微内容序化方法与应用实例*

张　乾　蔡淑琴　石双元

华中科技大学管理学院　武汉 430074

〔摘要〕以 Web2.0 技术产生的微内容杂乱无序、难以利用的问题为出发点，应用信息组织理论和序化思想，研究如何对微内容进行汇总、序化加工，形成有序的、易于理解与利用的综合信息，包括：设计对微内容加工的方法，构建 Web2.0 信息加工框架，为 Web2.0 网站建设及改进提供依据，并探索微内容信息利用的方式。

〔关键词〕微内容　序化　指标　信息加工

〔分类号〕G203

Ordering Methods and Its Application of Microcontents

Zhang Qian　Cai Shuqin　Shi Shuangyuan

School of Management, Huazhong University of Science and Technology, Wuhan 430074

〔**Abstract**〕The background of this paper is the problem that the microcontents are lack of order and difficult to use. Using information organization theory and ordering theory, this paper analyzes the way of collecting microcontents and ordering procession, then produces the ordered result and combination of information. Finally an information procession frame is build up in order to give advices on building and improving web2.0 websites, as well as the way of using microcontents.

〔**Keywords**〕microcontents　ordering　indicators　information process

Web2.0 技术使普通网民成为信息的接收者和信息的提供者。但是，广大网民提供的 Web2.0 信息（简称微内容）呈现无序化、去中心化的特点。对网民来说，序化的微内容是更有意义、更有利用价值的，因此微内容序化和中心化成为新的研究问题。

Cmswiki 将微内容定义为："微内容包括个人所形成的任何数据：比如一个简单的链接、一篇网文、一幅图画、一段音频、视频、收藏的书签、喜爱的音乐列表，等等"。熊回香认为，微内容是指在网络上至少拥有一个唯一地址或编号，以及只含有极少数中心概念的元数据和元数据的有限汇集[1]。

序化即增加系统的有序度，是指系统的所有组成元素按照特定的逻辑法则进行顺序排列的过程。张如法认为，有序化是指发现事物或现象之间的各种联系，而将它们作一定的排序和连接[2]。信息序化，又称为信息整序，和信息组织的概念密不可分。吴华欣、于雄杰指出，社会信息的生产和流通具有无序性，主要特点表现为信息从局部上看是有目的、有计划的，但从整体上看则不然。这种无序性对信息的利用造成了极大的障碍[3]。王松林认为，信息序化就是通过对信息外在特征和内容特征的表征和序化，实现无序信息流向有序信息流的转换，从而保证用户对信息的有效获取和利用及信息的有效流通和组合[4]。

在应用方面，如何挖掘微内容的商业价值也逐渐成为研究的热点。Chen 和 Xie 认为，微内容是由大众知识导向的，对其进行加工而得到的信息产品，对企业有重大的商业价值[5]。当无序、杂乱的微内容被加工后，Tam 和 Ho 认为，这些信息产品就可以极大地影响普通网民和企业的认知和决策[6]。

1　微内容的分类

根据微内容的表现形式的差异，将传统的微内容概念进行扩充，分为信息微内容和用户微内容两类。信息微内容即传统意义的微内容，包括用户创造的记录、对其他微内容的评论等。同时本文将用户微内容定义为：在网络中，标识每一个用户的多个角度的元数据集合。

* 本文系国家自然科学基金项目"微内容生产加工模式及其支持平台的研究"（项目编号：71071066）和教育部人文社会科学基金项目"基于互联网信息的企业危机事件识别研究"（项目编号：11YJA630098）研究成果之一。

收稿日期：2011－07－01　修回日期：2011－10－12　本文起止页码：38－41　本文责任编辑：易　飞

“用户”指在Web2.0网站中发布过微内容的用户(简称信息提供者),区别于普通网民(个体序化信息接受者,简称个体接受者)中不提供信息的用户,例如只在Web2.0网站浏览而不生产内容的用户,和对Web2.0信息有更高层次需求的企业序化信息接受者(简称企业接受者)等,如豆瓣网和优酷网。

相比于信息微内容,用户微内容有特定的结构和意义(如用户的注册信息等),但如果不对这些孤立的信息进行序化加工或信息组织,就没有实际的应用价值。将用户微内容与信息微内容区分为不同类型,一方面可以对用户信息进行序化,这同样应作为Web2.0信息序化必不可少的一部分,尤其在社交网站中应用较多;另一方面,由于网络中存在部分用户借助Web 2.0平台发布广告、虚假信息或不负责任的评论等,用户微内容在过滤这些信息的创造者方面亦发挥重要作用。

Web2.0用户分类和微内容分类如图1所示:

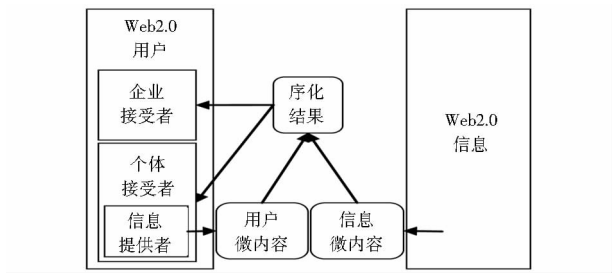

图1 Web2.0用户分类和微内容分类

综上所述,用户微内容来自Web2.0用户中的信息提供者自身,信息微内容来源于信息提供者发布的Web2.0信息。将这两部分微内容进行序化加工之后得到的序化结果,可以被两部分Web2.0用户所使用——作为个体接受者的网民注重从微内容中获得清晰、有序的信息,企业接受者则注重从微内容中得到具有商业价值的信息。

2 微内容序化指标

微内容是杂乱、无序的,微内容的序化即通过一定线索将杂乱的信息进行整合,从一个或多个维度将信息组织在一起,使孤立的信息处于不同维度的信息序列之中。这些维度就是对信息序化加工的依据,即微内容的序化指标。经过对多个主流的Web2.0网站的微内容进行汇总和抽象,可以发现,虽然各个网站的序化指标的名称各有不同,但本质上主要包括以下方面:

2.1 用户微内容的序化指标

2.1.1 *用户基本信息* 用户基本信息是最基本的用户微内容序化指标,一般是在注册时就需要用户提供,如姓名、性别、单位等,方便网民在Web2.0网站建立自己的人际网络。

2.1.2 *标签* 标签的内容是任意的。通过标签,可以构建一个多维的超空间(一般标签可以最多设置8至10个),每个用户都是这个空间的一个多维交叉点,在用户基本信息之外设置了新的用户序化角度,使用户在杂乱的Web2.0用户群中快速发现感兴趣的结果。

除了用户基本信息,还有很多指标可以刻画用户的网络行为特点,比如用户可信度、活跃度等,相关指标如下:

2.1.3 *认证情况* 现在大部分微博都提供了认证的功能,主要是为了确保微博的信息真实、准确,鼓励用户对自己言论的真实性负责。

2.1.4 *微内容数量* 用户创造的数量越多,表示用户在Web2.0网站中越活跃。

2.1.5 *被关注数* 一个用户的被关注数越大,就能代表他在人际圈中越活跃。

2.2 信息微内容的序化指标

2.2.1 *微内容的质量* “微内容的质量”用来度量微内容自身的优劣程度。不同的用户由于能力或者态度的差异,会创造出不同质量的微内容。通过设置这一指标,可以将质量高的和质量低的微内容区分开,从而使质量高的微内容被保留,质量差的微内容被淘汰。

2.2.2 *微内容的受关注程度* 一个质量很高的微内容也许并不是很受关注的,受关注程度高的微内容更容易引起用户的重视,对用户的价值更高。

2.2.3 *时间价值* 一般来说,距当前时间越短的信息越有价值。使用时间价值可以方便用户追踪最新的最有价值的新闻,避免用户接收时间久远的大量无效信息。

2.2.4 *标签* 类似于用户微内容的标签指标,信息微内容的标签也可以从多个角度,尤其是传统分类方法不能涵盖的角度对微内容进行刻画,以便对微内容进行序化。

2.2.5 *用户微内容的综合结果* 不同类型的用户发布的微内容的质量有很大差别,一般活跃用户发布的微内容质量较高。另外,通过这个指标也可以将一些恶意的Web2.0用户发布的微内容过滤掉。

3 微内容序化层次

在获得了微内容的序化指标之后,就可以从不同

维度对微内容进行序化加工。熊志云认为,信息整序过程,是根据人类已有的关于序化的知识体系,根据不同的序化目的,采用合适的方法,加工无序的对象并使之有序的过程[7]。目前,主流的 Web2.0 网站中将微内容的序化加工分为三层。

3.1 单一指标的加工

第一层是对单一指标进行加工,得到序化结果。比如将“微内容的质量”指标由高到低进行排列,就可以得到“口碑榜单”。这种形式的一种变体是将微内容的静态的指标变为动态指标。比如,以“被关注数”在一天内的增加值为指标进行排列,得到的就是“每日名人榜”。通过设置不同的排列方式,可以从各个方面刻画微内容的静态和动态的特征,满足用户不同的需求。

