基于有向带权图的页面聚类算法研究

方 杰，张结魁，周 军
（合肥工业大学，安徽 合肥 230009）

摘 要：聚类算法是数据挖掘中的一个重要的分析工具。Web使用挖掘中的聚类分析一般分为用户聚类和页面聚类。其中页面聚类是指导网站结构离线优化的重要方法。利用有向带权图表示用户的访问会话记录，对建立的有向带权图模型运用聚类算法实现页面聚类。选取真实数据对典型的聚类算法K－means算法、DBSCAN算法和COBWEB算法进行实验。实验结果表明，在选取的数据集范围内，COBWEB算法准确率要高于K－means算法和DBSCAN算法，时间性能与用户访问频率矩阵大小有密切关系。

关键词：有向带权图；聚类算法；页面聚类；K－means算法；DBSCAN算法；COBWEB算法
中图分类号：TP301.6 **文献标识码**：A **文章编号**：1673－629X(2009)09－0049－05

Study on Page Clustering Algorithms Based on Weighted Directed Graph

FANG Jie, ZHANG Jie-kui, ZHOU Jun
(Hefei University of Technology, Hefei 230009, China)

Abstract: Clustering algorithm is an important analytical tool in data mining. Clustering analysis is generally fallen into user clustering and page clustering in Web usage mining. Page clustering is an important methods for guiding for the structure of the site off－line optimization. This paper use weighted directed graph to describe user visit and conversation records, and use clustering algorithms to realize the page clustering by the weighted directed graph mode established. Select the real data carries on the experiment to the typical clustering algorithms K－means algorithm, DBSCAN algorithm and COBWEB algorithm. The experiments results indicate that in the selected data sets, the accuracy rate of COBWEB algorithm is higher than that of K－means algorithm and DBSCAN algorithm, and the time capability is closely related to the size of user visit frequency matrix.

Key words: weighted directed graph; clustering algorithms; page clustering; K－means algorithm; DBSCAN algorithm; COBWEB algorithm

0 引 言

互联网已经成为一个巨大的分布式全球信息服务中心。如何为用户快速、高效、准确地提供他们所需要的，并具有高度相关性的一簇Web页面，已经成为业界研究的主要内容。解决这个问题的一个有效途径就是对页面进行合理的聚类分析，从而更高效地进行Web信息的分类、存储、检索和集成。然而，若要真正实现高效的Web页面聚类，就必须要找出Web页面之间的内部链接关系，特别是页面之间的相似性更是尤为重要。同时，每个页面的重要程度、页面内容以及页面的访问情况也是非常重要的信息。因此可以利用有向带权图来表示用户会话，运用转移概率矩阵描述超链接的重要性，并且在建模的过程中结合站点的拓扑结构。这样不仅可以较好刻画用户的访问行为，还描述了网站的拓扑结构以及站点的结构特征数据。基于有向带权图的用户访问模型的有效性在文献[1]已经得到证明，故对于有向带权图模型的有效性试验不再单独验证。

Web使用挖掘中的聚类分析一般分为用户聚类和页面聚类两类。用户聚类是对用户的会话进行分析，根据用户的访问行为，寻找行为模式相似的用户[2]。将这些用户分为一组，则组内的用户可以共享一个用户配置文件，即该组用户访问频率较高的页面集合。页面聚类的分析对象为单个的页面。聚类的结果揭示的是页面之间的相互关系。对于事务数据库，用户聚类可以被称为纵向聚类，而页面聚类是横向聚

收稿日期：2009－01－18；**修回日期**：2009－03－28
基金项目：国家自然科学基金项目(70672097)；国家自然科学基金重点项目(70631003)
作者简介：方 杰(1983－)，男，硕士研究生，研究方向为Web挖掘；张结魁，讲师，博士，研究方向为数据挖掘。

