维普资讯 http://www.cqvip.com

基于概念格和关联规则Web个人化系统

张 涛,周爱武,谢荣传
(安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039;
安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:最近的一些研究提出将Web使用日志的挖掘技术应用于Web个人化系统中,用于克服传统个人化技术(如CF技术、基于内容的过滤技术)中存在的问题,如处理大数据量的能力较差,依赖于用户主观的登记信息,产生的用户描述是静态的,不能获取对象之间丰富的语义联系等。但是基于Web使用日志挖掘的个人化技术不能适用于用户的使用信息获取困难或者站点内容经常变化的情况。更有效的办法是将站点的内容特征和使用特征结合到一个Web挖掘结构中去,以备推荐引擎统一使用。提出了一个基于关联规则挖掘的个人化系统,它使用概念格作为存储频繁页面集的数据结构,并介绍了如何利用概念格实时地为当前活动用户产生推荐集。

关键词:Web使用日志挖掘;推荐集;概念格
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2008)02-0139-04

A Web Personalization System Based on Concept Lattice and Association Rule

ZHANG Tao, ZHOU Ai-wu, XIE Rong-chuan
(Educational Ministry Key Lab. of Intelligent Computing & Signal Processing,
Anhui University, Hefei 230039, China;
College of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: Recent proposals have suggested Web usage mining as an enabling mechanism to overcome the problem associated with more traditional Web personalization techniques such as collaborative or content-based filtering. These problems include lack of scalability, reliance on subjective user ratings or static profiles, and the inability to capture a richer set of semantic relationships among objects. Yet, usage-based personalization can be problematic when little usage data is available pertaining to some objects or when the site content changes regularly. For more effective personalization, both usage and content attributes of a site must be integrated into a Web mining framework and used by the recommendation engine in a uniform manner. In this paper present an association-rule-based recommendation system, which extract usage patterns from Web log file and use the concept lattice as its data structure storing frequent itemsets. And show how to use concept lattice generated to compute a recommendation set for the active user on real time.

Key words: Web usage mining; recommendation set; concept lattice

0 引 言

新信息、新产品、新服务每天都在不断被推上Web,同时,用户的种类、数量和关注点也在增加。一方面,用户疲于在信息海洋里搜寻信息,另一方面Web网上的服务商也在不断设法获取用户兴趣爱好,以填补用户和网站之间的信息鸿沟。但是由于Web是无结构的、动态的,并且Web页面的复杂程度远远大于文本文档,人们要想找到自己想要的数据有如大海捞针一般。解决这些问题的一个途径就是将传统的数据挖掘技术和Web技术结合起来,进行Web挖掘。

针对用户特性并向用户提供个性化服务已经成为Web技术的一个研究热点,个人化技术就是基于这种需要产生的。把Web站点的内容、语义信息与Web使用日志的挖掘结合起来是目前基于Web使用日志挖掘个人化技术研究的新趋势。

笔者提出了一个基于关联规则挖掘的个人化系统,它使用概念格(concept lattice)作为存储频繁页面集的数据结构。

收稿日期:2007-05-15
作者简介:张 涛(1982-),男,安徽临泉人,硕士研究生,研究方向为Web与数据库技术;周爱武,副教授,硕士生导师,研究方向为数据库技术;谢荣传,副教授,硕士生导师,研究方向为数据库、Intrenet应用、多媒体技术。

维普资讯 http://www.cqvip.com

1 基本思想

个人化系统的有效性可以从生成推荐集的覆盖度和精确性两个方面来评价:精确度度量表示系统生成正确推荐集的程度;覆盖度评价系统预测所有可能被用户访问的页面的能力。最近的一些研究已经将关联规则挖掘应用到个人化系统中。但是在极大程度上,这些研究依赖于生成推荐集之前发现所有的关联规则(因而要求在推荐阶段扫描所有的规则)或者在当前用户的邻居中在线生成关联规则,而不关心支持度或者用户历史记录的大小对推荐集的影响。这里为个人化系统提出了一个使用关联规则挖掘的灵活结构。特别的,提出了一种适用于个人化系统的存放频繁项目集的数据结构;推荐算法使用这个数据结构实时生成推荐集,而不需要从频繁项目集合中生成关联规则;期望这个结构可以克服基于关联规则的个人化系统的一些弱点,例如由于高的支持度和大的数据历史记录形成的低覆盖度、由于数据特征稀疏而造成的低精确性等。

2 数据预处理

Web 使用日志的挖掘并非仅是将新数据用于传统的挖掘算法,由于存在一些原因,原始的日志文件不能提供可靠的输入数据,因而首先需要对日志文件进行预处理,其流程如图 1 所示。

