编者按: 数据挖掘就是从大量的数据中发现隐含的规律性的内容, 解决数据的应用质量问题。充分利用有用的数据, 废弃无用的数据, 是数据挖掘技术的最重要的应用。该文对我国农业信息收集有一定帮助。

Web Usage Mining 在电子商务中的应用

宁小红 (浙江师范大学杭州幼儿师范学院, 浙江杭州 310012)

摘要　提出结合站点的拓扑结构和 Web 页面内容的改进算法。改进算法根据 Web 页面的内容链接过滤非内容页, 利用页组的组内链接度提高挖掘结果中频繁访问页组的机率, 以提高客户访问率, 进而能提高电子商务的效益。

关键词　网页; Web 使用挖掘; 改进算法; 电子商务; 应用
中图分类号　TP31　　文献标识码　A　　文章编号　0517 - 6611(2007) 13 - 04071 - 03

随着我国网络技术的发展和Internet 的迅速普及, 电子商务也得到了蓬勃发展。许多电子商务网站存在的共同问题是缺少个性化服务。客户在浏览基于 Web 的网络商务信息的过程中, 总会面临大量与己无关的信息。这正如人们所说的"99 % 的信息对99 % 的用户是无用的"。对于网站来说, 每一个来访的用户都是不同的个体, 每个个体都有不同的访问习惯和兴趣, 但同时人们的行为远比想象的容易并准确地被预测。因而, 对于电子商务网站, 必须改变过去对所有用户提供统一界面、同样内容的方式, 网站需要拥有好的自动辅助设计工具、针对不同用户的访问习惯和兴趣, 网站应提供不同的服务。这类网站比起其他同类网站更有可能吸引更多的用户。

1　WUM 技术

近年来, Web 挖掘技术在 Web 个性化方面得到了越来越广泛地应用。Web Mining 是对 Web 文档的内容、Web 上可利用资源的使用情况以及资源之间的关系进行分析, 从中发现有效的、新颖的、潜在有用的、并且最终可理解的模式。简单地讲, Web 挖掘指从 Web 服务器上的数据文件中提取人们关心的知识。Web 挖掘的一种比较流行的分类方法见图1。根

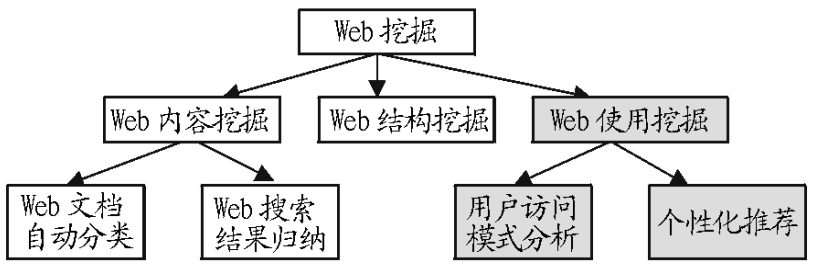

图1　Web 挖掘分类

据 Web 挖掘的数据对象, 将 Web 挖掘分为3 类: 内容挖掘(Content Mining) 、结构挖掘(Construct Mining) 、使用挖掘(Usage Mining) 。其中, Wed Usage Pattern Mining(WUM) 是对用户访问 Web 时在 Web 服务器留下的访问记录进行挖掘, 即对用户访问 Web 站点的存取方式进行挖掘, 从这类记录文件 Web Log 中抽取感兴趣的模式的过程。WUM 被认为是 Web 挖掘技术中最有前途的研究领域。目前, WUM 已成功地应用于 Web 个性化(Web Personalization) 服务、系统改善(System Improvement) 、推荐系统(Recommender) 、商业智能(Business Intelligence) 等领域。

基于 WUM 的个性化服务的基本思路是分析 Web 日志数据, 利用 Web 挖掘方法发现用户的使用模式, 从而向用户提供个性化服务。对于一个成熟的电子商务网站, 有大量的 Web 访问信息可以利用(如用户的访问日志、注册信息、成交意向、购买结果等) 。这些信息如不加以利用, 则会造成资源的浪费。利用 WUM 技术充分挖掘这些信息资源, 了解和掌握客户的情况、需求、能力、进度、兴趣等, 及时调整商务计划, 呈现符合客户需要的个性化信息资源。

WUM 基本过程(图2) 可分为2 个部分: 离线部分和在线部分。其中, 离线部分分为数据收集、数据预处理、模式发现; 在线部分主要包括模式分析与在线推荐。采用 WUM 技术对基于 Web 的网络商务系统实行数据挖掘可分为4 个阶段: 数据收集、数据预处理、模式发现、模式分析。

