

基于Web挖掘与社会网络分析的用户聚类研究^{*}

□ 张敏 刘海鹏 李鹏 / 西南大学计算机与信息科学学院 重庆 400715

摘要: Web用户聚类对于个性化服务、网站结构优化等具有重要意义。文章从用户的访问路径、项目评分等角度总结了用户聚类方法及算法,指出当前用户聚类研究存在的不足,提出了结合Web挖掘与社会网络分析方法的用户聚类的模型,分析了两者结合的必要性及结合策略,形成较为完善的用户聚类机制。

关键词: Web挖掘, 社会网络分析, 用户聚类

DOI: 10.3772/j.issn.1673—2286.2014.03.007

1 引言

用户聚类,是指将具有相似行为特征的大量个体用户聚合划分为一定的用户群,使得同一群内的用户特征相似,而不同群间的用户特征相异。当前的用户聚类,主要是指Web用户聚类,与之相近的概念还有用户聚合、社区划分、社区发现等。

Web中的海量信息仍在不断增长,新的Web服务形式不断涌现,Web环境中数据量巨大、数据类型丰富、网络结构复杂的特征极为明显。进入Web2.0时代后,Web环境更是呈现出新的特征:大量信息内容由用户生成(如博客、社会化标签、用户上传视频等),用户与系统之间、用户与用户之间的交互增强(如“关注”、“评论”、“回复”、“推荐”等功能),用户在现实生活中的社会关系在Web中再生复现(SNS、网络社群等),用户行为呈现出群体化、协作化的特征(如维基百科、团购等)。因此,在这种人人参与、去中心化的Web2.0环境下,将具有相似信息行为特征的用户聚合为用户群,对于提高个性化信息服务、过滤与推荐服务、合作信息检索效果、在线社区发现均有重要作用。

本文在分析已有用户聚类研究现状的基础上,发现当前用户聚类研究还存在一些不足,提出了改进的基于Web挖掘的用户聚类机制。然而,单纯的Web挖掘对于Web2.0环境中用户行为的作用有限,极少考虑用户

在Web2.0环境中出现的社会性、交互性的新特征。在讨论相关性与相似性关系的基础上,提出结合Web挖掘和社会网络分析方法的用户聚类机制。

2 用户聚类研究现状

本文将从用户聚类方法、用户聚类具体算法、用户聚类应用几个方面入手,分析目前用户聚类研究的现状。

2.1 用户聚类方法

通过文献调研发现,目前的用户聚类方法大体有两类:一类是通过Web日志挖掘用户访问路径,基于用户访问路径相似性或者在用户访问路径的基础上抽取用户兴趣、用户偏好的相似性聚类不同的用户;另一类是基于不同用户对项目评分的相似性对用户进行聚类,这一类方法主要用于协作过滤推荐系统。

2.1.1 基于访问路径的用户聚类

基于访问路径的用户聚类,一般是通过对Web日志预处理的基础上进行用户识别、会话识别和事物识别,使用用户点击次数、点击频率、相对访问时间、停留时间、访问频率、访问次序等变量构筑用户访问模式、

^{*} 本文为国家社会科学基金“网络学术社区的信息聚合与共享模式研究”(编号:11CTQ038)的研究成果。

用户访问路径相似矩阵或其等价矩阵等,对这些处理好的数据采用已有的聚类方法或者基于传统聚类方法改进算法进行聚类,识别具有较高相似度的用户群。具体的方法有:

使用几个变量共同表征用户兴趣或用户访问模式从而进行聚类。如文献[1]在用户聚类前,首先对用户路径进行约简,得到用户感兴趣的页面,然后根据路径相似度得到用户相似度矩阵,在此矩阵中使用编网法直接进行聚类得到最终的用户集合。文献[2]中提出改进的用户兴趣模式聚类算法,综合考虑了用户访问页面顺序,建立用户浏览路径相似度矩阵来获得聚类结果。文献[3]中通过数据预处理得到的用户点击流来表示用户的访问行为模式,并采用混沌蚁群算法(CAS-C算法)对用户访问矩阵进行聚类分析。文献[4]重新定义了相似性和聚类中心,综合考虑用户访问的次序与频率,提出了一种能对不规则用户访问路径进行访问模式聚类的方法。文献[5](通过页面浏览顺序与浏览时间表现用户访问兴趣,采用结合了粗糙集的Leader方法进行聚类。文献[6]同时考虑用户对URL的浏览时间和访问次数对用户进行聚类。文献[7]将反映用户浏览行为的页面点击次数、停留时间、用户偏好等因素用模糊多重集来刻画用户访问站点的兴趣度,建立模糊多重相似矩阵直接进行聚类。文献[8]对用户特征进行数学分析,将访问点击次数和访问时间视为数值型参数,并以这两个参数定义Web资源偏爱度和关联度,将访问路径视为过程型参数,简化为无向图,在此基础上提出基于Kruskal的Web用户聚类算法。