类。实际应用中,用户聚类结果可以应用于网页的个性化推荐技术,即在线优化;而页面聚类结果可以建立用户兴趣集,指导网站结构的离线优化。

聚类算法是数据挖掘中的一个重要的分析工具。目前已有许多聚类算法,不同的聚类算法可能使用不同的聚类准则、相似度度量标准和问题求解方法。但聚类问题的原始定义是一致的(即是把数据集中的数据点分成多个簇,在同一簇中的数据点之间尽可能相似,而不同簇中的数据点尽可能不相似),因而不同的聚类算法的算法框架是基本相同的。文中基于有向带权图模型用户访问行为,利用真实数据,选取三种不同类型的典型聚类算法实现页面聚类。对聚类算法实现页面聚类问题进行比研究。

1 基本概念

在描述算法之前,首先给出相关的概念和定义。

定义 1:有向带权图结构模型。设 $V = \{V_1, V_2, \cdots, V_n\}$ 是一个非空集合,$E = \{e \mid e = < V_i, V_j > V_i, V_j \in V\}$,其中 $< V_i, V_j >$ 是有序对,方向由 V_i 节点指向 V_j 节点,$W(E(G)) = \{w(e_h) \mid \forall e_h \in E(G)\}$,称 (V,E,W) 为有向带权图,记为 $G(V,E,W)$,V 中的元素为用户访问的页面,V 为页面集合,E 为超链接集合,W_{ij} 为相应链接的权值的集合。单个用户的每一次会话可以抽象成一个图,用 G_i 表示。

定义 2:图数据库 GDB。所有用户的访问历史记录,可以看作是一个图数据库,用 GDB 表示,GDB = $\{GID, G_i\}$,其中 $G_i = (V(G_i), E(G_i), W(G_i))$,表示单个用户的每一次会话。$N$ 表示图库的大小。GID 是 G_i 在 GDB 中的 ID,表示用户会话 ID。

定义 3:权值 W_{ij}。把权值 W_{ij} 定义为页面 i 和页面 j 的访问相关性,采用如下定义形式:

$$W_{ij} = \frac{S_{ij}}{\mathrm{Max}(S_i, S_j)}$$

式中,S_{ij} 表示所有用户会话中同时包含页面 i 和页面 j 的用户会话的个数,S_i 表示包含页面 i 的用户会话个数,S_j 表示所有用户会话中包含页面 j 的用户会话个数。

网站中的网页一般可分为索引页和内容页,索引页与包含具体内容信息的内容也不同,其主要目的是用于浏览导航。因此索引页可能经常被用户访问,但并不意味其中包含用户真正感兴趣的信息。采用最大值方法可有效避免由于对索引页的频繁访问而引起的边权值过大问题。对于页面 i 和页面 j 为同一页面的情况,设置访问相关性为 $W_{ij} = 0$。

定义 4:Hamming 距离。设 $X, Y \in \{0,1\}^n$,$n \geqslant 1$,那么,X, Y 间的 Hamming 距离 $H_d(X, Y)$ 定义为:

$$H_d(X, Y) = \sum_{i=1}^{\|X\|} | X_i - Y_i |$$

2 聚类算法及实现

聚类分析是数据挖掘中的一个很活跃的研究领域,将具体或抽象的集合分组成为由类似的对象组成的多个类的过程称为聚类。由聚类所生成的簇(cluster)是一组数据对象的集合。通常,聚类分析的目标是在某些限制条件下,将组内成员之间的差距最小化,将组间差距最大化。聚类分析的方法主要有五类:划分聚类方法、层次聚类方法、基于密度聚类方法、基于网格聚类方法和基于模型聚类方法。

基于划分方法的代表聚类算法有:k - means, EM, k - medoids, CLARA, CLARANS 等。常见的 k - medoids 算法有 PAM 算法、CLARA 算法、CLARANS 算法。

基于层次方法的代表聚类算法有 BIRCH, CURE, ROCK, Chameleon, AMOEBA, COBWEB, Clustering with RandomWalks 算法等。

