IFLA WLIC 2013
SINGAPORE

Submitted on: July 22, 2013

基于用户术语的 Web 资源分类

Translation of the original paper "Classification of Web Resources using User Generated Terms"

Margaret E.I. Kipp
School of information studies, University of Wisconsin-Milwaukee, Milwaukee, United States.
E-mail address： kipp@uwm.edu

Soohyung Joo
School of information studies, University of Wisconsin-Milwaukee, Milwaukee, United States.
E-mail address： sjoo@uwm.edu

Inkyung Choi
School of information studies, University of Wisconsin-Milwaukee, Milwaukee, United States.
E-mail address： ichoi@uwm.edu

Translated by <王兴兰，Xinglan Wang>, <中国科学院国家科学图书馆，National Science Library, Chinese Academy of Sciences>, <中国，China >

摘要:

在本文的研究中，我们提出了根据用户生成的社会标签信息进行 Web 资源分类的方法。我们试图研究在某个领域里，社会标签是否能成为对网站进行分类的工具。为此，我们将主成分分析法（PCA）和层次聚类法两种统计方法应用到消费者健康信息领域的网站分类中。首先，我们使用 PCA 方法识别所选网站的不同维度。使用 PCA 方法从网站中提取六个维度：女性、老人、儿童/育儿、药物、男性、研究。然后，我们使用层次聚类法在不同的层级对相似的网站分组。以上两种方法揭示了社会标签能够很好地表达健康信息领域的个人网站的特征。这个研究为使用社会标签进行 Web 资源自动分类提供了理论依据。

关键词: 网页源组织，分类，社会标签，自定义术语，数据挖掘，主成分分析，层次聚类

1 引言

随着 Web 资源的爆炸式增长，如何组织 Web 资源已经成为信息专家和研究者的一个重要话题。与传统的印本材料或电子期刊不同，大多数 Web 资源没有使用传统的组织方式。尽管对信息资源组织模型的开发得到了研究者的共同努力，但是目前还没有广泛适用于在线资源的标准元数据。另外，在线资源较传统资源增长速度快，对在线资源进行手工分类几乎是不可能的。虽然在线资源受到上述两方面的限制，但是 Web 资源组织仍有需求，并且有利于用户提高有效地访问网页信息的能力。网络资源分类能够更好地支持用户的浏览策略，帮助用户扩展对相关文献的搜索兴趣(Xie and Joo, 2012)。研究者已经意识到 Web 资源组织的重要性，他们试图寻找有效地 Web 资源自动分类方法，以减少网络信息的复杂性并提高网络信息的可访问性。其中，基于全文的机器学习分类是最常见的方法。然而，据我们所知，在 Web 资源分类系统中很少使用用户的术语。

社会标引已经成为一个热门话题，研究者不断研究它的特征和模式。使用社会标签服务来补充现有的元数据已经被多数图书馆和信息服务机构接受。在论文中，我们期待社会标签在组织 Web 资源中发挥另一个实践意义。本文探讨用户自定义术语能否在 Web 资源分类中发挥作用，是否能够成为基于全文的在线资源聚类方法的替代方案。由于社会标签包含被描述的 Web 资源的关键词，本文假定标签有得于实现 Web 资源分类。本文以消费者健康信息领域为例，使用主成分分析法（PCA）和层次聚类法检验社会标签用于分类的可能性。对社会标引的研究包括不同的方法，但只有少数研究涉及降维方法或层次聚类技术。本文通过标签的相似度实现 Web 资源分类。同时，本文使用主成分分析法（PCA）和层次聚类两种数据挖掘技术实现在线健康信息领域中基于用户标签的 Web 资源聚类。

2 研究综述

此前，研究者致力于对基于关键词分析的 Web 资源自动分类方法的研究。。为了实现机器网页信息组织(Ricca et al.2004; 2007; Tonella et al. 2003)，研究者不遗余力地从在线文档中提取主题词或是元数据。这些方法有助于随机分散的 Web 资源成为有组织的模式。然而，这些方法主要依赖于文档自身的内容，并未从用户的角度对资源进行解释。同时，对 Web 资源分类的全文分析需要系统负载。再者，由于元数据存在于特定的领域，使得元数据的覆盖范围有限，导致基于元数据的分类应用范围有限。

