Web点击流的频繁模式聚类算法

程舒通[1,2]
(1. 浙江大学 计算机学院,浙江 杭州 310027;
2. 杭州科技职业技术学院,浙江 杭州 310022)

摘 要:用户在访问Web站点时会碰到很多问题,主要原因是Web站点对用户需求缺乏适应性。为了提高Web用户的服务质量和用户的满意度,在用户访问网站点击流形成频繁序列模式的基础上,提出基于距离函数的聚类分析以及基于时间相似度函数的二次聚类分析算法。该算法可以求取频繁序列的相关性和反映用户对网页的兴趣的相似度,对下一步改善Web站点的结构及存在形式使站点达到更好的效果起先导作用。

关键词:频繁模式;相似度;聚类;数据挖掘
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2007)09-0018-03

Clustering Algorithm of Web Click Flow Frequency Pattern

CHENG Shu-tong[1,2]
(1. College of Computer Science of Technology of Zhejiang University, Hangzhou 310027, China;
2. Hangzhou Poly Technique College, Hangzhou 310022, China)

Abstract: Difficulties in navigation through the Web are very often encountered by users, the main reason is the lack of adaptation of a Web site to its visitors' needs. For the sake of promoting Web user's service quality and satisfactory, base on the frequency sequence pattern by the Web click flow frequency constitutes and adopt the analysis of the clustering algorithm according to the distance function and the time similarity function. The arithmetic which obtain the relativity of sequence and the similarity of user's interest on a webpage can give us advice how to improve the Web site's structure and ontology.

Key words: frequent pattern; similarity; clustering; data mining

0 引 言

频繁模式的挖掘是关系数据挖掘研究中的一个焦点问题[1],它可以完成相关性[2]、因果性[3]、顺序模式[4]、局部周期[5]等数据挖掘任务,目前已经有大量的研究来解决频繁模式挖掘中的效率问题[6,7]。在用户网站点击流的应用中,频繁模式的数据挖掘成为热点之一。主要是通过识别用户访问Web站点的特征,提高用户对浏览内容的满意度。用户对网站访问形成的点击流就是该用户对网站的访问所形成的序列。

文中研究的目的就是根据用户的点击流的频繁模式聚类来寻找访问序列的相关性以及具有相似的兴趣或者是动机的用户。

由于频繁模式数据挖掘存在着不少的问题,如:可用性上存在不足;频繁模式挖掘结果受阈值影响等。为此,需要解决下列关键问题:

(1)一个度量模式间的相似性;
(2)需要控制簇的质量。

收稿日期:2006-12-13
作者简介:程舒通(1976-),男,浙江杭州人,硕士,讲师,研究方向为人工智能、数据挖掘。

1 相关知识

1.1 序列距离函数

在文献[8]中,作者提出了一个频繁模式间的距离函数,对于给定的两个频繁序列 α 和 β,$T(\alpha)$ 和 $T(\beta)$ 分别为 α 和 β 的支持度列表。则 α 和 β 的距离定义为:

$$D(\alpha,\beta)=1-\frac{|T(\alpha)\cap T(\beta)|}{|T(\alpha)\cup T(\beta)|}$$

它有如下的四个性质:

(1) $D(\alpha,\beta)>0, \forall \alpha\neq\beta$

(2) $D(\alpha,\beta)=0, \forall \alpha=\beta$

(3) $D(\alpha,\beta)=D(\beta,\alpha)$

(4) $D(\alpha,\beta)+D(\beta,\gamma)\geqslant D(\alpha,\gamma), \forall \alpha,\beta,\gamma$

此度量函数反映了支持这两个模式的交易的相似程度,也可以说是这两个模式在同一条交易中出现的概率大小。当计算出来的 $D(\alpha,\beta)$ 的值越小,则说明两个序列 α 和 β 的距离比较小,支持 α 和 β 的交易列表比

较相似,也说明 α 和 β 出现在同一条交易中的概率比较大。但是在现实的数据挖掘过程中,往往只能得到模式和它们的支持度大小信息,如果要得到模式的支持度列表,所消耗的时间和空间都是非常巨大的,特别是对于数据量大和增长式的数据,这几乎是不可能的,所以需要寻找其它的可行的途径。

支持度的大小反应的是一个模式支持的交易的数量,在一定程度上,特别是对于支持度大小相仿的模式,如果支持度的大小相似,则这两个模式出现在同一条交易中的概率相对也是较大的。所以也考虑只用支持度大小来聚类,其距离公式如下:

$$D_1(\alpha,\beta) = \frac{|T(\alpha) - T(\beta)|}{\mathrm{Max}(T(\alpha), T(\beta))}$$

但是对于比较稀疏的数据来说,支持度相近的,但是实际上它们支持度列表不相近的可能性很大,所以定义了下面的距离函数,此距离将两个模式之间相同的项的权重看得比较重:

$$D_2(\alpha,\beta) = \frac{(\alpha-\beta)\cup(\beta-\alpha)}{\alpha+\beta} = \frac{\alpha+\beta-2\alpha\cap\beta}{\alpha+\beta}$$