基于密度方法的代表聚类算法有 DBSACAN, DBCLASD, OPTICS, DENCLUE 等。

基于网格方法的代表聚类算法有 STING, Wave Cluster, CLIQUE, MAFIA, OptiGrid 等。

基于模型方法的代表算法有 COBWEB 等。

迄今为止,Web 日志挖掘中的页面聚类研究主要有两个方向。Chen M S 等人[3]首先将数据挖掘技术应用于 Web 服务器日志文件,以期发现用户浏览模式。他们提出最大前向引用序列 MFR 的概念,并用它将用户会话分割成一系列的事务,然后采用与关联规则相似的方法挖掘频繁浏览路径。而 Han 等人[4]则根据 Web 日志建立数据立方体,然后对数据立方体进行数据挖掘和 OLAP Minnesota 大学的 WEBMINER 系统提出了一种通用的 Web 日志挖掘的体系结构,该系统能自动从 Web 日志中发现关联规则和序列模式等。这两种方法均要进行用户识别和会话识别,而用户识别和会话识别都要受到本地浏览器缓存、防火墙和代理服务器等的影响。实现时,基于 Web 事务的方法采用复杂的 Hash 数据结构存储候选项集,对其进行维护和搜索还会增加额外的负载。

聚类算法的算法框架是基本相同的,因此将不同聚类算法应用于页面聚类。具体实验中,每种算法参数可能代表不同的含义,但是聚类结果的思想是一致的。算法实现过程中以邻接矩阵 M 存储有向图 G 的信息,方向由图中顶点下标表示,以矩阵元素

$M[i][j]$保存所有用户会话中同时包含页面 i 和页面 j 的用户会话的个数。矩阵对角线元素 $M[i][i]$ 保存所有用户会话中包含页面 i 的用户会话的个数。采用这种方式,可以得到网站内任意页面之间的边权值。

下面简述选用的 K－means 算法、DBSCAN 算法和 COBWEB 算法。

2.1 K－means 算法

K－means 算法用簇重心作为簇代表,因而它只能发现"类球形"的簇。K－means 用误差平方和最小作为聚类收敛准则。K－means 的主要步骤:从数据集中随机选取 K 个数据点分别作为 K 个子簇的簇代表(簇重心);计算其余点分别到 K 个簇代表的距离,并将它们分配到与之最近的簇代表所表示的簇中去,重新计算 K 个簇的簇代表,并用它们来替换原来的簇代表,再重新执行该步骤,直至平方误差不再减少。因此,它的时间复杂度为 $O(nkl)$,其中 n 为数据集数据点的数量(以下几个算法中的 n 与此同含义),k 为簇的数量,l 为算法收敛时已迭代的次数[5]。K－means 对噪音和异常点敏感,因为即使少数这样的数据点对平均值的影响也相当大。

2.2 DBSCAN 算法

DBSCAN(density based spatial clustering of applications with noise)算法[6]是一个基于密度的聚类算法。它定义簇为密度相连的数据点的最大集合,因此该算法具有较强的发现任意形状簇的能力。它通过检查数据集中每个点的 ε—邻域来找核心对象,每个核心对象的邻域构成一个子簇。然后在这子簇中寻找核心对象直接密度可达的对象,并将该对象加入子簇中形成新的子簇。再重新执行该数据变换步骤,直至没有新的对象加入到子簇中[6]。该数据变换过程中可能涉及子簇合并,因此它的时间复杂度为 $O(n^2)$。如果采用空间索引,它的时间复杂度则为 $O(n\log n)$。

2.3 COBWEB 算法

COBWEB 算法是一种流行的简单增量概念聚类算法。它的输入对象用分类属性－值对来描述,以一个分类树的形式创建层次聚类。分类树的每个节点对应一个概念,包含该概念的一个概率描述,概述被分在该节点下的对象。在分类树某个层次上的兄弟节点形成了一个划分。为了用分类树对一个对象进行分类,采用了一个部分匹配函数沿着"最佳"匹配节点的路径在树中向下移动。寻找可以分类该对象的最好节点。这个判定基于将对象临时置于每个节点,并计算结果划分的分类效用。产生最高分类效用的位置应当是对象节点一个好的选择。但如果对象不属于树中现有的任何概念,则为该对象创建一个新类。

由于 CORWEB 算法是它基于这样一个假设:在每个属性上的概率分布是彼此独立的。而属性间经常是相关的,所以这个假设并不总是成立。此外,聚类的概率分布表示使更新和存储聚类相当昂贵。因为时间和空间复杂度不只依赖于属性的数目,而且取决于每个属性的值的数目,所以当属性有大量的取值时情况尤其严重。而且分类树对于偏斜的输入数据不是高度平衡的,它可能导致时间和空间复杂性的剧烈变化[5]。COBWEB 算法不适用于聚类大型数据库的数据。