从社会标签中挑选出的用户自定义术语能够成为全为标引和基于元数据分类方法的替代方案。社会标引作为一种 Web 资源组织手段受到研究者的关注，研究者探究在

不同的领域或环境下用户标签的实践意义。例如，Kipp 和 Campbell (2007)发现标签法与传统标引在某些方面的一致性，创建一个额外的个人维度来表达时间和任务术语。大量研究比较标签和受控词汇在标引中的角色(Yoon 2009; Kipp 2005; 2011)。少数研究探索标签在 Web 资源标引中的独特性。Cattuto et al. (2008)等使用网络分析法监测社会标签在网上书签系统中同时出现的次数。为了建立资源的加权网络和语义关系，他们介绍了基于用户集体标签行为的“资源距离”概念。Kipp 和 Joo (2010) 使用结构方程建模，分析了基于标签模式的网页空间的语义结构。然而，极少研究者尝试利用社会标签进行 Web 资源分类。过去的研究着眼于对标签实践能力的描述，忽略了 Web 资源分类中标签的实际应用。

3 方法

为了检测用户自定义术语是否能够应用于 Web 资源分类，本文以消费者健康信息领域的网页为测试样本。本文采集了 CAPHIS(消费者和患者健康信息科)推荐的 34 个消费者健康信息网站，将其作为数据集。本文从 Delicious.com 中选择标签，被选择的网站至少包含 50 个标签。CAPHIS 提供了一系列健康信息网站列表，并且这些网站分类到不同的群组，如女性、男性、老人、儿童&育儿以及药物。手工创建的 CAPHIS 分类模式用于评估基于标签的分类的能力。采集的数据去除通用术语和专业术语后，得到 6416 个不同的术语。本文选择至少同时在 5 个网站中出现的术语共 654 个，排除 79 个在网站中频繁出现的通用术语，将剩余 575 个术语用于下一步的分析。

首先，主成分分析法用于识别被选领域的 Web 资源结构。主成分分析法常用于探索数据集的结构实质，通常用于发现任何数据集的维度，以表达特定领域的结构(Jackson 1991)。它将包含变量的大集合压缩成一个小集合，但尽量保留原来集合的信息。降维后的集合更易于分析和解释数据集的结构。为了确保基于标签的分类的可能性，本文首先探索被选择的 Web 资源的维度。

然后，层次聚类法被用于检测基于用户术语的分类。这部分内容是本研究的主要目的。这里使用 Ward 学者的关联方式。它是一种凝聚（至下而上）层次聚类法，通常用于检测基于实体相似性的文本聚类的整体相似度。同时，Ward 的方法有利于在互斥的子集中形成等级群组。

通过这两种定量分析方法的使用，我们将基于用户术语的 Web 资源分类法结果与之前的 CAPHIS 分类法的结果进行比较，得到分类的准确率。在每一个网站中，我们增加了一个前缀：w_ (女性健康); s_ (老人健康); k_ (儿童健康); d_ (药物信息); m_ (男性健康); and r_(相关信息研究).

4 实验结果

实验结果显示用户术语有助于网页在线资源的正确分类。PCA 方法和层次聚类法能够像 CAPHIS 分类中专家构建的模式一样对目标网站进行分类。

PCA 的结果显示标签信息能够表示在某个特定领域 Web 资源的结构。在特征值为 1.87 时，PCA 能够识别六个维度：女性、老人、儿童/育儿、药物、男性、研究，这六个维度占总方差的 61.12%。另外，大部分网站的中度或高度因子载荷超过 0.45.这个结果提示了社会标签能够阐述 Web 资源的结构组织。

Table 1: 旋转后的成份矩阵 (方差最大旋转) Rotated component matrix (Varimax rotation)

	Dimension					
	1	2	3	4	5	6
w_Our_Bodies_Ourselves	.819					
w_healthywomen	.814					
w_NLM_Womens_Health	.798					
w_4woman.gov	.735					
w_fwhc	.706					
w_feminist	.679					
w_WebMD_women	.641					
w_menopause	.451					
s_healthinaging		.889				
s_cdc_aging		.887				
s_nihseniorhealth		.872				
s_firstgov		.716				
s_agingcare		.592				
s_aarp		.513		.445		
s_gmhfonline		.506				
s_medicare		.403		.339		
k_aap			.906			
k_dr_greene			.864			
k_kidshealth			.795			
k_virtual_pediatric			.667			
k_nichd			.616			.358
k_aacap			.528			
k_whattoexpect			.483			
d_nlm				.890		
d_rxlist				.887		
d_pdrhealth				.866		
d_needymeds				.706		
m_menshealth					.923	
m_mensfitness					.904	
m_WebMD_men					.795	
m_NLM_Mens_Health					.486	
r_plos						.894
r_biomedcentral						.843
r_entrez						.803