结合用户访问路径与网页内容进行聚类。如文献[9]在分析用户访问模式的同时,引入网页内容分析和浏览行为分析结果,通过对用户访问模式向量表示的改进,构造基于网页内容和用户访问兴趣的访问模式相似性,最后采用蚁群混合聚类算法实现基于Web浏览内容和行为的用户聚类。文献[10]分别从用户浏览路径的结构和内容两个方面实现网络用户聚类。其中以用户的会话作为用户聚类的特征,实现基于浏览路径结构的聚类;引入目标页和导航页的概念,先对目标页聚类,再利用其聚类结果进行用户聚类,实现基于浏览路径内容的用户聚类。文献[11]使用点击率与相对浏览时间来定义用户兴趣度,同时综合考虑网页内容的长度,使用粗糙聚类方法对Web用户聚类。

此外,对于基于访问路径的Web用户聚类可进一步细分。文献[12]中提出:部分用户聚类方法将用户浏览

的会话用矩阵的形式表示,可以称之为基于矩阵的聚类算法BOM(Based On Matrix);而另外一些用户聚类算法把用户会话浏览序列用集合来表示,称之为基于集合的聚类算法BOC(Based On Collection)。在文献调研的基础上,沿着这样的分类思路,本文提出基于向量的聚类算法BOV(Based On Vector),即将用户会话划分,生成会话向量。如文献[13]提出通过提取网络会话信息,表述为会话向量,实现对匿名用户的聚类;文献[14]同样采用了会话向量的方法。同时,这两个研究中都采用了利用页面层次性对会话向量降维的方法。文献[15]从会话向量中发掘频繁数据集,归一化为模式向量,采用SOFM模型进行聚类。

2.1.2 基于项目评分的用户聚类

基于项目评分的用户聚类主要使用于基于合作的过滤推荐服务中。多数基于用户聚类的合作过滤推荐均分为离线部分和在线部分,离线部分的主要任务就是将大量个体用户聚合为具有一定相似度的用户群。

文献[16-19]均利用用户对项目的评分进行用户聚类研究,从而用于过滤与推荐。文献[17]将用户评分看作数据流,提出利用金字塔框架进行预处理,从而体现用户兴趣随时间的变化;文献[18]则利用用户项目评分的概念分层实现多层相似性的用户聚类;文献[19]应用模糊聚类技术将单个用户对项目的评分转为用户相似群对项目的评分,构建模糊评分矩阵实现聚类分析。

除了上述的基于访问路径和基于项目评分的用户聚类方法外,文献[20]将用户聚类方法划分为静态方法和动态方法两类。同时,有一些研究比较关注用户聚类过程中的降维问题。如在上文中提到的文献[13,14]利用页面的层次性对会话向量降维,文献[21]提出利用页面规约的方法降维,文献[22]提出基于方向相似性的蚁群聚类算法,文献[23]提出了“比对降维”的思想,文献[24]提出使用FCC(Filter Coefficient Clustering)算法解决“维灾难”问题。

2.2 用户聚类算法

目前已有的用户聚类研究中,采用的具体聚类算法存在以下三类情况:

(1)直接使用某一种聚类算法。如K-Means算法、K-Medoids算法、Chameleon算法、leader层次聚类算

法^[13]、SOFM聚类、模糊C均值FCM (Fuzzy C Means) 算法^[25]、遗传算法、超图聚类^[21]、 λ 截聚类、非欧几里得的关系模糊聚类方法 (NERF)^[14]等。

