挖掘用户浏览网页的兴趣研究①

曹 易，张 宁

(上海理工大学 管理学院，上海 200093)

摘 要：通过挖掘网页的浏览记录来对用户群体兴趣进行分析。对访问网站的兴趣类别、时间、用户数进行统计，得到规律性的结论。其次提出一种改进的基于 HAC 和 k-means 的算法对用户根据兴趣进行聚类，挖掘用户的访问模式。最后验证了主导兴趣的稳定性即随着日志的增加，用户的最大兴趣是趋于稳定的。

关键词：群体兴趣；数据挖掘；层次聚类；k-means；主导兴趣

Study of the Uses' Interests Based on the Internet Browsing History

CAO Yi, ZHANG Ning

(Business school, University of Shanghai for Science and Technology, Shanghai 200093, China,)

Abstract：This paper analyses the users' group interests by mining the internet browsing history. To count the visiting information of the interests' categories, visiting time and the number of users, get the regularity of conclusion. Then, it has put forward an improved HAC (hierarchical agglomerative clustering) and k-means algorithm to cluster the users by their interests, to mine the users' access mode. Finally, it has proved the stability of users' dominant interests. That means the users' most important interests are stable as the time increases.

Key words：group interests; data mining; hierarchical cluster; k-means; dominant interests

随着 Internet 的迅猛发展，在当今信息爆炸的时代，Internet 和 WWW 都以指数形式在增长，用户越来越难在信息海洋中找到自己感兴趣的内容。“数据丰富，知识匮乏”，面对这些海量的信息，如何克服这个“数字鸿沟”，如何能准确、快捷、高效的获取有用信息，让人人都公平地享有信息资源，无疑是人们关注的一项问题，同时也是一个全球性难题。用户对互联网的浏览行为，是人们获取信息的一种重要方式，每次访问大都具有一定的访问动机，蕴藏着用户的某种兴趣。通过分析这些网页的浏览记录，对个性化服务技术等方面具有很大的实际应用价值。

目前对兴趣分析研究主流用的是聚类分析法和线性回归分析法，来挖掘用户的访问信息，可在此基础上提供个性化服务。国内外机构学者对该领域有：Dan 等人提出利用 Web 挖掘分类方法，基于 Web 访问信息挖掘用户建模技术[1]，Jose Borges 等人提出了挖掘用户的导航模式方法[2]，Perkowits 用聚类分析的方法研究了 Web 访问的自适应性[3]，中科院的高文教授提出了对 Web 访问路径进行聚类，每个聚类集就代表了该集合内用户的访问兴趣[4]。模糊聚类分析的基本思想是根据分类对象之间的模糊相似性来考量不同对象的异同程度，从而来实现模糊分类。De 等[5]使用模糊聚类理论对 web 事物进行聚类。Mitra 等[6]提出了一种进化的粗糙 k-means 算法聚类 web 用户，其中遗传算法用于阀值以及其他参数的调用，以便聚类效果达到最佳程度。

本文首先挖掘统计了网页浏览记录的一些信息，通过分析这些数据，得出了一些规律性的结论。其次，本文提出了一种改进的基于 HAC（凝聚层次聚类）和 k-means 的算法，用其来对用户根据访问兴趣进行分类。该算法弥补了 k-means 中事先确定分类数目的缺陷，因为往往在此之前我们并不知道确切的分类数目；

① 基金项目:国家自然科学基金(70971089);上海市重点学科建设项目(S30501)
收稿时间:2011-10-07;收到修改稿时间:2011-11-18

而且也克服了 HAC 中分裂和合并点的选择困难问题，因为一旦选择错误，不能弥补，以至于恶性循环，得到低质量的聚类结果。最后，我们发现了用户主导兴趣的稳定性，即随着日志的增加，用户的最大兴趣是趋于稳定的。基于上述结论，我们可在此基础上进一步地提供的个性化服务。

1　数据预处理

本文根据上海某高校网络中心服务器的网页浏览日志，该日志记录了大学校园网 2933 个 IP 用户的近 6 个月的访问信息，一条完整的记录的形式为：

表 1　记录形式

编号	访问时间	用户 IP	网站 IP	网站网址	网页	类别

对数据进行一下预处理，删除一些数据不完整或者没有意义的部分[7]。本文用访问时间、用户 IP、类别来进行分析，对一些必要的类别进行一些处理：拆分一些大的类别，合并小而相似的类别。部分类别要进行一下加权处理，一些访问量较大的例如搜索引擎，防止部分类别太突出，从而掩盖了相对关注度较小的类别，例如法律、宗教等，虽然访问量很少，但很明显地能反映出访问用户的兴趣，它们访问量乘以一个系数来作为权值，防止被其他热门类别所掩盖。