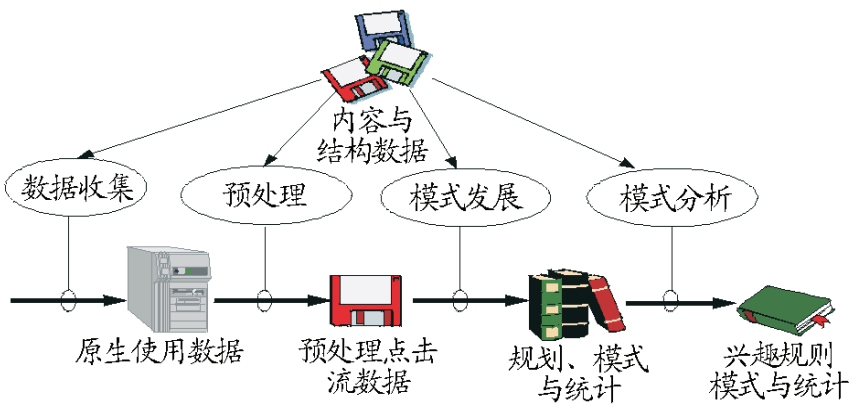

图2　WUM 的处理过程

通常利用兴趣性评价 Web 使用挖掘的结果, 被客户频繁访问的页面应该满足以下3 点: ①经常被大量的用户在其一次浏览过程中相继访问; ②页面内容丰富; ③网站内页面相互之间有尽可能少的超文档链接。

2　在电子商务中的应用

2.1　指导在线推荐模式　模式分析与在线推荐是 WUM 中最后一个重要步骤, 主要是通过选择和观察, 把发现的规则、模式和统计值转换为知识, 再经过模式分析得到有价值的模式, 即人们所需要的规则、模式, 再采用可视化技术, 以图形界面的方式表现出来, 进而指导实时 Web 个性化推荐服务, 实现 Web 个性化服务的目的。在线推荐时根据浏览页 p 的推荐系数产生推荐集, 引用浏览页 p 的推荐系数 Rec(S, p) 公式:

$$Rec(S,p) = weight(p,C) \times match(S,C) \quad (1)$$

其中, 用户当前会话 $S = \{s_1, s_2, \cdots, s_n\}$

基金项目　2006 年度浙江省教育厅科研项目(20060441) 。
作者简介　宁小红(1966 -) , 女, 浙江杭州人, 高级讲师, 从事 Web 挖掘方面的研究。
收稿日期　2007-01-22

根据当前的用户会话产生的实时推荐集:

$$W_k^c = \begin{cases} weight(p_i, C), if & p_i \in C \\ 0, otherwise \end{cases} \quad (2)$$

总体使用特征 $C = \{w_1^c, w_2^c, \cdots, w_n^c\}$ (3)

使用余弦相似性函数计算 C 和 S 之间的匹配系数:

$$match(S, C) = \frac{\sum_k (w_k^c \times s_k)}{\sum_k (s_k)^2 \times \sum_k (w_k^c)^2} \quad (4)$$

2.2 关联规则算法的改进 假设 FG_k 是包含 k 个页面的频繁访问页组的集合,其中每个页组的支持度都大于预先设定的阈值 T。在传统的页面聚类算法中,支持度是指包含页组中所有页面的用户会话的个数。在改进算法中,模仿和引用了最新的计算方案,将支持度的计算进行了扩展,一个页组 G 的支持度为:

$$Support(G) = O(G) \times H_{NCLR}(G) \times [1 - GILD(G)] \quad (5)$$

式中,O(G) 是包含 G 中所有页面的用户会话的数目(也就是传统算法中的 Support 的定义);$H_{NCLR}(G)$ 是 G 中所有页面 NCLR 的调和平均值;GILD(G) 是 G 的页组链接度。

研究发现,频繁访问页组是一个递归的过程。首先,将 FG1 初始化为支持度大于 T 的页面,FG2 是在 FG1 的基础上产生,FG3 又是在 FG2 的基础上产生,依此类推。

2.3 聚类算法的改进 为了提高算法的效率,Mannila 等引入了修剪技术来减小候选集 C_k 的大小,从而明显改进了生成所有频集算法的性能。算法中引入的修剪策略基于以下性质:一个项集是频集当且仅当它的所有子集都是频集。如果 C_k 中某个候选项集有一个(k - 1) 子集不属于 L_{k-1},则这个项集就可以被修剪掉不再被考虑(图3)。这个修剪过程可以降低计算所有的候选集的支持度的代价。

```
1: count the NCLR of all distinct pages appeared;
2: initialize FG1 as the top requested single
       page groups with Support >= T;
3: for (i=2; i<=k; i++) {
4:   Sort the pages of groups in FGi-1 in lexicographical order;
5:   for each group {x1, ..., xi-1} in FGi-1 {
6:     for each group {y1, ..., yi-1} in FGi-1 {
7:       if (x2=y1 and ...and xi-1=yi-2) {
8:         construct a new group G={x1, ..., xi-1, yi-1};
9:         if (G not already in CGi) {
10             test all other combinations of subgroups of G with size (i-1);
11:          if (all such subgroups are in FGi-1)
12:            if (Support(G) >= T) add G into FGi;
13:        }
14:      }
15:    }
16:  }
17:}
```

