·理论探索·

基于标签聚类的电子商务网站分类目录改善研究

张 红 甘利人 薛春香
(南京理工大学经济管理学院, 江苏 南京 210094)

〔摘 要〕本研究针对电子商务网站用户对商品概念认知与网站实际分类目录不匹配, 导致检索效率低下的问题, 提出了基于用户标签的电子商务网站分类目录改善方案, 即将用户标签进行多层聚类, 将聚类结果以层级结构的形式展示, 并实现标签聚类结果和网站分类目录的映射, 从而提高电子商务网站的分类检索效率和分类导航性能。

〔关键词〕网站分类目录; 用户标签; 标签聚类; 标签映射
DOI: 10.3969/j.issn.1008-0821.2012.01.001
〔中图分类号〕G250.7 〔文献标识码〕A 〔文章编号〕1008-0821 (2012) 01-0003-05

Research on the Improvement to Categories of E-commerce Sites Based on Tag Clustering

Zhang Hong Gan Liren Xue Chunxiang
(School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China)

〔**Abstract**〕The study proposed a method to improve the categories of e-commerce site which based on tags to solve the problem of users' concept does not match the actual categories. The method is that making the user tags clustered, then making the clustering results in the form of hierarchy, and mapping it to web site categories to improve e-commerce sites' categories search efficiency and category navigation performance.

〔**Key words**〕web site categories; user tags; tag clustering; tag mapping

南京理工大学信息管理系用户行为课题组 2009 年曾做过一项调查发现: 在电子商务网站中, 有近 80% 的用户倾向使用网站分类目录来查找商品, 但有部分用户通过网站分类目录查找不到指定的商品或用时过长 (超过 3 分钟)。由此可见, 当前电子商务网站的商品分类目录面临着一个突出问题, 即用户对商品的概念认知与网站实际分类架构组织体系的不匹配。其实质是一种用户心智模型与网站分类架构师心智模型差异的体现[1], 这种差异无疑会在很大程度上增加用户对网站分类的认知负荷, 从而降低网站信息传递效率。因此基于用户认知来探索网站分类目录改善的可能途径就变得十分有意义。

目前基于用户认知的网站分类目录改善思路主要有以下两个方向: 一是依据用户认知改善并调整网站现有分类目录, 比如对产品进行多重归属。但是用户需求总是处在不断变化中, 直接依据用户需求调整网站分类目录会为网站后台分类目录动态调整带来很大的压力和工作量。因此, 有学者提出第二条思路, 即直接按照用户认知来构建“商品分类目录”。用户标签就是当下在网络环境中用户认知和用户参与的一个重要体现, 这也是本研究采用的主要思路。

1 研究背景

1.1 相关概念

早在 1998 年美国人约舒亚·沙科特 (Joshua Schachter) 就提出了用户标签 (Tag) 这一概念。为方便检索和信息管

收稿日期: 2011-09-26
基金项目: 本文系国家社会科学基金资助项目“网站信息传递中符号表证的认知活动探索” (08BTQ036) (2009-2011) 的研究成果之一。
作者简介: 张 红 (1988-), 女, 硕士研究生, 研究方向: 信息行为用户研究。
甘利人 (1957-), 女, 教授, 博导, 研究方向: 网络信息资源管理。
薛春香 (1979-), 女, 副教授, 研究方向: 信息智能处理、知识组织系统等。

理，由网络信息的提供者或者用户自发为某类信息赋予一定数量的标识，这种标识就称为用户标签[2]。它显著的特点就是用户可根据自己的认知、理解与想法，以自由词汇作为标签对资源进行组织和利用[3]。

伴随用户标签发展的是一种新型的网站信息组织方式——folksonomy（公众分类法）。它的基本思想是：根据标签被使用的频次，选用高频标签作为该类信息类名的一种网络信息分类方法。与一般分类方法不同的是，它向社群参与者提供一种协同构建与共享各自网络资源标签的开放式平台，通过用户自身制定分类标准和提交标签来实现[4]。但由于标签是由不同用户根据自己的理解提出的，因此随意性大，与网站一般分类目录相比科学性明显不足。所以用户标签与网站一般分类目录的关系不应该是相互替代或是并行，在后期探讨网站分类目录改善方案中，可以在充分考虑用户心智模型的基础上兼顾网站建设的科学性，将两者予以整合。