2.4 页面聚类算法实现

算法过程如下:

(1)判断当前访问结点与引用结点(即会话中前一次访问的结点)是否属于不同的聚类,如果属于不同的聚类且两结点之间的边权值大于最小阈值 MinFreq(取值大小为边权值的平均值并在不同算法实验中保持一致),则合并两个聚类;

(2)以当前访问结点所属聚类作为初始聚类,进行重新划分。通过权值大于 MinFreq 的边,由当前访问结点作为起始结点对初始聚类进行广度优先遍历,将被访问节点的聚类标识设置为起始结点;

(3)如果初始聚类中存在未被访问的结点,则以其中任意结点作为起始结点对剩余结点重新进行广度优先遍历,将被访问结点的聚类标识设置为起始结点。重复第(3)步,直到初始聚类中所有结点被访问为止。

算法描述如下:

输入:邻接矩阵 M,聚类标识数组 ClusterID,当前访问页面 i,引用页面标识 j。

方法:

```
1)W_ij;
2)If((W_ij > MinFreq)&(ClusterID[i] ≠ ClusterID[j])) Then
3) 合并页面 i 和页面j 所在的聚类;
4)endif 将页面 i 所属聚类中除页面j 之外的节点组成集合 NC;
5)h = i; //h 为遍历起始点
6)ClusterID[i] = i;
7)clusterflag = i;//clusterflag 为区分不同聚类的标志
8) 构建堆栈 S 并初始化为空;
9)while(h ≠ NULL)
10)for(NC 中每一个边权值 W_h, n 大于 MinFreq 的节点 n)
11)ClusterID[n] = clusterflag;
12)remove(NC, n);
13)push(S, n);end for
14)if(S ≠ NULL) then
15)h = pop(S);
16)else if(NC不为空) then
17)h = pop(NC);
```

18)clusterflag = h;
19) else h = NULL;
20)return

算法中借助堆栈结构完成对聚类结果的遍历。通过对最小阈值取值的适当选择,可以将页面聚类中包含的页面数量控制在合理的范围内。经过 PageCluster 算法的处理会产生一个新的聚类标识数组 ClusterID。

3 实 验

为研究不同算法应用于页面聚类的结果,通过实验进行比较。实验环境为 Pentium 4 2.8GHz 的 PC 机,内存为 512M,硬盘 80GB。操作系统为 Windows XP professional。所有算法均用 MATLAB7.0 开发。

3.1 数据准备及参数选择

聚类之前,首先要对站点日志进行预处理,从中识别出用户事务。先进行数据清洗,即将日志转化为适合数据挖掘的可靠的精确的数据,删除日志中与数据挖掘不相关的冗余项。日志记录包括用户 IP 地址、用户 ID、用户请求访问的 URL 页面、请求方法、访问时间、传输协议、传输的字节数、错误代码等属性,而与数据挖掘相关的只有用户 IP 地址、用户请求访问的 URL 页面,其他属性可以去掉。URL 页面中除了用户关心的正文外,往往还有图像、声音、视频等辅助信息。挖掘日志的目的是找出用户的共同访问模式,关于辅助信息的记录是无用的,可以删除。这可通过检查 URL 的后缀来实现,后缀为 gif, jpeg, cgi…等无关请求的记录都去掉。当然,如果网站内容主要是图片的话,则须另行考虑,然后在此基础上进行会话识别[7]。

选用的实验数据是合肥工业大学管理学院网站服务器的日志文件,有 3 万多条记录。经过日子的预处理和页面的归纳化处理[8],得到用户访问频率矩阵[9](96×54),其中有 96 个页面向量,37 个用户向量。在这个矩阵的不同子集上进行实验。

数据集的大小(矩阵的行×列):96×54,82×40,60×34,48×23,36×8,分别对应 1,2,3,4,5 五个不同的实验用例。

聚类中最重要过程就是对象之间相似度分析。不同的聚类算法中页面相似度计算方法是不同的。基于统一实验考虑,页面相似性度量是根据 Hamming 距离[10]进行的,对于 $\forall M[i,j] = 1$,然后,计算向量间的 Hamming 距离。Hamming 距离越小,其相似程度越高。同时引入页面转移概率 P,用于 DBSCAN, COBWEB 算法中作为形成新类簇的标准。试验中,数据点数 $K = 3$,最小阈值 MinFreq = 0.42,转移概率 P = 0.35。

