计算机科学2004Vol. 31№. 12

基于模糊聚类的 Web 日志挖掘

李桂英　李吉桂
(华南师范大学　广州510631)

摘　要　本文使用模糊聚类的方法对 Web 日志进行数据挖掘，实现用户聚类和页面聚类，并设计与实现了一个基于模糊聚类的 Web 日志挖掘系统。

关键词　模糊聚类，数据挖掘，Web 日志挖掘，用户聚类，页面聚类

Web Log Mining Based on Fuzzy Clustering

LI Gui-Ying　LI Ji-Gui
(South China Normal University, Guangzhou 510061)

Abstract　This paper proceeds Web log mining by using the technique of fuzzy clastering to realize customers clustering and Web pages clustering. A Web log mining system based on fuzzy clustering is designed and realized in this paper.

Keywords　Fuzzy clastering, Data mining, Web log mining, Customers clustering, Web pages clastering

1　引言

Web 是信息发布、交互及获取的主要工具，Web 上的信息量正以惊人的速度不断增长，如何针对 Web 页面内容、页面结构、用户访问信息等数据，应用数据挖掘技术挖掘出有用的知识，更好地为用户服务，已经成为国际上的热门研究问题。

Web 挖掘(Web mining)就是从大量的 Web 资源中提取隐含的、未知的、对决策有潜在价值的知识和规则的过程，它把数据挖掘技术应用于对 Web 资源的挖掘上，是一个新兴的研究领域。根据挖掘对象的不同，Web 挖掘可分为 Web 内容挖掘、Web 结构挖掘和 Web 使用挖掘三类。Web 内容挖掘是指对 Web 页面内容进行挖掘，是从数以百万计的 Web 资源中发现知识的过程，主要包括文本信息挖掘和多媒体信息挖掘；Web 结构挖掘是指对 Web 页面之间的结构进行挖掘，是从 Web 的组织结构以及引用和被引用间的链接关系中推理知识的过程；Web 使用挖掘也称 Web 日志挖掘，是从 Web 访问日志中抽取知识的过程，主要算法有路径分析、关联规则和序列模式的发现、聚类分析等。

Web 是一个无统一结构、无标准查询语言和数据模型的分布式信息系统，对其进行挖掘是困难的。然而，每一个提供信息资源的服务器都有一个结构化较好的记录集，即 Web 访问日志。Web 访问日志是 Web 服务器用以记录用户访问该网站各页面情况的文件，用户每访问一个页面，日志文件就会增加一条记录。随着 WWW 技术的快速发展和时间的积累，Web 服务器日志文件越来越大，其中包含的用户信息也越来越多，对这些数据进行分析，有助于了解网络的通信状况、用户的访问动机、访问模式和访问趋势，协助网站管理者优化网站结构，提高信息服务质量。由于 Web 站点的内容是动态变化的；用户浏览 Web 时又往往没有明确的目的，具有模糊性和不确定性，因此，本文采用模糊聚类的方法对 Web 日志进行数据挖掘，实现用户聚类和页面聚类。

2　Web 日志挖掘中的用户聚类和页面聚类

Web 日志挖掘的主要目的是利用 Web 日志文件所记载的信息来发现用户访问模式，以改进信息服务质量和网站性能。Web 日志挖掘可分成原始数据收集、数据预处理、模式发现和模式分析四个阶段，如图1所示。

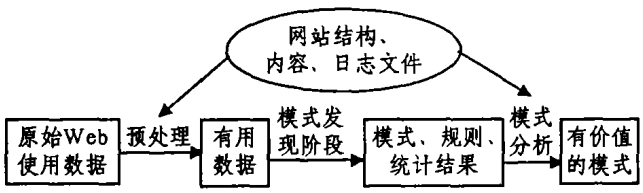


图1　Web 日志挖掘过程

2.1　原始数据收集

原始 Web 使用数据可从服务器端、代理服务器端或客户端获得，本文的原始数据主要来自 Web 服务器端的日志文件。不同 Web 服务器的日志文件格式并不完全相同，但通常都包括用户的 IP 地址(或域名)、访问日期和时间、访问方法(GET 或 POST)、被请求页面的 URL、HTTP 版本号、传输字节数、访问结果(成功、失败、错误)、引用页的 URL 等信息。