图中的原始使用记录数据包括:一般的日志文件,包含如下信息 IP 地址、请求时间、方法(如 GET)、被请求文件的 URL、HTTP 版本号、返回码、传输字节数;引用日志(referer log)文件:包含应用页面的 URL;代理日志,含有客户端 Web 浏览器及操作系统。

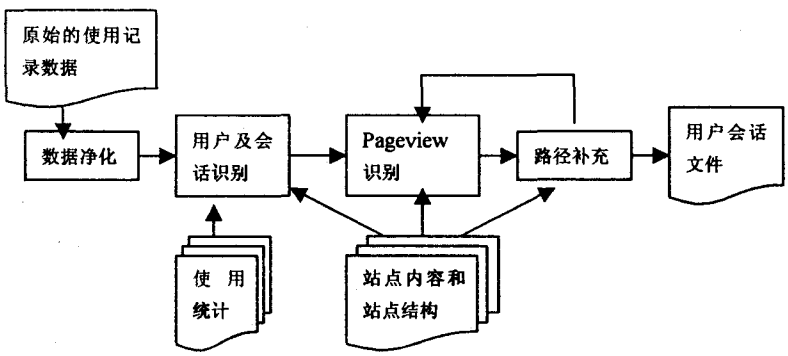

图 1 数据预处理的流程

对基于 Web 日志的挖掘而言,数据净化——从日志中删除挖掘过程中不需要的域很重要,因为挖掘结果的质量依赖于它所使用的挖掘数据的质量。HTTP 协议要求每一个从服务器申请的文件都对应一条独立的连接,因此,当一个用户请求浏览某个页时可能会生成几个日志条目,因为图形和脚本程序等也会随着 HTML 文件的下载而下载;用户模式的分析并不关心用户没有显式请求的文件,在许多时候,用户并不明确请求页面上的图形文件或其他附加文件,因此只有 HTML 文件的请求条目与用户会话相关。通过检查 URL 后缀删除那些被认为不相关的数据:辅助文件 gif,jpeg,wav,au;脚本文件 cgi 等。此外还要删除含有特殊 HTTP 返回码的条目,如返回码为 404。其次是删除来源于 spider,crawlers, robots 等代理程序的导航信息;最后,还要补充由于本地或代理服务器缓存而遗漏的页面,这将在会话识别后完成。将 URL 的表示规范化也是这个阶段的任务之一[1]。

用户识别是将用户及其请求的页面相关联的过程,最终将其表示为带权的 pageview 的集合或者序列,这个过程由于本地缓存、代理服务器和防火墙的存在而变得复杂。但是可以使用一些启发规则,例如 IP 地址相同,但是代理日志中的浏览器和操作系统变了,则认为每个不同的代理表示不同的用户。

在一个日志文件中,同一个用户可能已经多次访问了站点,会话识别的目的是将用户的访问记录分为单个的会话——从用户进入站点到他离开所实行的动作。最简单的是利用超时,即用户的某两次请求之间的时间差已经大于一个 time window,则认为是新的会话,一般取这个 time window 为 30 分钟。会话识别定义为:把活动正确的映射到不同个体 + 正确划分个体的行为。这里引入"真实会话"的概念。首先,认为 Web 日志 L 是一系列请求页面的列表,这是一个基于访问时间的序列,L 中的第 i 个实体表示为 $L[i]$。用户在一次访问中所请求的页面序列为一个真实会话 r,r 中的元素对应于 L 中的实体。R 是一组真实会话的集合,它有如下特点:(1)R 中每个会话的元素都按其时间标签排序,即保留了 Web 日志的时序;(2)R 中的会话形成了对 L 的划分。

pageview 的识别是用于确定在一次浏览中,访问了哪些页面文件及相关的对象。这个任务依赖于站点的结构信息。pageview 识别的另一个任务就是根据站点的领域知识及挖掘任务的目标,为识别到的每一个 pageview 赋以权重。例如,在一个电子商务的站点中,面向产品的事件(如,购物车的变化或者商品信息的浏览等)所对应的 pageview 被赋予较高的权值,因为它具有比其他类型页面更大的意义;一个提供目录(content)信息的网站,目录页面将比导航页面具有更大的权值。

路径补充是将由本地或代理服务器缓存而遗漏的

维普资讯 http://www.cqvip.com

页面补充完整,详见文献[2]。

3 模式发现

Web日志挖掘的个人化系统的第二个阶段就是从用户会话的集合中获取用户的使用模式,输入是第一阶段预处理结果产生的用户会话的集合,输出是一个适用于推荐阶段的含有频繁项目集的数据结构。