最后根据支持度大小和支持度列表的相似性,定义了如下的距离函数:

$$D(\alpha,\beta) = \sqrt{(D_1(\alpha,\beta)^2 + D_2(\alpha,\beta)^2)/2} \quad (1)$$

其中:

$$\forall \alpha,\beta, D_1(\alpha,\beta) \in (0,1), D_2(\alpha,\beta) \in (0,1)$$

$$\text{s.t.}\,(D_1(\alpha,\beta)^2 + D_2(\alpha,\beta)^2)/2 \in (0,1)$$

$$\text{s.t.}\, D(\alpha,\beta) \in (0,1)$$

此距离函数有效地将模式本身的相似信息和支持度大小信息结合起来,在不能够得到详细支持度列表的情况下,仍然能够得到比较理想的聚类效果。

1.2 序列的时间相似度

一个用户访问了 n 个页面可以表示成 n - 长度的规则序列,这里用 $\pi = [(\alpha_1,\tau_1^{\alpha})(\alpha_2,\tau_2^{\alpha})\cdots(\alpha_n,\tau_n^{\alpha})]$ 来表示,$\vec{\alpha} = <\alpha_1,\alpha_2,\cdots,\alpha_n>$ 是网站页面访问的序列,$\vec{\tau}^{\alpha} = <\tau_1^{\alpha},\tau_2^{\alpha},\cdots,\tau_n^{\alpha}>$ 是用户浏览相关页面上所花费时间的序列,即 α_j 所对应的时间为 τ_j^{α}。

如果存在两个序列 $\vec{\alpha}$ 和 $\vec{\beta}$,它们在 γ_i 页面中花费的时间是 $\tau_{l(i)}^{\alpha}$ 和 $\tau_{l'(i)}^{\beta}$,所以它们在这个页面上的相似程度可以通过最小 - 最大相似度进行计算:

$$s_i = \frac{\min(\tau_{l(i)}^{\alpha}, \tau_{l'(i)}^{\beta})}{\max(\tau_{l(i)}^{\alpha}, \tau_{l'(i)}^{\beta})} \quad (2)$$

平均相似度为:

$$S' = \frac{1}{L}\sum_{i=1}^{L} s_i = \frac{1}{L}\sum_{i=1}^{L} \frac{\min(\tau_{l(i)}^{\alpha}, \tau_{l'(i)}^{\beta})}{\max(\tau_{l(i)}^{\alpha}, \tau_{l'(i)}^{\beta})}$$

通过时间相似度,可以构造出一个相似度曲线 G_0,在这里 $G_0 = \{\pi_1,\pi_2,\cdots,\pi_P\}$。一些网站用户在浏览网页时表现相似的兴趣可以通过时间相似度曲线体现出来,部分人的浏览网页的过程不和大部分群体是一样的,他们在相似度曲线上表现出来的路径也很少能和其他结点有相关联系。通过这样的曲线,反映出不同的用户对于不同网页的关心程度。

2 聚类算法实现

针对用户访问网站的点击流形成的频繁序列模式,先对序列采用基于距离函数的聚类分析,求取序列的相关性;然后再对已经聚类好的序列分别采用基于时间相似度函数的聚类分析,反映用户对网页的兴趣的相似度。

2.1 根据距离函数聚类

根据公式(1)距离函数,就可以对输入的序列模式进行聚类。目前主要的聚类方法有基于划分的方法(partitioning method)和层次的方法(hierarchical method)。在 Web 分析领域中也使用到图分法[9]的聚类,但算法过于复杂。本算法中需要将距离相近的模式聚类到一起,所以最常用的聚类方法都可以用到这里来。本算法中不过多叙述聚类方法的细节,所以实现了距离函数的 k - 中心方法下的聚类。

算法:将闭合序列模式聚类

输入:

闭合序列模式集 $P = \{\alpha_1,\alpha_2,\cdots,\alpha_m\}$

聚类簇数 K 输出:需要连接的序列位置列表。

输出:k 个闭合序列模式集,$C_1,C_2,\cdots,C_k$

(1) load pattern $\alpha_1,\alpha_2,\cdots,\alpha_m$

(2) initialize $C_1,C_2,\cdots,C_k$

/* 计算两个序列模式之间的距离 */

for each α_i

for each $\alpha_j, j < i$

$\mathrm{DIST}_{ij} = D(\alpha_i,\alpha_j)$

/* 对模式进行聚类 */

(3) repeat

(4) DistributeSamples();/* 指派每个剩余的对象给离它最近的中心点所代表的簇 */

(5) CalcNewClustCenters();/* 重新计算该簇的中心点 */

(6) until 每个簇的中心都不发生变化;