2.2　数据预处理

由于 Web 日志文件中存储的仅仅是原始的访问资料，这些数据并非是专门用来做数据挖掘的，因此在进行数据挖掘之前必须对其进行数据清理、筛选和转换等预处理。

在 Web 日志文件中，只有用户 IP 地址、访问日期和时间、被请求页面的 URL 在 Web 日志挖掘时有价值，其它数据均可去掉；在 Web 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，Web 服务器日志会将用户访问单个 Web 页记录为访问文本、图像、声音、视频等多个文件。但在大多数情况下，用户一般不会显式地请求页面上的图像、

李桂英　讲师，硕士，主要研究方向：数据挖掘、计算机网络与通讯、远程教育等。李吉桂　教授，硕士导师，主要研究方向：数据挖掘、人工智能、计算机网络与通讯、远程教育等。

声音、视频等类型文件，因此，应把 URL 的后缀名为 JPG、GIF、MP3、MPG 等的记录删除。但是，在具体应用中，应根据站点的具体情况灵活处理，例如，对一个主要包含图形文件的站点，日志中 URL 的后缀名为 JPG、GIF 等的记录可能代表了用户的显式请求，此时就不能将这类记录删除；另外，日志文件中的错误日志记录和不完整的用户行为记录也应该删除。经过以上处理后，把日志文件中有用的数据转化为一个格式如表1所示的关系数据库 RZ. DBF。

表1 RZ. DBF 的格式

DATE	TIME	IP	URL
2003-06-16	06:24:23	64. 68. 82. 28	/index. html
2003-06-16	06:27:12	64. 68. 82. 28	/6076-6. html
2003-06-16	06:27:12	64. 68. 82. 28	/ads/banner. js
2003-06-16	06:29:43	210. 187. 125. 251	/index. html

2.3 模式发现

模式发现可以采用众多领域(如统计学、数据挖掘、机器学习、模式识别等)的方法和算法，用于 Web 领域的主要算法有统计分析、关联规则、聚类、分类、序列模式等。

对于一个特定的 Web 站点，其拓扑结构、页面个数都是已知的；虽然同一用户在不同时期可能有不同的浏览模式，但其长期趋势应该是稳定的。因此，分析一定时期内用户的访问信息便可以了解用户的访问模式，实现用户聚类和页面聚类。用户聚类主要是把用户划分成若干组，具有相似浏览模式的用户分在一组，这类知识在电子商务和为用户提供个性化服务等应用中特别有用；页面聚类则可以找出具有相关内容的网页组，这对网上搜索引擎和调整页面结构等应用很有用。

2.3.1 *用户聚类* 假设某网站共有 n 个页面，用 Y 表示该网站各页面的 URL 的集合，$Y=\{Y_1,Y_2,\cdots,Y_n\}$；在某段时间内共有 m 个用户访问该网站，用 X 表示这些用户的集合，$X=\{X_1,X_2,\cdots,X_m\}$，此时，X 为待分类对象。对数据库 RZ. DBF 进行统计，计算出每一用户 $X_i(i=1,2,\cdots,m)$在这段时间内访问各页面 $Y_j(j=1,2,\cdots,n)$的次数 $T(X_i,Y_j)$，以及每一用户 X_i 在这段时间内访问该网站所有页面的总次数 $\sum_{k=1}^{n}T(X_i,Y_k)$，然后用 $\dfrac{T(X_i,Y_j)}{\sum_{k=1}^{n}T(X_i,Y_k)}$ 表示用户 X_i 和页面 Y_j 之间的关联度，并用它作为特征数据，得到一个 m 行 n 列的原始数据矩阵 A：