3.2 多个指标的加工

第二层是将两个以上的指标进行汇总,采用综合指标作为序化依据进行序化加工。由于微内容含有很多度量的维度,对单一的指标进行序化,虽然简单高效,但有时序化包含的信息过杂,比如得到的“最新话题”榜单,如果没有区分地域,就会将国内外的话题都包含进来,而很多用户可能并不关心国外的新闻;或不能兼顾多个维度的序化需求,如微博中的热度榜,如果仅按粉丝数量排序会有所偏颇,应该再加入生产微博的数量和近期生产微博的频率等指标,综合评价。因此,在依据单一指标进行序化的基础上,有必要再引入其他的指标,构成序化指标体系,更有针对性地对微内容进行序化。比如,将“时间价值”和“微内容受欢迎程度”结合在一起,就可以得到在某个时间段内“最热”的微内容。

3.3 对维度全部信息的加工

第三层是将某一维度的全部信息进行汇总整理,并将整理的结果展示出来。如新浪微博的“经典语录”功能可以通过某种内部算法,从全天的微博中选出经典的部分汇总起来,用户浏览这个应用中的少数微博,就能够把握一天中最重要的信息,增加了用户在“信息过载”情况下抓取重要信息的便捷性。或者,将大量微内容的标签抽出,分类汇总并进行统计分析。在大量样本的情况下,尽量减少单一信息对结果的影响,获得比较均匀、可信的统计结果,从中可以分析出现时流行的标签,代表了主流用户的兴趣,可以在此基础上进行更深入的研究。

4 微内容序化应用实例

随着 Web2.0 网站的快速发展,很多研究关注于某个 Web2.0 网站的体系结构或运营模式,或是从 Web2.0 网站中选取样本数据进行实证分析,如王玉珏、郭玉锦比较了土豆网和抓虾网的运作模式[8],张英杰、冷伏海通过抽取国内“类 Twitter”网站中的微信息,进行用户关系网络研究[9]。

本文通过一个实例,从前文提出的微内容序化指标中选择合适的指标,对电影评价信息微内容进行第二层序化加工,获得综合排名榜,以提升用户寻找高品质电影的效率,并将结果和主流的电影排序结果相比较,在一定程度上验证了本文中序化方法的合理性;进一步,将这个结果中的优秀电影的特点抽取出来,即对电影微内容的“标签”指标进行第三层加工,近似获得大众的兴趣点,用以辅助相关网站进行商业决策,比如视频网站如何选择购买正版电影和如何给电影播放前后插播的广告定价的问题。

4.1 面向个体接受者的序化加工实例

4.1.1 *构建指标体系* 本案例设计的排行榜旨在帮助用户发现综合水平最高、获得广泛好评的电影。“微内容质量”和“微内容受关注程度”两个指标与本文排行榜的目标有直接联系。因为,一部电影的优劣,最直接的评价标准就是网民对电影的评分;另外,一些电影质量很高,但由于某些原因,不被大众接受,在考虑外部性的情况下,引入“微内容受关注程度”指标,作为“微内容质量”的补充。本文的样本数据为豆瓣网中的 100 部电影,其中,“微内容质量”是电影的评分,“微内容受关注程度”可以近似看成是参与评分的人数。

4.1.2 *指标抽取和归一化* 即将所选的用户对电影的评分和评价人数两项指标抽取出来。在抽取时注意对用户进行初步的筛选,把用户信息极不完整、在豆瓣网发布信息数量极少、对电影评价次数极少的用户界定为不可信用户,将不可信用户的评分记录去除,以提高电影评分的可信度。之后进一步对数字进行归一化。因为,评分与评价人数之间的数字绝对值之间差距很大,如果直接按原始数据进行加工,则会导致结果中评价人数指标的影响过大。

4.1.3 *电影排序* 将指标归一化之,按照赋权求和的方法计算出每部电影的结果值,将结果值从大到小排列,即可得到电影的综合排名。本文中选取评分的权重为 0.8,观看人数的权重为 0.2,原因有两点:①电影评分是本序化加工框架的主要因素,而评价人数是作为辅助的指标;②评价人数会受一些其他因素的影响,如国内是否引进、是否有翻译版本等,而这样的影响与电影本身的优劣是无关的,所以,评价人数的权重过高

会导致不客观的结果。

将本文的排名与豆瓣250和时光150的排名进行对比,结果如表1所示:

表1 排序结果对比(前15位)