/* 除掉每个聚类中的冗余项 */

for each Cluster $C_i (i = 1,\cdots,k)$

for each Pattern α_n in Cluster C_i

if(IsCovered(α_n) = True

remove α_n from C_i

(7) return $C_i (i = 1,\cdots,K)$

2.2 根据时间相似度聚类

根据公式(2)时间相似度函数,实现了时间相似度函数在 k - 中心方法下的聚类。

算法:将闭合序列模式聚类

输入:

闭合序列模式集 $P = \{\beta_1, \beta_2, \cdots, \beta_m\}, \beta_i = \{\alpha_1, \alpha_2, \cdots, \alpha_k\}$

聚类簇数 N 输出:需要连接的序列位置列表。

输出: n 个闭合序列模式集, $C_1, C_2, \cdots, C_n$

(1) load pattern $\alpha_1, \alpha_2, \cdots, \alpha_n$

(2) initialize $C_1, C_2, \cdots, C_n$

/* 计算两个序列模式之间的时间相似度 */

for each α_i

for each $\alpha_j, j < i$

$ST_{ij} = s(\alpha_i, \alpha_j)$/* 对模式进行聚类 */

(3) repeat

(4) DistributeSamples();/* 指派每个剩余的对象给离它最近的中心点所代表的簇 */

(5) CalcNewClustCenters();/* 重新计算该簇的中心点 */

(6) until 每个簇的中心都不发生变化;

/* 除掉每个聚类中的冗余项 */

for each Cluster $C_i (i = 1, \cdots, n)$

for each Pattern α_k in Cluster C_i

if(IsCovered(α_k) = True

remove α_k from C_i

(7) return $C_i (i = 1, \cdots, K)$

3 结 论

在用户点击流序列的基础上,利用距离函数聚类算法和序列时间相似度函数聚类算法,可以研究序列的相关性和用户对网页内容所表现出来的兴趣的相似性。这对于开发一些目的性强的网站具有较大帮助(如电子商务网站),可以提高 Web 用户的服务质量,使用户享用到满意的个性化服务。下一步工作是在这样算法系统基础上,减少时间复杂度和空间复杂度。

参考文献:

[1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]// In: Proc of ACM SIGMOD International Conference on Management of Data. Washington D.C.: [s.n.], 1993:207 - 216.

[2] Brin S, Motwani R, Silverstein C. Beyond market basket: Generalizing association rules to correlations[C]// In: Proc of 1997 ACM SIGMOD Int'l Conf on Management of Data. Tucson, Arizona, UAS: ACM Press, 1997:265 - 276.

[3] Silverstein C, Brin S, Motwani R, et al. Scalable techniques for mining causal structures[C]// In: Proc of 1998 ACM SIGMOD Int'l Conf. on Management of Data. Seattle, Washington, USA: [s.n.], 1998:343 - 353.

[4] Agrawal R, Srikant R. Mining sequential patterns[C]// In: Proc International Conference on Data Engineering. Taipei, Taiwan: [s.n.], 1995:3 - 14.

[5] Han J, Dong G, Yin Y. Efficient mining of partial periodic patterns in time series database[C]// Int Conf Data Engineering (ICDE 99). Sydney: IEEE Press, 1999:106 - 115.

[6] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation[C]// In: Proc of 2000 ACM SIGMOD international conference on management of data. Dallas, Texas, United States: [s.n.], 2000:1 - 12.

[7] Ayres J, Flannick J, Gehrke J, et al. Sequential Pattern mining using a bitmap representation[C]// Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. Edmonton, Alberta, Canada: [s.n.], 2002:429 - 435.

[8] Xin D, Han J, Yan X, et al. Mining compressed frequent - pattern sets[C]// Proceedings of the 31st international conference on very large databases. Trondheim, Norway: [s.n.], 2005:709 - 720.

[9] Strehl A, Ghosh J, Mooney R. Impact of similarity measures on Web - page clustering[C]// In Proc. 7th natl Conf on Artificlal Intellgigence: Workshop of Artificial Intelligence for Web Search(AAAI 2000). Austin, TX, USA: [s.n.], 2000: 58 - 64.

(上接第 17 页)

[4] Murino V. Structured neural networks for pattern recognition [J]. IEEE Transactions on System, Man, and Cybernetics part B: Cybernetics, 1998, 28(4):53 - 56.

[5] Han Hong, Yang Jing - yu. Combination of neural network classifiers[J]. Journal of Computer Research & Development, 2000, 37(12):1488 - 1492.

[6] Flandrin P, Rilling G, Gonealves P. Empirical Mode Decomposition as a Filter Bank[J]. IEEE Signal Processing Letters, 2004, 11(2):112 - 114.

[7] Slavik P, Govindaraju V. Equivalence of Different Methods of Slant and Skew Corrections in World Recognition Applications [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2001, 23(3):323 - 326.

[8] Pandya A S, Macy R B. Pattern Recognition with Neural Networks in C++[M]. [s.l.]: IEEE Press, 1999:156 - 172.