1.2 基于用户标签的电子商务网站分类改善研究现状

目前，一些电子商务网站开始采用公众分类法为用户提供标签服务。本研究在对著名电子商务网站——亚马逊和淘宝网的考察中发现：用户标签在电子商务网站中的主要功能是通过标签云图的形式为用户提供商品推荐，同时方便用户查找其他具有相同特性的商品，并对自己感兴趣的商品进行组织。在这些网站中，标签云图与网站一般分类目录形成了两大并行体系，其目的都是为了方便用户检索相关产品。

可以说与电子商务网站一般分类目录相比，网站使用标签云图为用户展示热门商品已经完全考虑到了用户的心智模型，且标签管理也已相当成熟，但是当下的标签云图仍存在很多的问题：（1）用户标签所组成的类目是非等级平面结构，难以揭示信息之间复杂的关系。（2）缺乏对语义尤其是同义词的控制。（3）通过标签云图检索到的产品不是五花八门就是不够全面。

对此，国内外学者提出了一些改进措施，试图改善电子商务网站中标签云图存在的不足。Heymann P 等人提出将大量的标签转化为可导航的层次结构的分类目。将标签按其所标注的资源的次数表示成向量的形式，同时用余弦相似性计算得到标签的相似图，最后得到潜在层级的分类法[5]。国内也有学者提出可以通过标签聚类技术，将标签进行层级处理。西安电子科技大学的窦永香等利用著名的 Porter 算法对英文标签进行词根提取，然后根据用户的精确度要求对相关标签进行聚类[6]。广东商学院的王翠英在对标签进行共现分析的基础上，提出基于共现信息的标签聚类算法[7]。此外，武汉大学的曹高辉等提出通过凝聚式层次聚类算法，利用相关标签的权重，计算标签之间的相关度，从而实现标签聚类[8]。

对于目前基于用户标签的电子商务网站改善方案，研究大多集中在标签聚类的问题上，由于在实际复杂的电子商务网站中，无论是用户、标签还是资源都是海量的，这导致了用户标签的随意性和不科学性，也给基于用户标签的电子商务网站分类目录改善从理论走向实际应用带来了许多困难。此外，对于用户而言仅依靠用户标签和标签云图是无法满足网站分类搜索这一需求的，而当前的改善思路很少考虑到将网站一般分类目录与标签云图整合起来研究。

本研究尝试利用网站现有分类目录的科学性，同时考虑用户的心智模型，在用户标签聚类的基础上，将网站一般分类目录与用户标签系统两者进行有机结合，从而改善电子商务网站分类目录的现状。

2 基于用户标签的电子商务网站分类目录改善方案设计

本研究拟采用如下方法来改善电子商务网站分类目录：对基于用户认知所提出的标签进行聚类，形成具有层级关系可导航的标签云图，同时在网站现有分类目录与具有层级关系的标签云图之间建立映射，使用户可以直接依据用户标签云图实现商品的分类搜索。整个系统实现思路包括：标签预处理、标签聚类、标签与网站分类映射 3 个部分，如图 1 所示。

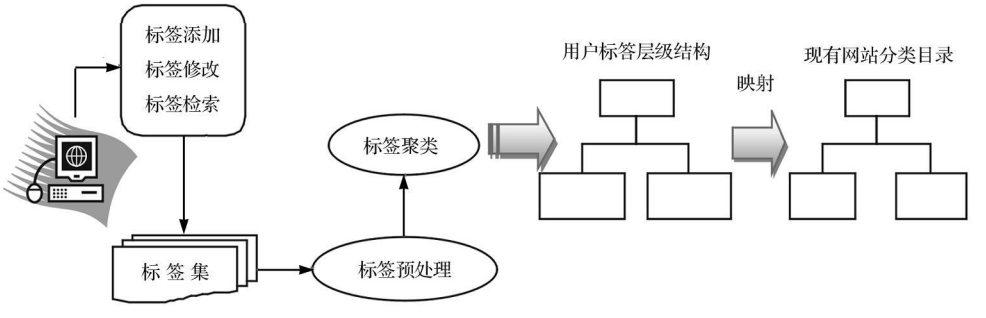

图 1 基于用户心智模型分析的网站分类改善方案思路图

2.1 标签预处理

主要目的是通过构建同义词表来达到同义词控制。

2.2 标签聚类

通过对用户标签同义词的控制，我们提出了对用户标签进行层级聚类的构想。标签聚类基本思想是通过对用户标签数据的词频统计以及共现分析（与 tag_i 共现次数最多的 tag_j 被认为与 tag_i 强相关），将用户标签聚类成一个符合用户个人认知习惯的商品分类体系。根据该思想，标签的聚类过程如图 2 所示:

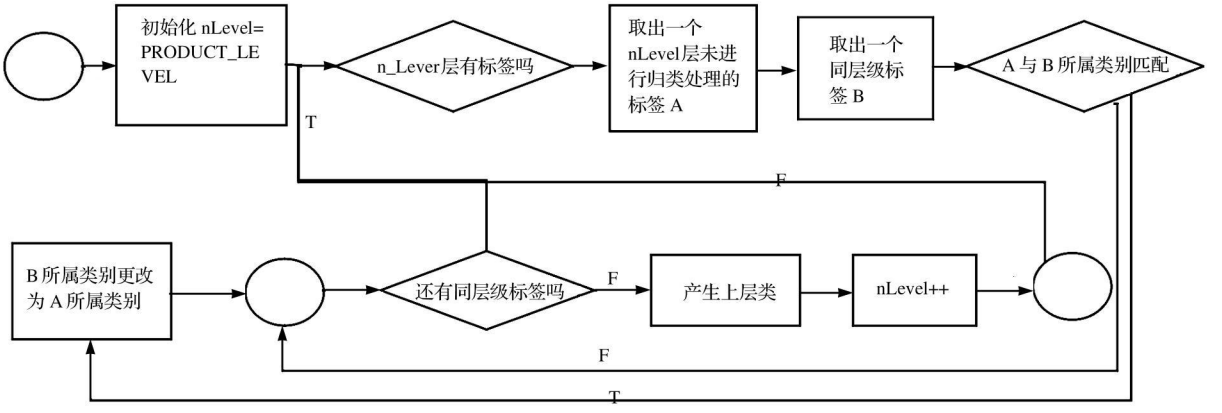

图 **2** 标签聚类流程图

2.3 标签映射

主要目的是将用户标签聚类结果与网站现有分类目录之间建立映射关系，通过点击标签云图上的用户标签能够迅速定位到网站现有分类目录相应类别上。具体用户标签映射实现思路如图 3 所示。

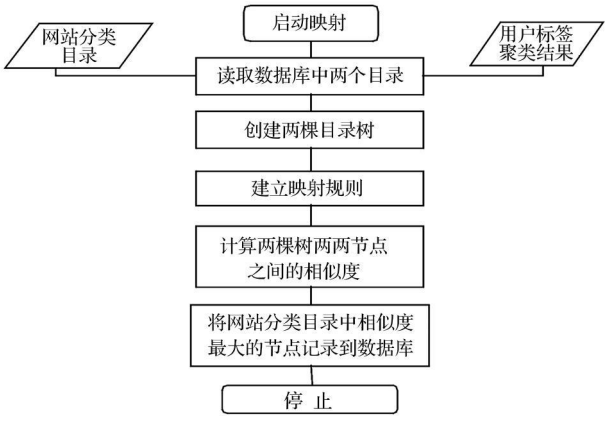

图**3** 用户标签映射网站分类目录算法

3 实验验证与系统实现

在上文论述的基于用户标签电子商务网站分类改善总体方案设计的基础上，本研究还模拟了实际用户对商品添加标签的情景，进行实验探索。

3.1 实验设计

由于不同用户背景、知识、经验各不相同，导致添加标签的结果存在一定差异。因此本研究选取了经管院和计算机院大三和大四 2 个年级共 188 名学生参与我们的实验。由被试对本实验中提出的 4 种商品（本实验主要指定了电子词典和 U 盘类目下的 4 种具体产品，分别是“诺亚舟 NH6080 电子词典”、“名人牛津搜索王”、“金士顿 U 盘”、“忆捷优盘 U5”）进行标注，分别提交 3 个标签。

实验共回收有效问卷 185 份，涉及到的标签概念 355 个。将用户提交的标签输入到专门为本实验模拟建立的电子商务网站用户标签平台上。如图 4 所示:

图 **4** 电子商务网站用户标签添加实验平台

3.2 用户标签预处理实验探索

本研究从实验标签集合中随机抽取 142 个概念作为建立同义词表的数据集。按照一定的同义词表构建依据，手工构建同义词表，并选用同义词组中使用频次较高的词作为标准词。然后通过设计计算机程序利用字面匹配和字面相似度计算的方法，将用户标签与同义词表中的词进行匹配，并用标准词对该标签进行表征。