经过以上操作，我们得到了搜索引擎、新闻门户、教育、财经、IT 相关、游戏、其他等 35 个类别，这里我们把一些不常见的、访问量较少的类别归于其他类。

2　挖掘网页浏览记录的时间统计规律性

为了挖掘日志中一些信息，如在此期间内访问的用户数，访问的网页以及网页的类别，以及访问时间等等[8]。进行了以下两个方面的统计：

a．分别抽取一天、一个星期、一个月期间日志，统计总的类别访问量，结果见图 1~3。

b．分别抽取一小时、一天、一个月期间日志，统计总的访问人数，结果见图 1~3。

从下图中得出以下结论：

1) 类别访问量与时间的关系。随着访问时间的增长，类别访问量趋近于线性增加。

2) 类别访问量比重与时间的关系。随着访问时间的增长，类别访问量比重趋近于稳定值。当访问时间增长到一定程度时，类别访问量比重变化较小，几乎为稳定值。因为每天上网的人，相对来说还是挺固定的，整体的兴趣是趋向于稳定。本数据中，搜索引擎一直占据着最大访问量的类别，与用户使用搜索引擎的习惯有很好的吻合。

3) 用户数与时间的关系。一般地，日志中用户总数是基本固定的，所以随着时间的增长，用户数增长速度越来越慢，最后趋于稳定，基本保持不变，类似于对数增长。

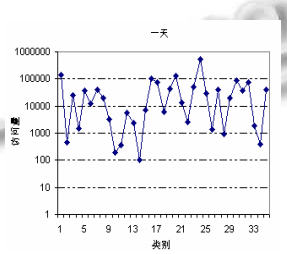

图 1　一天访问量

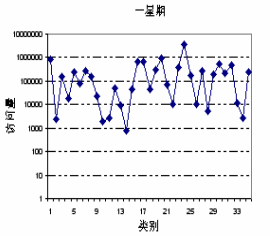

图 2　一星期访问量

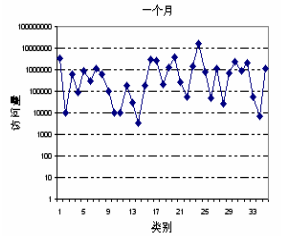

图 3　一个月访问量

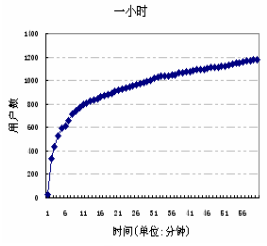

图 4　一小时用户数

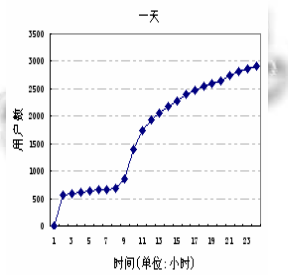

图 5　一天用户数

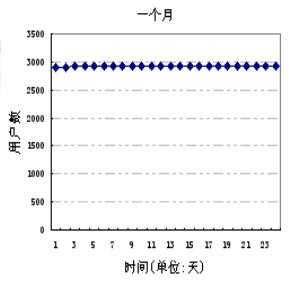

图 6　一个月用户数

3　用改进的基于HAC和k-means算法对用户根据兴趣进行聚类

3.1　问题描述

每个用户访问一些感兴趣方面的网页，比如新闻、军事、国际时事等等，但是每个方面占的比重又不同。我们把所有待做聚类的对象集合设为 $X=\{x_1,x_2,...,x_n\}$，该集合中的每个对象 $x_i(i=1,2,...,n)$ 用有限个指标来衡量，每个指标来代

表 x_i 的某个特性。因此对象 x_i 可用向量 $P(x_i)=(x_{i1},x_{i2},...,x_{im})$ 来描述，其中 x_{ij} 表示对象 x_i 中第 j 个特性的值。$P(x_i)$ 称为 x_i 的特征向量或者是模式矢量。聚类就是分析对象集合X中n个对象 $P(x_i)$ 所对应特征向量之间的相似性来划分成多个不相交的子集 $X_1,X_2,...,X_m$，其中 m 为分类数目。需要满足以下条件：

$$X_1\cup X_2\cup...\cup X_m=X, X_i\cap X_j=\varnothing$$

$$1\le i\ne j\le m$$

隶属函数 $\theta_{ij}=\begin{cases}1 & 1\le \mathrm{i}\le \mathrm{n}\ \text{ and }\ 1\le \mathrm{j}\le \mathrm{m}\\ 0\end{cases}$，