(2) 结合使用几种聚类算法。如文献[9]采用蚁群算法与K-Means混合聚类算法, 文献[5]将粗糙聚集中的粗糙度的概念引入leader聚类算法中。

(3) 基于一般聚类算法的改进或设计新的适用于Web用户聚类的算法。如文献[17]在AntClass的基础上引入金字塔框架设计的AntStream算法, 文献[10]提出的改进的用户浏览路径聚类算法 (UBPC), 文献[3]提出的混沌蚁群算法 (CAS-C), 文献[26]提出的ISODATA (Interactive Self Organizing DATA) 算法, 文献[7]的多重模糊集CAFM (Clustering Arithmetic based on Fuzzy Multisets) 聚类算法, 文献[8]基于Krushal的Web用户聚类算法 (Krushal-Based Algorithm of Clustering Web-User, K-Bacer) 等。

2.3 用户聚类应用

当前Web用户聚类主要用于: (1) 协同过滤推荐。主要是电子商务网站、娱乐网站等基于用户聚类的过滤与推荐, 提高用户服务质量。(2) 网站结构优化。如文献[27]提出基于用户聚类分析网站导航结构的智能优化问题。(3) 个性化服务。如文献[25]试图利用用户聚类结果实现图书的个性化推荐。(4) 改善人机交互, 提高信息检索效率。如文献[28]提出以用户为中心的用户事务聚类方法, 认为聚类结果可用于分析和理解用户可能的查询意图, 在用户查询与搜索引擎返回结果的人机交互过程中, 引导用户更快速准确定位自己所关注的内容。文献[29]中提出一种基于用户行为聚类的搜索引擎, 可以通过分析不同的用户行为将搜索用户聚类成不同的用户组, 为每组用户返回其喜欢的结果, 优化查询结果。

3 当前用户聚类研究中存在的不足

用户的网络行为产生于其信息需求的驱动。而用户的信息需求一般来源于两个方面: 解决任务或问题所需的信息和满足用户个人兴趣、休闲所需的信息。对应的信息行为主要有两类: 信息检索行为与信息浏览行为, 这两者在一定程度上代表着用户对不同信息资源的偏好和访问模式的不同。而当前用户聚类研究中, 一般只是直接利用Web日志挖掘用户的访问路径模式, 在预

处理的过程中没有对这些原始素材作出分类以完成不同目的的聚类分析。

用户的兴趣或者行为特征具有多样性, 因此其聚类结果有一定的模糊性, 即用户既可以属于某一用户群, 也可以同时聚类到另外的用户群。当前的聚类算法多为硬聚类算法, 聚类结果唯一, 降低了聚类结果的准确性和可用性。

用户的行为特征会随时间、任务、兴趣变化等变化, 新的用户也在不断涌现。目前多数用户聚类实现研究中均未考虑到用户聚类结果的更新问题, 如何采用一定的方法保证用户聚类数据有效地更新, 是有待解决的一大问题。文献[22]提出了一种基于用户聚类模型维护库的增量式用户聚类方案。

用于表征用户访问路径模式的变量有点击次数、点击频率、相对访问时间、停留时间、访问频率、访问次序等, 一般只是简单地将其中的几种变量结合, 未能对这些变量的可表征程度作出系统的分析, 变量结合中未能给出对应的权重等。

访问路径分析与页面内容分析结合不足。用户访问路径模式在一定程度仅代表了用户访问网络的一种使用习惯, 而其所访问的具体页面 (特别是内容页) 的具体内容的语义才真正足以表述其兴趣。