概念格最初是作为信息获取和知识发现被提出的,其关于两个集合之间关系的描述是一些概念聚类方法的基础。在对于基于Web使用日志挖掘个人化算法的研究中发现,概念格对于分类属性间具有一定偏序关系的对象非常有效,这个特点对于分析一个主题网站的使用情况很有效,因为网站的页面在语义上可能存在一种偏序关系,例如,产品列表页相对于每个产品的页面,就存在一种"上级"关系;concept lattice可以把这种偏序关系反映到它的网格结构中。此外,概念格的结构比较有利于信息的可视化,例如可以按信息的较高分类将信息分为若干块,由用户选择需要详细了解的信息块,起到较好的导航作用;概念格的结构与当前流行的语义Web的结构具有同构性,将它应用到Web使用日志的挖掘中可以更容易地与Web的语义信息结合起来,不仅得到用户的使用模式,还可以解释为什么得到这样的模式。文中概念格的结构作为一种存储频繁项目集的数据结构引入个人化的应用中,思想类似于文献[3]中提出的关联规则挖掘的思想。

3.1 概念格

假设E和E'是两个有限集,R是E和E'之间的二元关系,即$R \subseteq E * E'$,符号xRx'表示元素x与元素x'之间存在二元关系,其中$x \in E, x' \in E'$;

由文献[4]中定义,3元组(E, E', R)是一个正规表达式,它对应了唯一的一个概念格。

定义1:概念格中的每个元素C是一个偶对(X, X'),它满足如下的特征:

$X \subseteq E, X' \subseteq E'$,

$X' = f(X) = \{x' \in E' \mid (\forall x \in X), xRx'\}$且

$X = f(X') = \{x \in E \mid (\forall x' \in X'), xRx'\}$

C被称作一个概念(concept),其中,X称作C的外延,X'称作C的内延。

概念之间的偏序关系"<"如下定义:

定义2:$C_1 = (X_1, X_1'), C_2 = (X_2, X_2'), C_1 < C_2 \Leftrightarrow X_1' \subset X_2' \Leftrightarrow X_2 \subset X_1$。

定义3:一组拥有这种偏序关系"<"的概念集合是一个完备网格,称为概念格,标记为$CL(E, E', R)$。

概念之间的偏序关系用于构建概念格的*Hasse*图,图中每个结点C代表一个概念,边(C_1, C_2)表示$C_1 < C_2$,方向是自上而下的,且概念格中不存在其他结点C_3使得$C_1 < C_3 < C_2$,其中,C_1称作C_2的双亲,C_2是C_1的一个孩子。通常,E是一组对象的集合,E'是对象属性的集合,在应用中,E是用户会话的集合T,而E'是站点中页面的集合P。在这里,每个用户会话都对应一个唯一的session ID,每个页面也都被赋予一个唯一的page ID。

概念格中,每个概念代表了一组对象及这组对象共同特征,图的层次揭示了对象分类之间一般/特殊的层次关系。两个概念在层次上相距越近,就表示这两个概念具有越强的语义相关性。

3.2 使用属性之间偏序关系

假设E'中的元素之间存在偏序关系"<",将这种偏序关系加入生成的概念格中,只需要将E和E'之间的二元关系R重新定义即可。

定义4:$xR^+ x'$:如果存在y'使得$y' \leqslant x'$且xRy。

概念在关系R^+下重新定义为:

定义5:$C = (X, X')$

$X \subseteq E, X' \subseteq E'$,

$X' = f(X) = \{x' \in E' \mid (\forall x \in X), xR^+ x'$且$\neg \exists y'(y' < x'$且$(\forall x \in X, xR^+ y'))\}$

$X = f'(X') = \{x \in E \mid (\forall x' \in X'), xR^+ x'\}$

新定义函数f, f'从概念中删除了对于这个概念不具有特殊性的元素,这种删除是由E'中元素之间存在的偏序关系"<"决定的。使用新定义的f和f'生成的结果概念格标识为$G_<$。

3.3 频繁项目

预处理结果确定出页面集合$P = \{p_1, p_2, \cdots, p_n\}$和用户会话的集合$T = \{t_1, t_2, \cdots, t_m\}$。$T$中的每个用户会话$t_j$是$P$的子集,表示在这个会话中用户访问过的页面。