$$A=\begin{bmatrix} \dfrac{T(X_1,Y_1)}{\sum_{k=1}^{n}(X_1,Y_k)} & \dfrac{T(X_1,Y_2)}{\sum_{k=1}^{n}T(X_1,Y_k)} & \cdots & \dfrac{T(X_1,Y_n)}{\sum_{k=1}^{n}T(X_1,Y_k)} \\ \dfrac{T(X_2,Y_1)}{\sum_{k=1}^{n}T(X_2,Y_k)} & \dfrac{T(X_2,Y_2)}{\sum_{k=1}^{n}T(X_2,Y_k)} & \cdots & \dfrac{T(X_2,Y_n)}{\sum_{k=1}^{n}T(X_2,Y_k)} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{T(X_m,Y_1)}{\sum_{k=1}^{n}T(X_m,Y_k)} & \dfrac{T(X_m,Y_2)}{\sum_{k=1}^{n}T(X_m,Y_k)} & \cdots & \dfrac{T(X_m,Y_n)}{\sum_{k=1}^{n}T(X_m,Y_k))} \end{bmatrix}$$

根据原始数据矩阵 A，用模糊聚类方法对用户进行聚类，其主要算法见图2。

2.3.2 *页面聚类* 页面聚类的算法与用户聚类的算法相似。此时，待分类的对象为 $Y=\{Y_1,Y_2,\cdots,Y_n\}$。对数据库 RZ. DBF 进行统计，计算出每一页面 $Y_i(i=1,2,\cdots,n)$在这段时间内被各用户 $X_j(j=1,2,\cdots,m)$访问的次数 $T(Y_i,X_j)$，以及页面 Y_i 在这段时间内被所有用户访问的总次数 $\sum_{k=1}^{m}T(Y_i,X_k)$，然后用 $\dfrac{T(Y_i,X_j)}{\sum_{k=1}^{m}T(Y_i,X_k)}$ 表示页面 Y_i 和用户 X_j 之间的关联度，并用它作为特征数据，得到一个 n 行 m 列的原始数据矩阵 A：

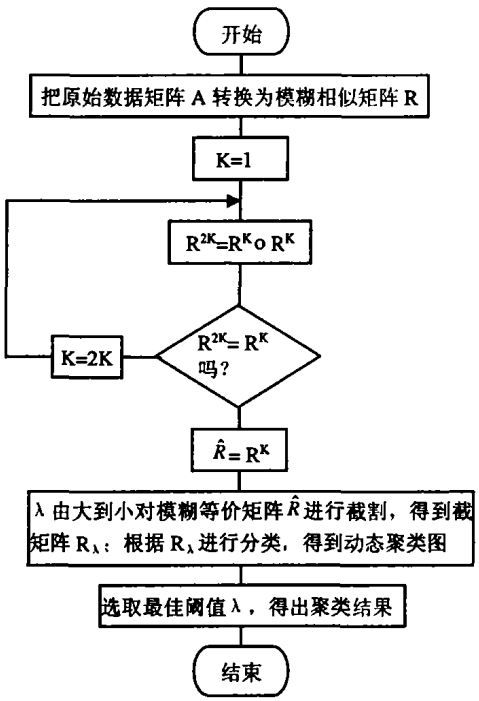


图2 模糊聚类流程图

$$A=\begin{bmatrix} \dfrac{(T(Y_1,X_1)}{\sum_{k=1}^{m}T(Y_1,X_k)} & \dfrac{T(Y_1,X_2)}{\sum_{k=1}^{m}T(Y_1,X_k)} & \cdots & \dfrac{T(Y_1,X_m)}{\sum_{k=1}^{m}T(Y_1,X_k)} \\ \dfrac{T(Y_2,X_1)}{\sum_{k=1}^{m}T(Y_2,X_k)} & \dfrac{T(Y_2,X_2)}{\sum_{k=1}^{m}T(Y_2,X_k)} & \cdots & \dfrac{T(Y_2,X_m)}{\sum_{k=1}^{m}T(Y_2,X_k)} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{T(Y_n,X_1)}{\sum_{k=1}^{m}T(Y_n,X_k)} & \dfrac{T(Y_n,X_2)}{\sum_{k=1}^{m}(Y_n,X_k)} & \cdots & \dfrac{T(Y_n,X_m)}{\sum_{k=1}^{m}T(Y_n,X_k))} \end{bmatrix}$$