当 $x_i\in X_j$ 时 $\theta_{ij}=1$，否则 $\theta_{ij}=0$。其中隶属函数要满足条件：$\begin{cases}0<\sum_{i=1}^{n}\theta_{ij}<1,\forall j\\ \sum_{j=1}^{m}\theta_{ij}=1,\forall i\end{cases}$ 也就是说每个对象都属于且只属于一个类别，而且每个类别为对象集合 X 的非空真子集。另外相似度的建立方法有数量积法、余弦幅度法、相关系数法、最大最小值法等等[9]，本文用的是最大最小值法，具体算法公式为：

$$r(x_i,x_j)=\frac{\sum_{k=1}^{m}x_{ik}\wedge x_{jk}}{\sum_{k=1}^{m}x_{ik}\vee x_{jk}} \tag{1}$$

上述问题，可用经典的凝聚层次聚类算法（HAC）和 k-means 算法来对这些用户根据兴趣来进行分类。

3.2 改进算法以及具体实现

HAC 算法的原理是：这是一种从下到上聚类策略，事先设定某个阀值，然后设 $X=\{x_1,x_2,...,x_n\}$ 中每个对象 x_i 都独自为一类，其次根据某种相似度来合并这些类，依次类推，直到聚类途中遇到阀值条件而终止，就得到了所求的聚类结果[10]。

k-means 算法原理是：与 HAC 算法不同，k-means 仅对对象进行一层划分，将文本聚合成 k 个类。按照某种方式得到 k 个聚类中心，计算每个对象 x_i 与聚类中心的相似度，归类于具有最大相似度的那类。然后根据 k 类中对象的均值重新设为聚类中心，再次聚类。依次类推，直到所有聚类中心均达到稳定才终止[11]。

HAC 算法虽然简单，但是在聚类过程中的合并点的选择很困难，一旦选择不合适，会造成恶性循环，分类结果质量低下。另外，该算法不适合大量的对象聚类，因为每一次聚类都要比较大量的对象，效率比较低。

相对于 HAC 算法来说，k-means 能处理大量的对象聚类，结果常常呈现凸状，容易造成局部最优解。需要预先确定分类数目，而这往往是在那些不充分熟悉的对象进行聚类之前最大困难之一，因为这分类数目是未知的。

因此本文用一种改进的二次聚类算法，它结合了以上两种算法的优点，弥补了不足。经过用聚类有效性评价法 F-measure[12]实验证明，该算法虽然复杂度提高了，但是聚类结果更好准确，避免了 k-means 算法的陷入局部最优与事先定义分类数目的缺陷。具体实现步骤如下：

Step 1：将待聚类的对象集合 $X=\{x_1,x_2,...,x_n\}$ 中每行 x_i 各自组成一个分类，这 n 个分类构成了 X 的一个聚类 $X=\{X_1,X_2,...,X_n\}$。

Step 2：计算 $X=\{X_1,X_2,...,X_n\}$ 中每个类之间的相似度用公式 1。

Step 3：设定一个阀值 λ，选择相似度最大的值 $\max=\max_{x_i,x_j\in X}(r(x_i,x_j))$。若 $\max\ge\lambda$，则将 x_i,x_j 合并，组成 $x_i=x_i\cup x_j$，从而就构成了 $X=\{X_1,X_2,...,X_{n-1}\}$，重复上述步骤 Step2,3。若 $\max<\lambda$，则终止算法，得到有 k 个分类的聚类 $X=\{X_1,X_2,...,X_k\}$。

Step 4 将 k 和 $\{X_1,X_2,...,X_k\}$ 均值中心点分别作为 k-means 的分类数和聚类中心的两个参数值进行聚类。一旦聚类中心稳定，就结束算法，得到新的聚类结果 $X=\{X_1^{'},X_2^{'},...,X_k^{'}\}$，为最终结果。

根据上述改进算法，我们可以控制阀值 λ 的大小，来得到粗细不同的聚类。λ 的取值应该就具体问题具体分析，不同阀值的选定对聚类结果产生很大影响。阀值越大，分类越精，分类数目越大，这就要求类内部的相似度就越大。反之，阀值越小，分类越粗，分类数目越小，类内部的相似度就越小。