给定一个用户会话的集合$T = \{t_1, t_2, \cdots, t_m\}$,$P$中的每个项目$p_i$的支持度为:

$$\sigma(p_i) = \frac{|\{t \in T : p_i \in t\}|}{|T|}$$

通过设定最小支持度,可以删除那些支持度过低的页面,得到一组频繁项目。

4 模式应用

推荐引擎将上一阶段产生的频繁项目集合作为输入,并将当前用户会话与已经发现的模式相匹配,为之提供推荐集(这个集合可以是链接、广告、产品等),被推荐的对象将添加到当前用户会话的最后一页发送给

维普资讯 http://www.cqvip.com

用户。

在一个会话中,用户可能访问几条路径,因而一个用户会话中可能包含几个独立的信息片,称这些独立的信息片为子会话,这种情况下,使用前一个子会话计算当前子会话的推荐集显然是不合适的。在当前用户会话的上定义一个固定大小的 sliding window 以获取当前用户的历史深度;尺度为 n 的 sliding window 意为仅允许最后 n 个被访问的页影响推荐集[5]。

如前一节所述,概念格中每个概念是一个偶对 $C=(f,\text{pageSet})$,pageSet 是频繁页面集,f 是这个页面集出现的频率。这里使用关联规则的可信度作为推荐项目的推荐分值。

关联规则 $r:X=>Y$ 的可信度 confidence($X=>Y$) $=\frac{\sigma(X\cup Y)}{\sigma(X)}$,假设当前活动会话的窗口大小为 k,下面讨论在概念格中如何搜索这个活动会话的推荐集。

(1) 当 $k=1$ 时:

a. 从 index$P$ 中找到这个页面所对应的结点 H;

b. 设置一个页面集合 is 存放得到的推荐集,is 初始化为空集;

从这个结点 H 出发,宽度搜索 Hass 图,且对搜索加入这样的限制:设置可信度阈值为 threshold,考察搜索中遇到的结点 Hc,如果 Hc.f/H.f >= threshold,将 is 替换为 is ∪ Hc.pageSet,则继续从这个结点出发宽度优先遍历,否则停止考察这个结点。

(2) 当 $k>1$,需要考虑到 k 个页面"同时"出现的情况:

由概念格的特性知,如果这个 item 的集合曾经在构建概念格的数据中出现过的话,那么这个集合在 Hass 图中一定对应唯一一个频率最高的结点。从这个结点出发,搜索与其语义相关的概念结点,构建推荐集,后一过程类似于 $k=1$ 的情况。

① 找到这个结点。

Search(ItemSet items) 模块用于处理当当前活动会话窗口的大小 $k>1$ 的情况,它的输入 items 是当前用户会话的活动窗口,是一个页面的集合;输出是这个页面集合在 Hass 图中出现频率最高的结点。这个方法是在建立的概念格中保留一个数组 C,$C[i]$ 是所有 pageSet 域长度为 i 的结点集合。Search 模块的输入是当前用户在一个活动窗口内的访问序列 list,输出是这个集合出现最高频率的结点。

```
Node Search(ItemSet items)
{
Int n=输入集合 s 的长度;
Exist=false;
Int k=n;
While (! exist){ //在已经建立的 Hass 图中寻找这个集合
For(c[k]中的每个结点 H) If(! exist &(sH.pageSet))exist=
ture;
K=k+1;
}
return H;
}
```

②生成推建集。

search 返回的结点标明了这个集合出现频率最高的结点位置,从这个结点出发,象处理 $k=1$ 的情况一样处理计算它的推荐集合。算法如下:

```
ItemSet getRecommend(ItemSet items,int threshold)
/* items 是输入的预测 item 集合;
threshold 是定义的可信度阈值
*/
{
ItemSet is=new ItemSet(); //用于放置已经预测到的 items
Node H=search(items);
int f=H.f; //把 f 赋值以 p 的频率宽度访问 Hass 图
建立结点队列 v;
H 入队;
while (队不空) {
取队头 H;
Is.<-H.pageSet-items;//把 p 结点中除 items 以外的 item 加
入
for(H 的每个孩子 Hc){
if(Hc.f/f>=threthold)Hc 入队;
}//for
}//while
}//getRecommend;
```

这种方法最明显的缺点就是,作为挖掘结果的网络结构与参与挖掘的对象数目有显然相关的联系;第二,在线搜索相关属性,且搜索的时间并不固定。

优点在于,根据用户的访问页面集合,对用户做了一个层次性的划分;这种结构可以较轻松地与当前的语义 Web 页结合使用,附以语义的解释,可以对这个划分有一个语义上的说明;而且概念格的结构也可以较方便的层次可视化用户的访问。