然后再用类似的方法对页面进行模糊聚类。

3 基于模糊聚类的 Web 日志挖掘系统

用上述方法进行用户聚类和页面聚类，其操作过程相当繁杂和机械，手工计算难以胜任，为了解决这个问题，我们设计并实现了一个基于模糊聚类的 Web 日志挖掘系统。该系统以 Windows98 为系统平台，以优秀的可视化程序设计语言 Delphi 7.0为开发工具，其系统功能如图3所示。

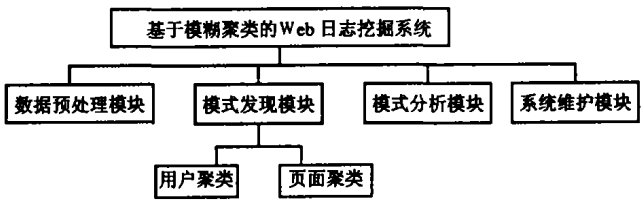
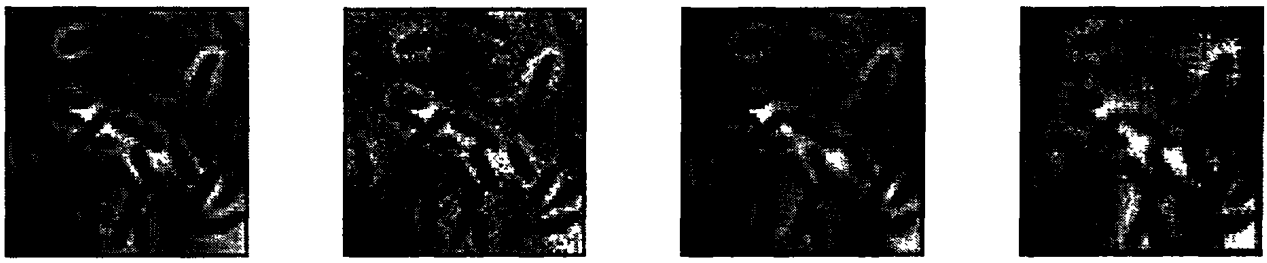


图3 系统功能结构图

(下转第153页)

这个推理在噪声均值为0的时候非常合理。



(a) (b) (c) (d)

(a) 原始图像 SNR=31.511 (b) 泊松噪声模糊图像 SNR=13.032
(c) wiener 滤波器降噪结果 SNR=22.65 (d) 交叉检验降噪结果 SNR=14.7312

图3 交叉检验方法复原泊松噪声图像示意图

实验中的另一个尝试是在得到训练停止点以后，立即对整个图像进行训练，实验证明，此时的网络与交叉检验时的网络有相近的性质。其学习步长为交叉检验训练中步长的十分之一，停止点比交叉检验方法中的停止点稍大(较多的训练步数)，这个时候能获得比较满意的效果，这种方法中不用进行图像的配准。

实际算法还可以利用 FFT 技术进一步优化。本文中径向基函数矩阵是一个 Toeplitz 矩阵，如果利用 Toeplitz 矩阵的快速算法，则可以进一步提高网络的训练速度，同时由于快速算法中只需要存储其基函数矩阵的第一列，因此还可以很大地减小算法对存储空间的需求。有关 Toeplitz 矩阵的快速算法的基础知识可以参见有关数学文献，笔者在文[5]的基础上发展了一种用 FFT 实现的 Toeplitz 快速算法。

结论 本文提出了基于 RBFNN 的交叉检验的图像降噪技术，描述了 RBFNN 应用于图像处理的基本原理与方法，并较详细介绍了交叉检验的原理和算法。实验证明所采用的算法可以很好地进行盲目图像降噪，对于信噪比较高的图像，其不但能有效抑制噪声，还能按需要保存图像细节(损失降噪效果)，其降噪效果较 wiener 滤波器要好，但是对于信噪比较低的图像，该方法不易收敛到极小点而不能获得满意的效果。最后，本文还对此方法进行了简单优化，给出了部分优化原理和经验数据，相信对改进和发展交叉检验方法的进一步研究有一定的帮助。