3.3 实验分析

用上述改进算法，通过实验分析 λ 取值 0.3、0.5、0.7 效果较好，因此根据该 λ 取值来对用户进行聚类，分布结果如下图：

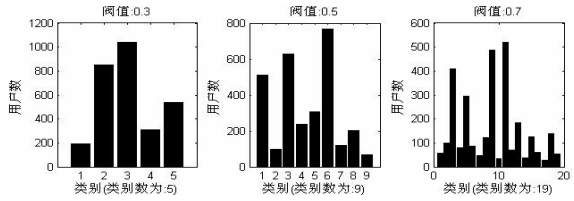

图 7　用户聚类

从上图可以看出，聚类数目随着 λ 增加而增加。用户根据访问类别的聚类结果大部分分布在几个“突出”的聚类中，说明很多用户的兴趣类似，更倾向于这几种浏览行为。而有些聚类，用户相对较少。这些结果与实际情况是比较吻合的。由于在本文所用到的高校 Web 日志数据中，大部分是 IP 用户是学生，因为热门类别搜索引擎、新闻门户、论坛博客等占据了大量的访问记录，使得把这些用户集中分布在少数某些热门类别中；而一些法律、宗教、政府组织等类别关注度低，访问人次较少，因此某些冷门类别分布的用户数较少。网站管理者可以根据以上结论，在网络上给用户进行个性化推荐时，只需要提供少数几个稳定的兴趣就能达到良好的效果。

4　用户主导兴趣的稳定性

一个用户可以有很多兴趣，但是其中必有一个比重最大的几个方面。设一共有 m 个兴趣类别，用户 x_i 访问不同的类别值分别为：$x_{i1}, x_{i2}, ..., x_{im}$，设兴趣类别权重最大的前三项的值为向量 $x_{i\max 3} = \max 3(x_{i1}, x_{i2}, ..., x_{im})$，称为主导兴趣[13]。随机选取 3 个用户分别计算出一天、一星期、一个月的每个兴趣类别的访问次数 $x_{i1}, x_{i2}, ..., x_{im}$ 进行分析。

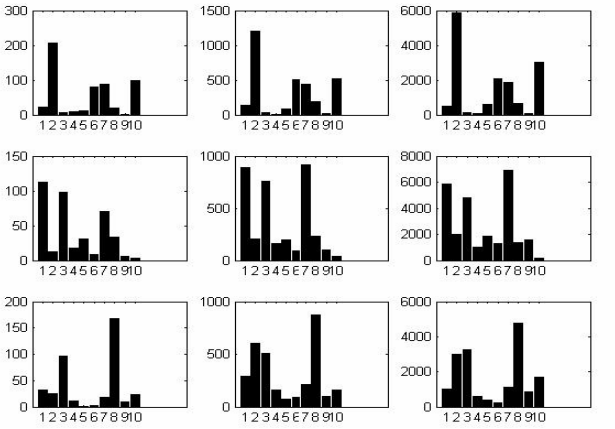

图 8　主导兴趣分析

上图中，每一行为一个用户分别在不同时间段内访问的兴趣类别的数据统计，可以得出结论：每个用户的兴趣集中分布在几个类别中，也就是一段时间内，每个用户访问的网页都集中落在少数几个突出类别中，其余的网页则分布在相对“冷门”类别中。说明用户访问动机有主次之分，具有很强的集中性。随着时间的增长，用户主导兴趣落在相同的三个类别中，趋向于稳定。

5　结语

本文根据分析网页浏览记录，首先对用户数以及访问类别进行统计分析，发现了一些群体行为的普适性统计规律。其次，提出一种改进的基于 HAC 和 k-means 的算法对用户根据访问兴趣进行聚类，能比较准确的聚类，从而挖掘用户的访问模式。最后，发现了用户主导兴趣的稳定性即随着日志的增加，用户的最大兴趣是趋于稳定的。通过以上研究，对建立个性化网站、对用户进行个性化推荐等等都具有很强的实用性。对网页信息挖掘的不够精确，是本文的不足。在将来的工作中，我们将在利用本文统计出来的结果的基础上结合关联规则挖掘以及时间序列分析，来进行个性化推荐技术等方面的研究，从而更好地开发利用网页浏览记录，在理论意义的基础上产生实际应用价值。

参考文献

1 Murray D, Durrel E. Inferring Demographic Attributes of Anonymous Internet Users. Proc. of WEB- KDD’99. San Diego, CA, 1999,August:15–18.

2 Borges J, Levene M. Data Mining of User Navigation Patterns. Proc. of WEBKDD’99. San Diego, CA, 1999, August: 15–18.

3 Perkowitz M, Etzioni O. Towards adaptive Web Sites: Conceptual framework and case study. Artificial Intelligence, 2000,118:245–275.

4 王实,高文,郎金文.在 Web 站点中的知识发现.计算机研究与发展,2002,22(4):482–486.