4 基于Web挖掘与社会网络分析的用户聚类

根据上文提出的当前基于Web挖掘的用户聚类研究的不足, 本文提出了一种改进的基于Web挖掘的用户聚类机制。如图1所示。

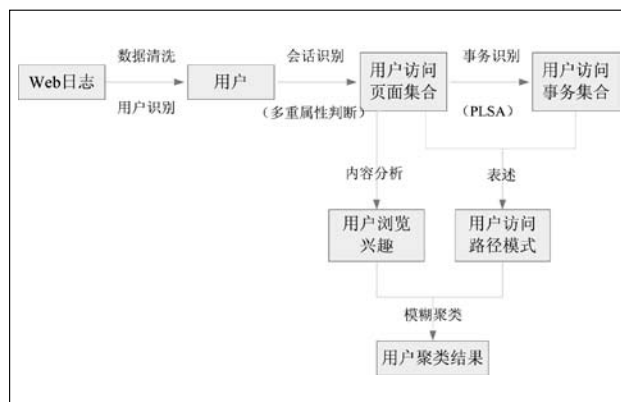


图1 改进的基于Web挖掘的用户聚类

通过对Web日志的预处理,分别实现用户识别、会话识别和事务识别。其中,在会话识别过程中采用综合多重因素评价的方法,以期判别出用户在一定的时间阈值(Timeout)内访问的页面集,排除偶然点击的链接等噪声;在事务识别过程中,采用概率潜在语义分析,以期识别出最有代表性的语义概念,减少主题词维度。预处理过后,利用获得的页面集合与事务集合构建事务-页面矩阵,用于表述用户访问路径模式;同时,对会话进行文本分析,获取用户浏览兴趣。最后,结合用户访问途径模式与用户浏览兴趣,采用具有更好模糊性质的软聚类算法进行用户聚类,得到用户聚类结果。

然而,这种改进后的用户聚类机制还是未能适应Web2.0环境的特征。Web2.0的社会性、去中心化、群体性、用户之间紧密的协作与联系,使得Web2.0中客观存在着这样一些具有相关关系的用户群。用户与用户之间的交互增强(如“关注”、“评论”、“回复”、“推荐”等功能),用户之间越来越多地表现出协作共享行为,而这些相关用户群中,较为稳定的便是用户所处的社会网络。

本文认为,相对于通过Web挖掘的用户间的相似性而言,用户在Web2.0环境中相关性是一种较强的“关系”或“连带”,相似性是一种较弱的“关系”;相关性在一定程度上,是一种显在的“关系”,极易浮出水面,而相似性是一种隐性的“关系”,需要通过一定的手段进行挖掘^[30,31]。Web2.0环境中用户聚合而成的群体,或者称用户社群,在进行形式化的描述时可以认为是社会关系网络与兴趣图谱的组合,即一个用户与其他用户的相关性中可能包含有兴趣相似性。因此,仅仅通过Web挖掘手段识别用户间的相似性来聚类用户是不尽完善的。

(1)传统的Web挖掘技术,无论是基于用户使用记录或项目评分,都是通过内容相似度计算,分析用户之间的关联(主要表现为兴趣相似性或人口统计学相似性),从而进行用户聚类,主要体现的是一种基于内容的聚类。Web2.0环境下,用户之间的相互影响与作用机制更为明显,用户之间的关联或者关系的内涵表现得更加丰富,除兴趣关系外,还有可能由交互行为产生相互影响关系(关注、链接、转发)等,单纯的Web挖掘技术无法揭示更丰富的用户关联层次,而社会网络分析关注的焦点是关系和关系的模式,有别于传统的统计分析和数据处理方法,采取结构分析路径,着眼于行为

体间的“位置及其相对关系”。社会网络分析主要从结构和用户行为两方面的因素进行分析,其中测量用户聚类节点重要性和节点间关系的常用方法主要有两类:一类是中心性分析;另一类是拓扑链接结构分析^[32-34]。通过与已经存在于Web中的社会关系网络的比对,可以以一个用户群内某个用户节点的中心势,确定聚类中心,删除一些无关节点,识别关键用户聚类节点和核心用户,改善用户聚类效果。由此可见,相比于传统的内容聚类,社会网络结构的方法是一种关系聚类,能从整体上把握社交网络的总体特征和用户交互情况,揭示用户之间的关系结构、关系性质及强度。

(2)非社交网络的网站中,例如电子商务网站,内容内在的联系是很重要的聚类原则,因为用户访问内容的数据相对稳定,因此相似度不但计算量较小,同时也不必频繁更新,因此传统Web挖掘技术在性能与复杂度方面具有一定优势,然而在对于新闻、博客或者微内容等网络系统,用户访问内容数量是海量的,同时也是更新频繁的,所以从复杂度的角度看基于用户关系的聚类更具优势。