本文排序	豆瓣250排序	时光150排序
肖申克的救赎	肖申克的救赎	肖申克的救赎
这个杀手不太冷	美丽人生	海豚湾
阿甘正传	海豚湾	盗梦空间
海豚湾	这个杀手不太冷	阿甘正传
美丽人生	阿甘正传	辛德勒的名单
盗梦空间	霸王别姬	教父
霸王别姬	盗梦空间	天堂电影院
辛德勒的名单	辛德勒的名单	乱世佳人
机器人总动员	机器人总动员	机器人总动员
三个傻瓜	三个傻瓜	这个杀手不太冷
海上钢琴师	海上钢琴师	霸王别姬
教父	教父	博击俱乐部
天堂电影院	乱世佳人	三个傻瓜
乱世佳人	天堂电影院	教父2
放牛班的春天	博击俱乐部	美国往事

通过比较可以发现,3个排行榜的前15位电影中,共有11部电影相同,从前15位的排名顺序来看,也有一定的相似度。但是,网络中的各种电影排行榜五花八门,每个人心中对电影的优劣评价又各有不同,很难有统一的判定标准,本实例只是将构想的序化模型予以实现,目的在于提出一个微内容序化的原始框架,是否有更合理的序化指标和序化方法,还有待进一步的研究。

将上文的权重值从(0.8,0.2)调整到(0.9,0.1)、(0.7,0.3)。从结果中可以发现,除了个别电影的排名有变化之外,总体的前15部电影排序结果是完全一样的,说明即使放松对权重值的限制,排序结果仍固定不变。表明通过选取两个指标对电影信息进行第二层序化的结果是稳定的。

4.1.4 局限性分析 网络中难免有不负责任的评价信息,如用户随意对电影进行评分,或是网络营销人员刻意抬高某个电影的评分等。针对此问题,用户微内容可以为过滤这些用户提供参考,如收集用户以往的评分与大众综合评分的差异或曾经在网站上发布过不良信息的记录等。通过对用户信息的审核,减少此类用户的评价信息对研究的影响。但鉴于用户的详细信息网站并不公开,所以本研究具有一定的局限性。

4.2 面向企业接受者的序化加工实例

4.2.1 标签的抽取和过滤 从豆瓣网中将上文获得的前15名最佳电影的标签抽出,每个标签都代表了电影的某个特点。将其中无意义的标签去除,如"经典"、"电影"和与电影名重合的标签名。另外,上文中的电影排行榜是将各个年代的优秀电影一起排序,电影的出版年代没有意义,所以,将年份的标签也去除。

4.2.2 标签的聚类结果 将排行榜中的前30部电影的标签作为样本,按上文的过滤规则筛选,将其中属于近义词的标签转变为同一形式,统计所有标签数量,将标签按数量由大到小排序,给出前15位使用最多的标签。结果的含义是:当电影中具有以标签为特点的内容时,比较能吸引观众,观众反响较好。

4.2.3 评估最新电影并辅助决策 本文从2011年5月上映的电影中选取5部,从豆瓣网中抽取网民给予该电影的标签,并与上文得出的结果相比较,按标签匹配的数目从大到小将电影排序得到的结果为:

功夫熊猫2(4个) = 加勒比海盗4(4个) > 最爱(3个) > 速度与激情5(2个) > 不再让你孤单(1个)

根据本文的分析,可以从一定程度上推断,按上面的顺序由前至后采购这5部新片,可以使网站获得较高的收视率。

参考文献:

[1] 熊回香. Web2.0环境下的网络信息组织[J]. 情报资料工作,2007(5):29-32,50.

[2] 张如法. 编辑学研究的误扩与回归[J]. 编辑学刊,1996(2):11-16.

[3] 吴华欣,于雄杰. 文献信息开发的四个层次[J]. 图书馆建设,1993(3):23-24.

[4] 王松林. 信息组织论[J]. 图书馆学刊,2005(6):1-4.

[5] Chen Yubo, Xie Jinhong. Online consumer review: Word-of-mouth as a new element of marketing communication mix[J]. Management Science, 2008,54(3):477-491.

[6] Kar Y T, Shuk Y H. Understanding the impact of web personalization on user information processing and decision outcomes[J]. MIS Quarterly, 2006,30(4):865-890.

[7] 熊志云. 信息整序漫谈[J]. 湖北大学学报(哲学社会科学版),2003,30(5):117-120.

[8] 王玉珏,郭玉锦. 两个Web2.0网站运作模式的比较分析[J]. 北京邮电大学学报(社会科学版),2009,11(3):26-31.

[9] 张英杰,冷伏海. Twitter类网站微信息组织及用户关系网络研究[J]. 图书情报工作,2010,54(16):116-119.

〔作者简介〕张 乾,男,1988年生,本科生。
蔡淑琴,女,1955年生,教授,发表论文70篇。
石双元,男,1962年生,副教授,发表论文40篇。