本文使用层次聚类法测试协同标引在 Web 资源组织中的实际作用。在聚类研究中，使用 Ward 的关联方法和 Minkowski 的测试方法。最后使用树状图来解释聚类结果。从图示可知，在第 13 级聚类层，识别出六个群组：男性，药物，研究，儿童&育儿，老人，女性。同时，树状图说明具有相似特征的网站所在位置相邻。这个结果提示了用户标引可以用于 Web 资源自动组织。

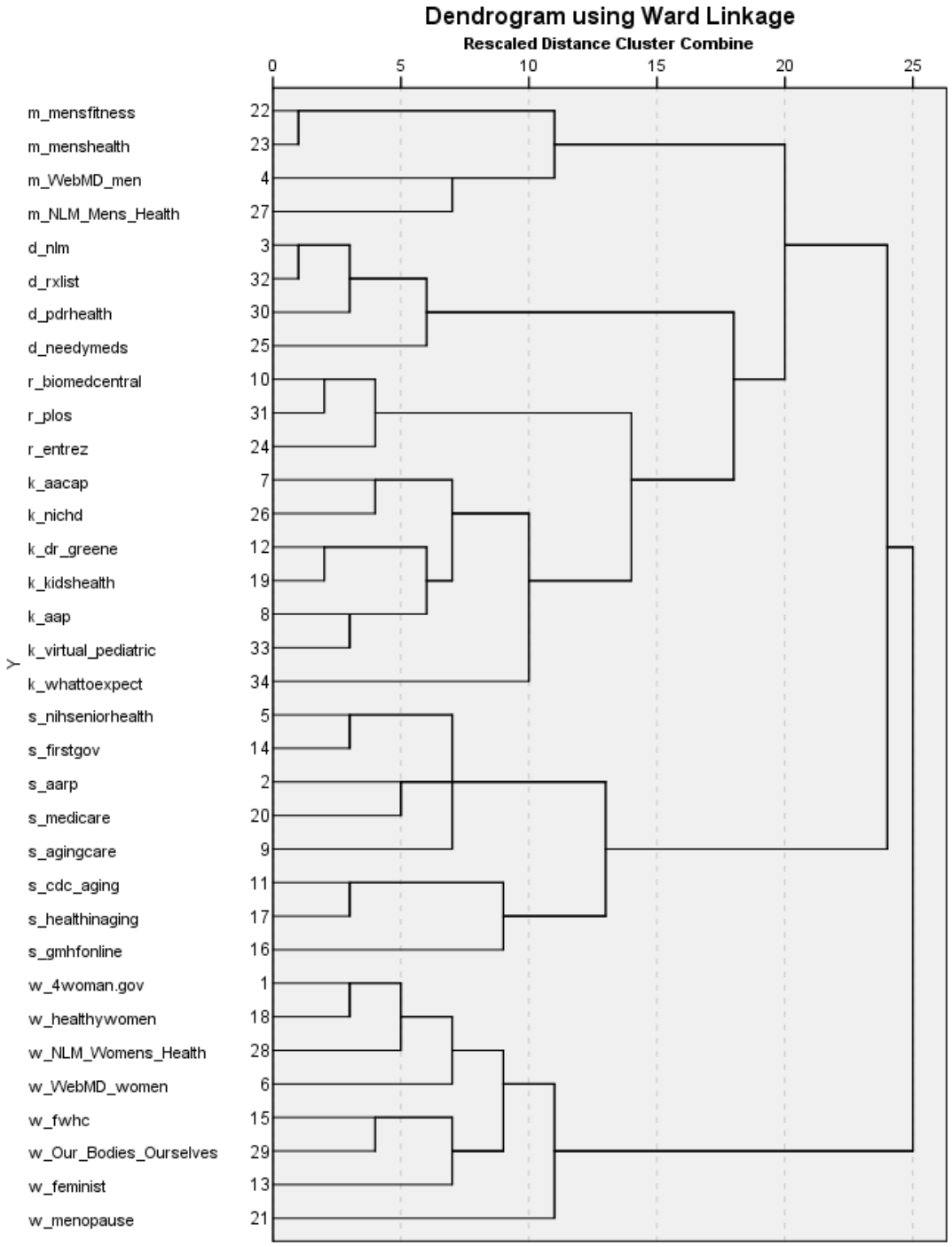

Figure 1: 层次聚类分析结果树状图

5 讨论/结论

本研究探讨了用户自定义术语中能否用于 Web 资源分类。我们从消费者健康信息网站中采集用户的术语，并用层次聚类法对其分类。然后将聚类结果与手工创建的分类模式进行比较，我们发现基于标签的聚类结果显示 100%精确的分类性能。尽管这个研究使用特定的数据集，这一发现仍说明用户标签有助于实现 Web 资源的自动组织。