3.2 结果分析

聚类算法有很多,经过比较,在实验中使用效果较好的 K - means 算法、DBSCAN 算法和 COBWEB 算法。

图 1 所示为三种算法的准确率比较。测试算法的准确性,首先对数据预处理结果进行分析,得到相关 Web 页面。然后采用三种不同算法进行处理,并根据算法结果和分析结果计算相应的准确度,得到图 1 所示准确度曲线。图 1 表明:在五个不同的实验用例下,页面聚类的准确率最高可达 71%;COBWEB 算法的准确率整体要高于 K - means 算法和 DBSCAN 算法;而 K - means 算法与 DBSCAN 算法的准确率相差不大且互有高低。主要原因是,CORWEB 算法不需要用户输入参数来确定分类的个数,它可以自动修正划分中类的数目,因而聚类的准确率较高。而 K - means 算法与 DBSCAN 算法在整个聚类过程中使用固定的参数,使得对于真实环境数据集的聚类,往往其聚类的效果不好。例如 K - means 算法,由于其定义的密度的传递性质,往往将绝大多数的数据点都聚集在非常少的几类中(通常是一类),因而准确率比较低。

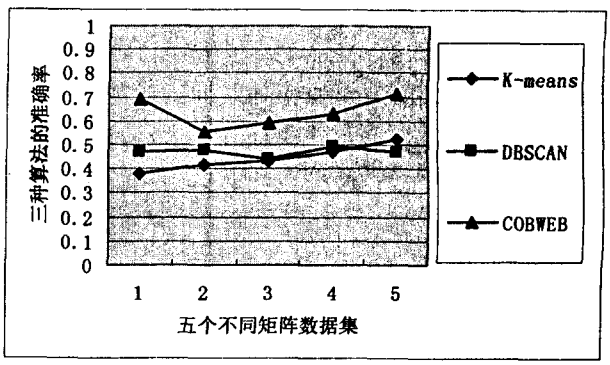

图 1 三种算法的准确率比较

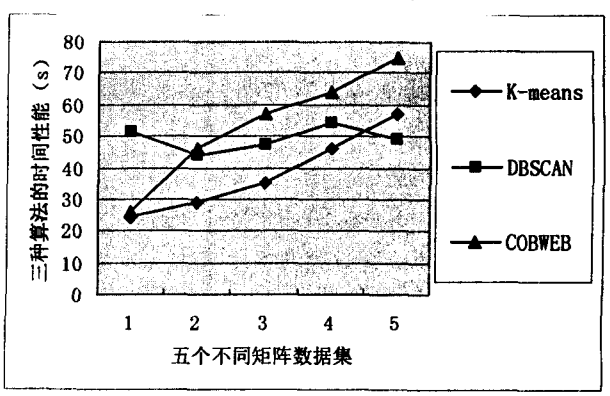

图 2 三种算法的时间性能比较

图 2 所示为三种算法的时间性能比较。从运行时间上来看,K - means 算法和 COBWEB 算法的时间复杂度与实验用例的大小即计算的页面数有直接关系。

所以随着实验用例中页面数的增加,时间复杂度也迅速增加。而 DBSCAN 算法的时间复杂度与实验用例的大小变化并不明显,只是在一定的水平上下波动。主要原因是 DBSCAN 算法是一种通过对局部密度分析,将相邻点聚集在一起的聚类算法[7]。在整个算法进行过程中,它只对数据库进行一次扫描。因此只要 DBMS 对相邻点的查询效率很高,DBSCAN 的效率将会很高。而 DBMS 的查询效率目前已经完全满足这一点[11]。

4　结束语

采用有向带权图表示用户会话,使得挖掘的结果包含更多的信息。运用不同的聚类算法实现页面聚类,可以为网站结构的离线优化提供更多的研究依据。实验结果表明,COBWEB 算法在准确率和时间性能上总体优于 K-means 算法和 DBSCAN 算法。但是实验数据并不具有权威性。同时,文中的有向带权图模型是否适用于大数据量,是否能应付异常数据等还需要继续研究。

参考文献:

[1] 周　军,余智学,姜元春. 基于有向带权图的 Web 用户浏览行为模型[J]. 情报理论与实践,2008,31(5):795-798.