图3 挖掘频繁访问页组的改进算法

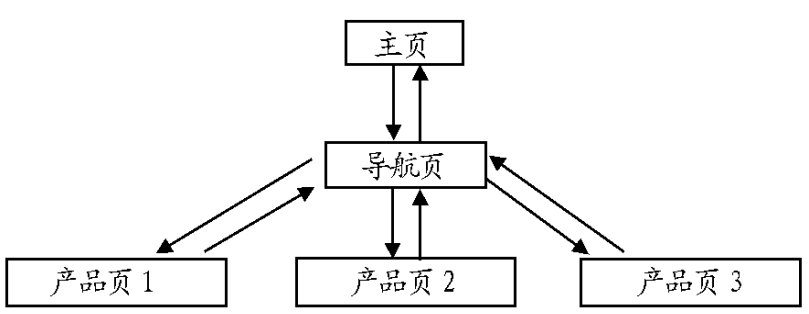

图4 站点的拓扑结构

在电子商务网站的基本网页中,一般的设计至少有主页、导航页、产品页1、产品页2、产品页3 等基本页面(图4)。

由于主页与导航页有很高的访问量,如果将 SpeedTracer 中的页面聚类算法应用到该日志文件上,得到的结果可能包括2 个页组,即主页、导航页、产品1 和导航页、产品页1、产品页2 等。但是,第1 个页组站点的拓扑结构显示这3 个页面已经是相互链接的,而且挖掘算法不必报告主页和导航页有大量用户访问这一事实。相反,第2 个页组(产品页1,产品页2 和产品页3) 也许会由于选定的阈值较高而未被发现。

一个好的聚类算法应当发现用户真正感兴趣的页组,即挖掘出的频繁访问页组中,页面间的相互链接程度尽可能地低。为此,引入页组的组内链接度的概念。一个页组 G 中页面的超链接关系是一个有向图 Graph(G),页组内的页面是有向图的节点,页面之间的链接对应有向图的边。如果有向图的边集合为空,那么这些页面之间相互不能直接到达,只能依赖于站点的其他页面间接到达;反之,如果这些页面之间为全互连,那么从任何一个页面都能够到达另外任何一个页面。所以,引入组内链接度来刻画组内页面间的链接紧密程度。

2.3.1 组内链接度(GILD, Group Inter-Link Degree)。引入定义:

$$GILD(G) = |Graph(G)| / (|G| \times (|G| - 1)) \quad (6)$$

式中,$|G| > 0$,$|G| \neq 1$。

|Graph(G)| 是有向图 Graph(G) 中的边数,|G| 是页组 G 中的页面数。

当页组内的任意2 个页面之间都没有链接,则其 GILD 为0;反之,页组内的任意2 个页面都是相互链接的,则其 GILD 为1。这样的页组就没有必要出现在挖掘算法的结果中。定义中没有考虑边的位置,可以利用子图的个数来定义页组链接度,但是这种方法消耗较大的CPU 时间和内存。

2.3.2 聚类算法改进后的效果。为了对该文提出的聚类算法进行评价,将它和搜索引擎 Yahoo、K- 均值聚类算法进行了比较。试验中采用的 Web 信息集是用搜索引擎 Yahoo 从 Internet 上搜索得到的。整个试验在 PⅣ450 计算机的 Windows XP 平台上进行。

2.3.2.1 算法的准确性。用 Yahoo 根据不同的主题进行了50 次搜索,下载每次搜索到的前20 个信息构成由 Yahoo 产生的50 个信息类,每个类中包含有1 000 个信息;然后,用人工方法剔除无关信息,同样将它们分成50 个类,并依此作为分类准确性的基准;最后,分别采用该文算法和 K- 均值算法对这个信息集进行聚类。