(3)两者的结合方式上,可以采用多重聚类或者层次聚类的方式。多重聚类即分别对内容和关系结构进行Web挖掘和用户关系分析,对用户聚类的结果综合分析对比。层次聚类即先通过Web挖掘,进行用户聚类分析,然后在聚类类别中采用社会网络分析方法,分析成员的关系性质、亲密度、角色等。

基于这样的思路,在改进的基于Web挖掘的用户聚类的基础上构建了如图2所示的用户聚类机制。图中还需要说明的一点是,对于多数Web2.0用户而言,为了能

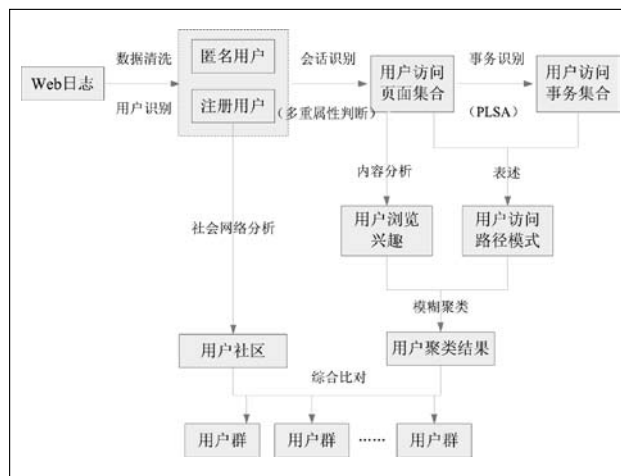


图2 基于Web挖掘与社会网络分析的用户聚类

够维持与其他用户特别是出于自己的社会关系网络中的用户的联系与交互,一般均为某些Web2.0服务的注册用户,因此很容易进行用户识别。然而,由于某些站点为了提高用户体验,匿名用户在一定范围内也可以浏览。因此,经过识别后将用户区分为两种,主要针对注册用户进行社会网络分析。社会网络分析的数据源为这些Web2.0站点的服务器。

5 结语

本文在分析用户聚类研究现状的基础上,提出了改进的基于Web挖掘的用户聚类机制。然而,改进后的用户聚类机制在作用于Web2.0环境中的用户时仍有其局限性,因此进一步提出结合Web挖掘和社会网络分析的用户聚类方法,同原有机制相比在理论上比较完善。本文的不足在于,没有每一个关键步骤提供具体可行的算法,也未能利用一定的系统与实例进行验证,有待进一步的补充和验证。