实验表明，如果基于 RBFNN 的交叉检验图像降噪技术与其他降噪技术结合，可以更好地提高降噪效果，甚至在降噪的同时提高图像分辨率。

本文提到的这种技术基本不需要噪声的相关信息，完全利用退化图像自身的特点进行降噪，因此具有较好的适用性。本文将交叉检验方法应用到图像的表示上，这种技术解决了只有极少信号量，却要进行建模的难题，这种思想可以应用于盲目信号恢复和盲目系统辨识等其他领域，有着广阔的发展前景。

参 考 文 献

1 Bishop C M. Neural Networks for Pattern Recognition. Oxford University Press, Oxford, 1995
2 Guo P, Lyu M R. A Pseudo inverse Learning Algorithm for Feed forward Neural Networks with Stacked Generalization Application to Software Reliability Growth Data. Neuro computing, 2004, 56: 101～121
3 Park J, Sandberg I W. Universal approximation using radial-basis-function networks. Neural Computation, 1991, 3: 246～257
4 Park J, Sandberg I W. Approximation and radial basis-function networks. Neural Computation, 1993, 5: 305～316
5 邹谋炎. 反卷积和信号复原. 北京：国防工业出版社，2001

(上接第131页)

结束语 随着 Web 技术的发展，各类电子商务网站风起云涌，建立电子商务网站的关键是如何使其有效益，要有效益就必须吸引客户，提高能带来效益的客户忠诚度。电子商务网站每天都可能有成千上万的在线交易，生成大量的数据库记录和日志文件记录，如何分析和挖掘这些数据，充分了解客户的喜好和购买模式，从而设计出满足不同客户群体需要的个性化网站，是电子商务网站取得效益的关键。

作为电子商务中一个重要环节的客户关系管理(CRM)，是一种旨在改善企业与客户之间关系的新型运作机制，其研究和应用已成为全球关注的热门话题。本系统的聚类结果可作为客户关系管理的依据。我们利用本系统对某电子商务网站的服务器日志进行数据挖掘，取得了良好的效果。根据用户聚类的结果，销售商可以有目标地展示自己的产品，将对客户群体最有价值的信息推到用户眼前，对特定的客户群体提供有针对性的服务。一方面，发送销售邮件时不再是千篇一律，而是根据客户群体的需求，对不同的客户群体发送不同的销售邮件。实践表明，由于销售邮件正好满足客户个性需求，客户喜欢销售商通知他购买他真正感兴趣的东西，从而节约个人时间，获得满意的服务；另一方面，销售商可为不同的客户群体提供个性化的界面。根据客户群体的需求，对不同的客户群体定制不同的广告，实践表明，这能大大地提高广告的单击率(上站人数与单击该广告次数的比率)。另外，根据页面聚类的结果，网站管理人员可改善网站结构，更好地为客户服务。

参 考 文 献

1 汪培庄，韩立岩. 应用模糊数学[M]. 北京：北京经济学院出版社，1989
2 刘同明. 数据挖掘技术及其应用[M]. 北京：国防工业出版社，2001
3 朱明. 数据挖掘[M]. 合肥：中国科学技术大学出版社，2002
4 党齐民，李晓聪. 电子商务与信息处理技术[M]. 上海：上海人民出版社，2002
5 Han Jiawei, Kaamber M. Data mining. 高等教育出版社，2001
6 Srivaastava J, et al. Web usage mining: Discovery and application of usage patterns from web data. SIGKDD Explorations, 2000, 1 (2)
7 Fu Y, Sandhu K, Shih M - Y. Clustering of Web users bsed on access patterns[A]. In: Proc. of the 1999 KDD Workshop on Web Mining[C], San Diego, CA. Springer - Verlag, 1999
8 Mobasher B, Cooley R, Srivastave J. Automatic personalization based on Web usage mining [R]: [Technical Report TR99010]. Department of Computer Science, DePaul University, 1999