文勇.TCP/1P网络原理与技术[M].清华大学出版社，1998

[8]蒋华.我国网络安全现状分析与对策[J].科技进步与对策,2001

(上接第81页)

3.1 由系统收集每本书的标签

3.2 系统整合收集到的标签

a.根据标签的频率来判断一个标签的流行程度。也就是说，如果一个标签出现的频率很大，就会有很多人乐于用它，这里我们将设定一个频率阈值来判定。

b.判断选出的流行标签的描述能力。流行的标签其描述能力不一定好，这里的判断标准主要是：一.对同义词的控制，如出现了“快乐男声”和“快男”这两个标签，就应舍弃其中一个。二.专指度的选择，如出现“小说”和“中国古典小说”，应尽量选择专指度比较高的下为概念“中国古典小说”。三.对可能出现词意混淆的词应加以标记。如上文提到的小说《项链》跟实物“项链”，应分别标注为“小说项链”和“饰品项链”，以免混淆。当然，对一些错误的标签，系统能通过其后台的纠错机制予以改正。

3.3 系统显示整合后的标签

把上述经过系统整合后的标签，按出现的频率大小显示出来，以供读者查看。

尽管现在“大众分类法”还存在一些问题，但我们也欣喜地看到作为一种新的分类方法，它越来越多地受到了人们的喜爱，有越来越多的学者对其进行研究，以期使其更加完善。Samantha、Hemalata Iyer、DianeNeal、Abebe Rorissa和JungWon Yoon等人就如何构建具有高效查找与搜索功能的大众标注系统开展了一系列研究。在不破坏豆瓣网图书网络标注的“大众化”特性的情况下，对其标签进行适当的规范化处理是必要的。正如“世界上没有绝对的自由一样”，我们相信在一定的规范下，“大众分类法”会朝着更好的方向发展的。

参考文献：

[1]羌丽,张学莲等. 图书大众标注评介——以豆瓣网为例[J]. 中国索引,2009(1)

[2]丛鲁丽.基于大众分类法的中文博客分类方法[J].情报杂志,2009(9)

[3] http://tech.qq.com/zt/2010/cnnic25

[4] http://www.sina.com.cn/

[5]http://book.douban.com

[6]孟连生等.标注及其演化研究[J].图书情报工作,2008(52)