参考文献

- [1] 王华,王治和,王平.Web用户聚类研究[J].甘肃联合大学学报(自然科学版),2010,24(1):79-82.
- [2] 陈峰.基于Web日志的用户兴趣聚类研究[D].合肥:合肥工业大学,2008.
- [3] 郭晓磊.基于Web日志挖掘的网络用户聚类研究[D].北京:北京邮电大学,2009.
- [4] 杨彦玲,任燕,段隆振,等.基于不规则路径聚类算法的用户访问模式挖掘[J].计算机与现代化,2010(4):91-93.
- [5] 陈敏,苗夺谦,段其国.基于用户浏览行为聚类Web用户[J].计算机科学,2008,35(3):186-187,255.
- [6] 宋江春,沈钧毅.一种新的Web用户群体和URL算法的研究[J].控制与决策,2007,22(3):284-288.
- [7] 宋麟,王锁柱.基于模糊多重集的Web页面与用户聚类算法[J].计算机工程与设计,2008,29(1):213-215.
- [8] 吴跃进.综合多重评价因素的Web用户聚类算法[J].计算机工程与应用,2006(28):147-149,210.
- [9] 彭艳,王小玲.基于Web浏览内容和行为的用户聚类算法研究[J].计算机与信息技术,2007(9):39-40,44.
- [10] 付志涛.基于Web日志的网络用户聚类研究与实现[D].南京:南京理工大学,2007.
- [11] 纪洲鹏,周军,何明.基于变精度粗糙集的Web用户聚类方法[J].计算机工程,2010,36(3):44-46.
- [12] 王凯丽.一种基于集合的Web用户会话实时聚类算法[J].价值工程,2010(13):182-183.
- [13] 宋斌,王玲,张宏,等.基于Web日志的匿名用户聚类算法研究[J].南京理工大学学报,2006,30(5):583-586.
- [14] 黄松,刘晓明,宋自林.基于归纳化会话的网络用户的聚类[J].计算机研究与发展,2001,38(10):1224-1228.
- [15] 徐涌,陈恩红,王熙法.基于神经网络的Web用户行为聚类分析[J].小型微型计算机系统,2001,22(6):699-702.
- [16] 邓晓懿,金淳,等.基于情境聚类和用户评级的协同过滤推荐模型[J].系统工程理论与实践,2013,33(11):2945-2953.
- [17] 王卫平,寇艳艳.基于AntStream用户聚类的协同过滤推荐[J].计算机系统应用,2010,19(12):180-184.
- [18] 李涛,王建东.基于多层相似性用户聚类的推荐算法[J].南京航空航天大学学报,2006,38(6):717-721.
- [19] 温会平,陈俊杰.基于用户模糊聚类的个性化推荐算法[J].计算机与数字工程,2008,36(2):13-17.
- [20] 白雪,田双亮.一种新的基于用户浏览模式的聚类算法[J].楚雄师范学院学报,2005,20(6):4-6.
- [21] 杨明花,古志民.基于超图聚类的用户行为模式挖掘[J].广西师范大学学报(自然科学版),2006,24(4):163-166.
- [22] 张斌,苏一丹,曹波.基于蚁群聚类模型的增量式Web用户聚类[J].微计算机信息(管控一体化),2008,24(15):231-233.
- [23] 颜端武,罗胜阳,成晓.协同推荐中基于用户-文档矩阵的用户聚类研究[J].现代图书情报技术,2007(3):25-28.
- [24] 业宁,李威,梁作鹏,等.一种Web用户行为聚类算法[J].小型微型计算机系统,2004,25(7):1364-1367.
- [25] 孙守义,王薇.一种基于用户聚类的协同过滤个性化图书推荐系统[J].现代情报,2007(11):139-142.
- [26] 刘国营.基于路径聚类的Web用户访问模式发现算法[J].情报杂志,2005(7):18-20.
- [27] 郑玲霞,李大学.基于用户聚类分析的网站导航结构智能优化研究[实现[J].重庆邮电大学学报(自然科学版),2005,17(6):763-767.
- [28] 唐晓兵,刘震,李盛恩.基于用户特征的聚类方法在人机交互中的应用研究[J].网络与通信,2008(4):28-29.
- [29] 郑双阳,林锦贤.基于用户行为聚类的搜索[J].计算机与数字工程,2009(12):28-30,75.
- [30] 约翰·斯科特.社会网络分析法[M].刘军,译.重庆大学出版社,2007.
- [31] 陈克寒,韩盼盼,等.基于用户聚类的异构社交网络推荐算法[J].计算机学报,2013(2):349-359.
- [32] 王连喜,蒋盛益,等.微博用户关系挖掘研究综述[J].情报杂志,

2012(12):91-97.

[33] HANNON J, MCCARTHY K, SMYTH B. Finding useful users on twitter: twittomender the followee recommender [C]// Advances in Information Retrieval 33rd European Conference on IR Research,

2011.

[34] 张丹,何跃.基于聚类分析的SNS网络研究[J].情报杂志, 2012(5):62-65.

作者简介

张敏 (1974-), 女, 西南大学计算机与信息科学学院副教授。E-mail: zhangwu@swu.edu.cn

刘海鹏 (1989-), 男, 西南大学计算机与信息科学学院2013级图书馆学专业硕士研究生, 研究方向: 信息检索与用户研究。

李鹏 (1987-), 男, 西南大学计算机与信息科学学院2010级情报学专业硕士研究生。

User Clustering Based on Web Mining and Social Network Analysis

Zhang Min, Liu Haipeng, Li Peng / Computer and Information Science, Southwest University, Chongqing, 400715

Abstract: Web user clustering has great significance in personalized service, website structure optimization, etc. In this paper, user clustering method and algorithm based on user access paths and item rating are summarized. The shortage of these studies is pointed out, and then an improved model combined with social network analysis and Web mining is put forward. With an analysis of the necessity and the application strategy, a relatively perfect mechanism of user clustering has come into being.

Keywords: Web mining, Social network analysis, User clustering

(收稿日期: 2014-01-10)