5 De S, Krishna P. Clustering web transactions using rough approximation. Fuzzy Sets and System, 2004,148(1):131–138.

6 Mitra S. An evolutionary rough partitive clustering. Patterns Recognition Letters, 2004,15(12):1439–1149.

7 宋擒豹,沈钧毅.Web 日志的高效多能挖掘算法.计算机研究

(下转第 109 页)

如果是中心域向普通域发送数据，则随机数由中心域生成，中心域把随机数和计算得出的普通域子钥按照相同的规则生成 64bit 过程密钥，普通域接收到报文信息后，提取出随机数，按照相同的规则把随机数和自身的子钥生成过程密钥解密数据。

在数据交换管理系统的域模型中，只有相互信任的普通域之间才能够进行通信。假定普通域 A 和普通域 B 是中心域 C 下的相互信任的域，则 A 和 B 拥有各自的机器特征信息。A 向 B 发送数据之前，A 首先向中心域申请与 B 通信的过程密钥，中心域接收到申请，验证普通域 A 和 B 是相互信任的域后，用 A 的机器特征，B 的子钥和中心域 C 生成的随机数按照指定的规则生成 64bit 的过程密钥。

⑤$Key_{过程密钥}$ = F(F(B 子钥+A 特征) + C 随机数)

中心域 C 用 A 的子钥把过程密钥 DES 加密后发送给 A，将随机数发送给 B。普通域 A 用解密后的过程密钥加密要发送的数据，普通域 B 接收到数据后，等待从中心域 C 发送过来的随机数，普通域按照相同的规则生成过程密钥解密数据。如果 B 接收到 A 的数据后，在给定的时间限制内没有接收到中心域 C 的随机数，则此次通信作废。B 向 A 发送通信失败的反馈信息，域 A 重新发起与域 B 的通信。

客户端与普通域和非域节点之间通信时，如果客户端作为发起端，则过程密钥为客户端子密钥和客户端生成的随机数通过 DES 算法生成的 64bit 的 DES 密钥，客户端将随机数放在报头中与数据一同发送给作为接收端的普通域或非域节点，接收端提取出随机数和客户端子钥通过 DES 方法生成 DES 密钥解密数据。如果是客户端作为接收端，则构成过程密钥的随机数是由普通域或非域节点生成的，随机数和客户端子钥通过 DES 方法生成 64bit 的 DES 密钥，客户端提取接收到的随机数，取出自身子钥生成 64bit 的 DES 密钥解密数据。

非域节点域与普通域之间通信的过程密钥与客户端与普通域节点之间通信的过程密钥的形成方式类似，由数据传输的发起端生成随机数和非域节点的子钥生成 64bit 的 DES 密钥，随机数同数据一起发送给接收端，接收端将随机数和非域节点的子钥生成 DES 密钥解密数据。

3 结语

密钥管理体系作为系统数据传输安全的保障，是为域管理和系统管理服务的，是流域水环境管理系统重要组成部分。本文从系统域模型管理的需求出发，主要研究密钥管理体系中密钥的生成、存储、分散以及过程密钥的生成与使用等问题。目前密钥管理体系主要是满足密钥管理的基本需求，将进一步研究密钥的动态更新、销毁和跨中心域的密钥管理问题以及使用不同加密算法的密钥管理体系。

参考文献

1 闫鸿宾.密钥管理关键技术研究.南通纺织职业技术学院院报,2010,10(4):5–7.
2 向进.加密算法的安全性分析.吉首大学学报,2011,32(1):42–44.
3 Stallings W.孟庆树,王丽娜,傅建明,等译.密码编码学与网络安全:原理与实践.第 4 版.北京:电子工业出版社,2006.43–63,183–205.
4 庞辽军.秘密共享技术及其应用研究[博士学位论文].西安:西安电子科技大学,2006.
5 杨晓明.一种基于 DES 和 RSA 混合加密算法的研究. 电脑学习,2011,1:1–3.
6 马秀芳,时和平,时晨.基于密钥管理的密钥分发解决方案探析.电信快报,2004,28(14):53–56.

(上接第 68 页)

与发展,2001,38(3):328–333.
8 张宁.群体兴趣网的统计特性研究.上海理工大学学报,2008,30(3):243–246.
9 刘靖,陈福生.结合粗糙集和模糊聚类方法的属性约简算法.计算机应用软件,2004,21(11):72–74.
10 卜东波.分类聚类技术研究.北京:中科院研究生院,2000.
11 行小帅,潘进,李焦成等.基于免疫规划的 k-means 聚类算法.计算机科学,2003,(5):605–610.
12 张惟皎,刘春煌,李玉芳.聚类质量的评价方法.计算机工程.2005,31(20):10–12.
13 郭岩,白硕,杨志峰,张凯.网络日志规模分析和用户兴趣挖掘.计算机学报,2005,28(9):1483–1496.