[2] 杨怡玲,管旭东,尤晋元. 基于页面内容和站点结构的页面聚类挖掘算法[J]. 软件学报,2002,13(3):467-469.

[3] Chen M S, Park J S, Yu P S. Data nining fo r path traversalpatterns in a Web environment[J]. In: Proc of the 16th Int'l Conf on Distributed Computing Systems. Hong Kong: [s.n.], 1996:385-392.

[4] 宋擒豹,沈钧毅. Web 日志的高效多能挖掘算法[J]. 计算机研究与发展,2001,38(3):328-333.

[5] 孙吉贵,刘　杰,赵连宇. 聚类算法研究[J]. Journal of Software,2008,19(1):48-61.

[6] Ester M, Kriegal H P, Sander J. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// In: Simoudis, Evangelos, Han Jia-wei, Fayyad U M. KDD'96 Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. [s.l.]: AAAI Press, 1996.

[7] 王　庆,周俊梅,吴红伟,等. XML 文档及其函数依赖到关系的映射[J]. 软件学报, 2003(7):1275-1281.

[8] 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.

[9] 郭　岩. 用户兴趣空间的 Web 页面聚类[J]. 微电子与计算机,2003(8):10-14.

[10] 宋摘豹,沈钧毅. Web 页面和客户群体的模糊聚类算法[J]. 小型微型计算机系统,2001,22(2):229-231.

[11] 苏　中,马少平,杨　强,等. 基于 Web-Log Mining 的 Web 文档聚类[J]. 软件学报,2002,13(1):99-104.

(上接第 48 页)

阶段。人们识别基于 P2P 协议报文的困难变得越来越大,一些传统的识别方法,比如端口识别法、应用层协议内容识别法的识别效果已经不是很好了。文中提出的方法能够检测使用动态端口,加密传输的 P2P 流量,在算法的执行效率上也有一定程度的提高。挖掘更深层次的流特征,把各种流特征有机地结合起来,运用人工神经网络的思想,对 P2P 流量进行识别,是最新的研究方向。

参考文献:

[1] 李江涛,姜永玲. P2P 流量识别与管理技术[J]. 电信科学,2006(7):57-60.

[2] 石　萍,陈贞翔,荆　山. 基于对等特征的 P2P 流量识别方法[J]. 中国教育网络, 2007(2):36-38.

[3] Sen S, Spatscheck O, Wang Dongmei. Accurate, scalable in-network identification of p2p traffic using application signatures[C]// In WWW '04: Proceedings of the 13th international conference on World WideWeb. New York, NY, USA: ACMPress, 2004:512-521.

[4] Karagiannis T, Broido A, Faloutsos M. Transport layer identification of p2p traffic[C]// In Proceedings of the 4th ACM SIGCOMM conference on Intrnet measurement. New York, USA: ACM Press, 2004:121-134.

[5] 蒋海明,张剑英,王青青,等. P2P 流量检测与分析[J]. 计算机技术与发展, 2008,18(7):116-119.

[6] 温　超,郑雪峰. 基于流量分析的 P2P 协议识别方法的研究[J]. 微计算机应用,2007(7):714-717.

[7] Constantinou F, Mavrommatis P. Identifying Known and Unknown Peer-to-Peer Traffic[C]// Network Computing and Applications, 2006. NCA 2006. Fifth IEEE International Symposium onVolume, 2006:93-102.

[8] Bolla R, Canini M, Rapuzzi R. On the Double-Faced Nature of P2P Traffic Department of Communication[C]// In Proceedings of the Sixteenth Euromicro Conference on Parallel, Distributed and Network-Based Processing (PDP'08). [s.l.]:[s.n.],2008:524-530.

[9] 柳　斌,李之棠. 一种基于流特征的 P2P 流量识别实时识别方法[J]. 厦门大学学报,2007(11):56-60.