3.3 用户标签聚类实验探索

在用户标签预处理的基础上，本研究通过计算机编程尝试实现用户标签的层级聚类。具体标签聚类步骤如下:

3.3.1 将每个商品下的标签进行聚类，取出使用频次最高的标签作为初始的聚类中心

例如产品“名人牛津搜索王”的所有标签中，“名人”的使用频次最高，“名人”就是该商品的聚类中心。

3.3.2 将从属于每个聚类中心点的最底层标签进行两两相似度判断

相似度判断依据有两点：首先，根据两聚类中心的最底层标签的字面匹配度来确定聚类中心是否相似。其次，根据最底层相似的个数，如果两聚类中心下相似标签的个数达到一定的阈值，那么这两聚类中心所代表的商品即为同类商品，它们会有一个共同的上层目录（父目录）。

3.3.3 上层目录（父目录）的确定

我们结合两个方面来确定上层目录：该标签在同类商品中出现的概率，以及在每个商品中出现的频次。对于同类商品，标签 A 都被标注或标注的概率很大，且出现的频次非常高，通过权重计算，我们可以判定 A 是该同类商品的上层目录。例如，对于“诺亚舟 NH6080 电子词典”以及它的同类商品“商品 A“商品 B”“商品 C”……来说，标签“电子产品”“电子词典”在上述 4 中产品中出现的频率非常高（分别为 75%、100%），且使用频次也比较大（分别为 65 次、70 次）。经过权重计算，最终结果是电子词典 > 电子产品，那么电子词典就是该同类商品的共同上层目录。按该方法继续由下往上聚类，即可形成多层类目体系。

按照上述算法步骤，最终程序实现聚类效果如图 5 所示。

图 5 用户标签聚类结果界面

3.4 用户标签映射实验探索

按照上节标签映射的基本思想，我们按照一定的映射规则将用户标签聚类结果与网站现有分类目录之间建立了映射关系。实验中具体实现步骤是：

3.4.1 建立标签树

根据网站自身分类目录和标签聚类结果分别建立网站分类目录树（如图 6）和聚类标签树（如图 7）。其中聚类结果将以具有层级结构树状结构（仅显示两层结构）在云图中展现。

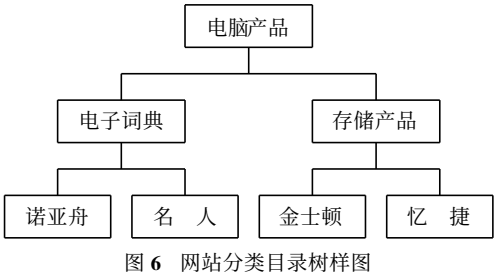

图 6 网站分类目录树样图

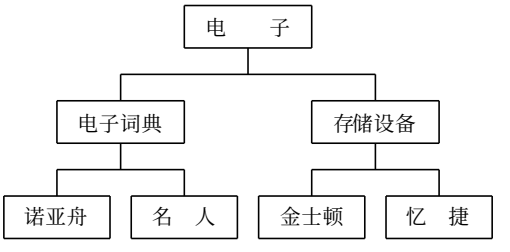

图 7 聚类结果标签树样图

3.4.2 建立映射规则

我们将两个树中的每一个目录标签作为一个实体，建立了 4 条映射规则。①如果描述两个数据项语义的语义树（就是它所处的目录列别的层级以及其子孙节点，兄弟节点，父亲节点）完全相同，则两个数据项语义相等，可直接映射，即实体间的一对一映射（如我们实验网站用户标签层级结构下的“电子词典”到网站分类目录下“电子词典”的映射）。②标签通常被人们用来作为实体的惟一标识

(名字)，因此若待比较的两个实体的标签相等，则认为两实体相等。③同样地，若两个待比较实体拥有相同的 URI (即层级目录中所指的相对应的商品展示页面相同)，则认为两实体相等。④拥有相同实例的两个实体，被认为相等。