由于不同的聚类算法产生类的数目很可能不同,为了使比较更趋公平,选用各自质量最好的40 个类进行比较。图5 即为由不同算法产生的40 个类的平均精度的对比情况。由

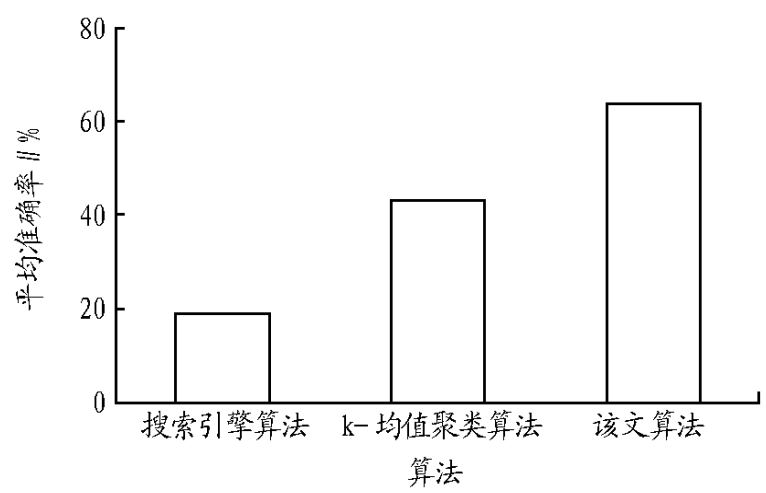

图5 不同聚类算法的精度对比

于该文算法允许类间重叠,并采用了类确认技术,因此平均聚类精度最高。

2.3.2.2　算法的扩展性。同样用 Yahoo 在 Internet 上搜索前面50 个主题的相关信息,但是下载的是每个主题的前10 个信息,并以5 的增幅逐步递增到前45 个信息,形成信息数量依次为500、750、1 000、1 250、1 500、1 750、2 000 和2 250 的8 个信息集;然后,分别采用该文算法和K- 均值算法对这8 个信息集进行聚类,并计算它们的平均聚类时间。图6 表明,随着信息数的增加,2 种算法的平均执行时间都在增加,但是该文算法平均执行时间的增幅较小,增长趋势较为缓慢,说明该文算法的扩展性较好。其原因在于,该文算法用主题表示信息,降低了信息特征向量的维数;同时,又以主题为事务项,减少了数据处理的工作量。

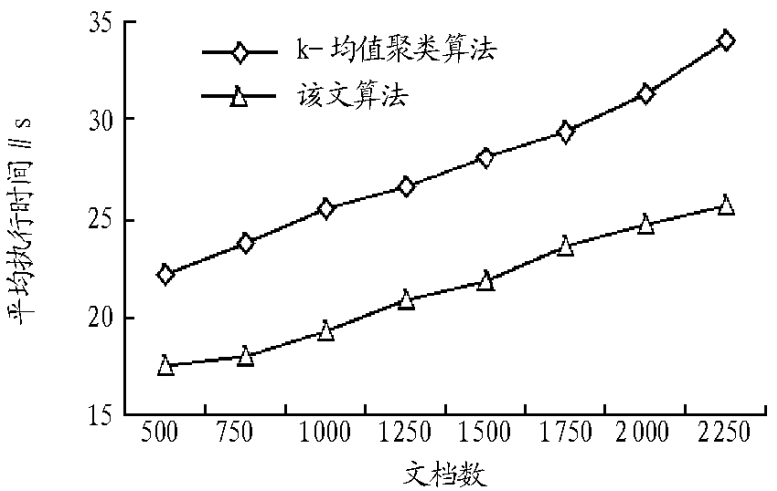

图6　不同聚类算法的扩展性对比

试验表明,改进的页面聚类算法的性能高于一般算法,挖掘结果有明显地改善。

3　结语

电子商务在企业和商贸领域占据着越来越多的市场份额。个性化推荐服务是电子商务领域中非常重要的新技术。它在帮助用户快速定位感兴趣商品的同时,也为企业实现了增值,所以将会成为未来电子商务网站的关键模块。现阶段个性化推荐服务面临如何发现客户行为的个性化特征以及 Web 重要页面的组织等问题。近年来兴起的 Web 挖掘技术主要用于商品的市场定位和消费分析,以辅助制定市场策略,还可以用来分析购物模式,预测销售行情。将 Web 挖掘技术应用于实现个性化推荐服务势必会推动电子商务的发展。

参考文献

[1] 韩家伟,孟小峰,王静,等. Web 挖掘研究[J]. 计算机研究与发展,2001,38(4):405-413.

[2] 刘培刚. Web 挖掘技术在电子商务中的应用研究[J]. 情报学报,2002(6):40-45.

[3] 周惠宏,柳益君,张尉青,等. 推荐技术在电子商务中的运用综述[J]. 计算机应用研究,2004,21(1):8-11.

[4] 严华云. Web 挖掘在网络教育中的应用研究[J]. 湖州师范学院学报,2003(6):72-75.

[5] EIRINAKI M, ANAND S, BUCHNER A. Web mining for web personalization [J]. ACM TOIT, 2003, 3(1):2-27.

[6] MULVENNA M, ANAND S, BUCHNER A. Personalization on the net using web mining[J]. CACM, 2000, 3(8):123-125.