PCA 分析法揭示了标签信息可以清晰、精确地解释 Web 资源的隐藏结构。根据本文的分析可知，在合适的定量方法的辅助下选择的标签信息能够生成一个合理的聚类。层次聚类法说明使用用户标签能够将 Web 资源准确地分类到人工分类模型预定义的六个群组（“女性”，“老人”，“儿童/育儿”，“药物”，“男性”，“研究”）。从实验结果来看，由上述两种方法得到的结果与 CAPHIS 建议的网站分类得到的结果相似。因此，用户使用标签对 Web 资源进行描述的行为对分类是有意义的。

网页信息组织是图书情报领域的一个挑战，从这篇论文可以看出用户自定义术语可以用于 Web 资源组织。尽管本研究使用的是特定领域的网站，但可以将本研究拓展到其他类型的 Web 资源，如书本、网页文档、图像以及其他多媒体资源。例如，使用用户术语可以将来自 Librarything(www.librarything.com/)的资源进行分类。采集 Flickr (www.flickr.com/)中的标签，根据相似度原理可对图像文档分组。本文使用两种数据挖掘方法，试着探索用户标签在组织特定领域的 Web 资源中的作用。为了扩大它的实用性和探索应用到组织工具中的标签，下一步研究的目的是在术语表的辅助下探索用户标签组织 Web 资源的可能性。在机器学习的分类方法中，需要大量数据、训练过程及成本，而社会标签作为一种组织工具为 Web 资源分类提供了一个更简单和容易的方式。

本研究中介绍的两种数据挖掘方法适用于特定领域的 Web 资源分类。PCA 方法和层次聚类法都验证了社会标签可以做为组织 Web 资源的一种工具。初步研究发现用户自定义术语可用于 Web 资源的自动分类。我们研究说明大量的统计方法可用到生成基于标签的 Web 资源聚类。这些方法与机器学习方法相比，所做的分析和训练过程较少。

参考文献

Jackson, J. E. (1991).*A user's guide to principal components* (Vol. 244). Wiley-Interscience.

Kipp, M. E. I. (2005). Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator, and Intermediary Keywords. *Canadian Journal of Information and Library Science, 29*(4):419–436.

Kipp, M. E. I. (2011). Tagging of Biomedical Articles on CiteULike: A Comparison of User, Author and Professional Indexing. *Knowledge Organization 38*(3): 245-261.

Kipp, M. E.I., & Campbell, D. G. (2007). Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. *Proceedings of the American Society for Information Science and Technology,43*(1): 1-18.

Kipp, M. E. I. and Joo, S. (2010). Application of structural equation modelling in exploring tag patterns: A pilot study. *Proceedings of the American Society for Information Science and Technology*, *47*: 1–2. doi: 10.1002/meet.14504701325.

Ricca, F., Tonella, P., Girardi, C., & Pianta, E. (2004). An empirical study on keyword-based web site clustering. In *Program Comprehension, 2004. Proceedings. 12th IEEE International Workshop on*, 204-213.

Ricca, F., Pianta, E., Tonella, P., & Girardi, C. (2008). Improving Web site understanding with keyword-based clustering. *Journal of Software Maintenance and Evolution: Research and Practice*, *20*(1): 1-29.

Tonella, P., Ricca, F., Pianta, E., & Girardi, C. (2003, September). Using keyword extraction for web site clustering. In *Web Site Evolution, 2003. Theme: Architecture. Proceedings. Fifth IEEE International Workshop on*, 41-48.

Xie, I. and Joo, S. (2012). Factors affecting the selection of search tactics: Tasks, knowledge, process, and systems. *Information Processing & Management*, 48(2): 254-270.

Yoon, J. (2009). Towards a user-oriented thesaurus for non-domain-specific image collections. *Information Processing & Management 45*(4): 452-468.