3.4.3 标签映射

在聚类效果达到比较好的基础上，参照上述映射规则，我们分别采用字面匹配的方法计算两棵树中的各节点 (父节点、子孙节点) 的相似程度，相似度最大的作为其在另一棵树中的映射节点，例如：当用户在用户标签分类目录中选择“U 盘”这一标签时，系统通过对标签分类体系中“U 盘”目录的父节点和子孙节点标签所对应的具体商品进行统计，发现这些商品在网站传统分类目录中属于“电脑产品”目录下的“存储产品”目录，那么就可将用户标签分类目录中“U 盘”与网站分类体系中“存储产品”目录相映射。当用户点击标签“U 盘”时，其返回的结果为网站分类体系中“存储产品”目录下的产品。

本研究最终希望达到的效果是：实验建立一个电子商务用户标签平台，实现标注功能，标签聚类结果以层级结构 (两层) 的形式作为用户构建的“商品目录”展示在标签云图上，并实现标签聚类结果和网站一般分类目录的映射，从而改善电子商务网站检索效果。最终程序实现映射效果如图 8:

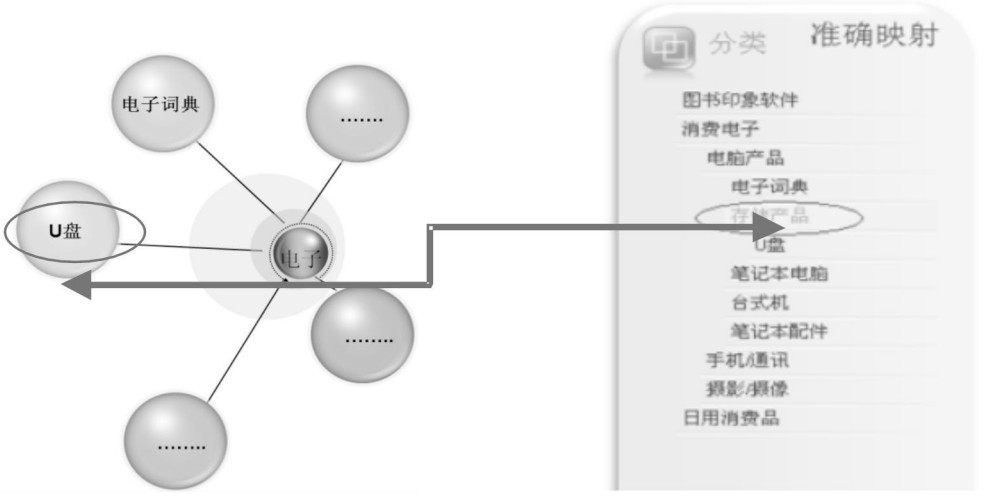

图 8 用户标签与网站分类目录映射效果图

4 结 语

本研究提出了一种基于用户标签聚类的电子商务网站分类目录改善方案，并通过实验验证了在电子商务平台上该方案的可行性。但是由于时间和人力有限，本研究在实验验证中只选取了 4 种产品、3 个层级、2 类产品，这与电子商务网站实际情况还存在很大差距。尤其是在真实情境下，面对大规模用户标签以及成千上万种商品时用户标签如何进行更好的语义控制、产品边界概念如何界定、标签云图展示哪些标签及如何合理的分布等问题还有待进一步研究。

参 考 文 献

[1] 朱晶晶. 电子商务网站分类体系理解的用户心智模型研究 [D]. 南京理工大学，2010.

[2] Thomas Vander Wal. Folksonomy Explanations [EB/OL]. http://www.vanderwal.Net/random/entrysel.php?blog=1622，2006-11-02.

[3] 乐庆玲. 基于协同机制的 Tag 资源自动分类研究 [J]. 现代图书情报技术，2007，155 (9)：58-61.

[4] 周荣庭，郑彬. 公众分类：网络时代的新型信息分类方法 [J]. 现代图书情报技术，2006，(3)：72-75.

[5] Heymann P，Garcia-MolinayH. Collaborative creation of communal hierarchical taxonomies in social Tagging systems [R]. Technical Report Info-Lab. Department of Computer Science，Stanford：StanfordUniversity，2006.

[6] 窦永香，苏山佳，赵捧未. 基于 Porter 算法的英文标签聚类方法研究 [J]. 现代图书情报技术，2009，(9)：40-44.

[7] 王翠英 (编译). 标签的聚类分析研究 [J]. 现代图书情报技术，2008，(5)：67-71.

[8] 曹高辉，焦玉英，成全. 基于凝聚式层次聚类算法的标签聚类研究 [J]. 现代图书情报技术，2008，(4)：67-71.