5 结 论

介绍了一个基于关联规则挖掘的个人化系统,并论述了它的流程:数据预处理、模式发现和模式挖掘。使用概念格作为关联规则的存储结构,给出了概念格的基本概念的介绍及它的相关变形结构,以及如何利用概念格从用户会话中获取用户的访问模式并把这个模式用于生成推荐集的算法。

(下转第 158 页)

维普资讯 http://www.cqvip.com

控制结点扩展的策略,这种策略优先扩展深度小的结点,把问题的状态向横向发展。

2.2 破解算法

首先,在获取到公钥后,根据先根序遍历序列和后根序遍历序列构造出所有可能应用的加密树(假设有 N 种),形成一个可能加密森林。建立一个节点 root,建立一个基于这个可能加密森林的 N 叉可能加密树。

然后根据密文,在应用广度优先搜索算法的基础上,对上述 N 叉可能加密树进行广度优先算法查找出所有的对应的密文与二进制明文之间的对应关系,同时把所有论域中所有的加密符号对应的二进制明文全部按加密树森林的对应顺序存储起来,形成一个密文与明文二进制相对应的破解库。

最后,把二进制明文转换成自然语言的明文,存储到相应的密文与明文的破解库中,替代原来的二进制明文,按照密文的先后顺序,找到查找密文与明文破解库中的对应节点,形成一个按照密文顺序排列的破解链表,再按照自然语言的语义,就能够得到对应密文的明文了。

2.3 举 例

由第一节给出的加密例子给出对应的破解过程。

根据先根序遍历序列和后根序遍历序列,也就是公钥:a1a2a4a2a5 和 a4a2a3a5a1,得出有可能的加密树如图 2 所示。

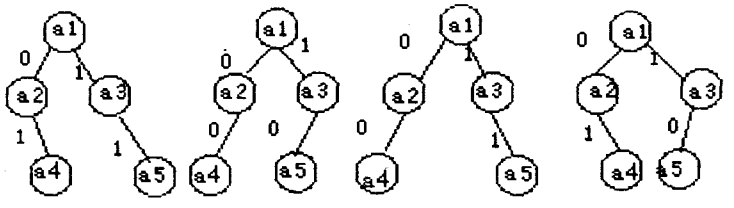

图 2 可能的加密森林

根据以上的可能的加密树,由广度优先的破解方法,要生成可能加密森林的一个 N 叉树,因此,根据图 2 可能的加密树得出加密森林,如图 3 所示。

在可能的加密 N 叉树上应用广度优先搜索算法,

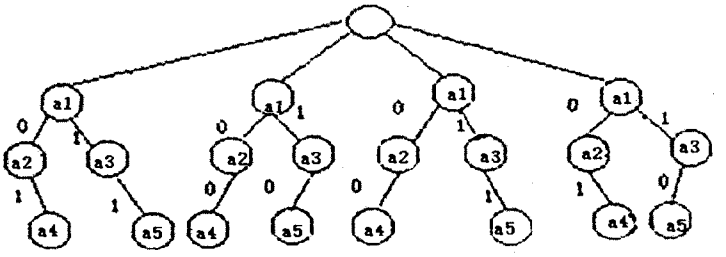

图 3 可能的加密 N 叉树

把遍历到的节点按对应顺序写到解密链表中,如图 4 所示。

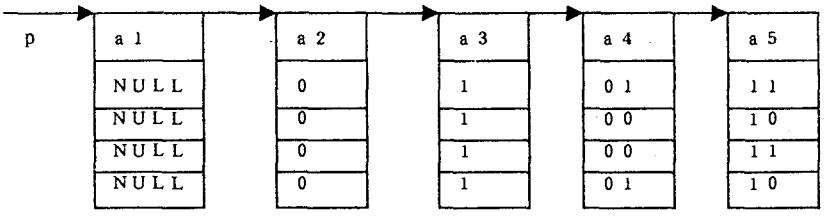

图 4 解密链表

然后根据广度优先搜索算法,将这个链表中的二进制转换成自认语言,最后按照自然语义得到相应的明文。这样就达到了破解的目的。

3 总 结

通过分析了二叉树加密算法的加密与解密及其传输过程,总结出了基于图的广度优先搜索算法的破解算法,主要针对在基于二叉树的先根序搜索序列和后根序搜索序列能够得到一个加密树森林,通过加密森林得到一个 N 叉加密树,再对 N 叉树进行广度优先搜索,逐步建立解密链表,达到破解的目的。

参考文献:

[1] 严蔚敏,吴伟明.数据结构[M].北京:清华大学出版社,2003.

[2] 臧武军.数据加密技术[J].教育信息化,2005,12:28-30.

[3] 陈 伟,付宇洁,秦 科.基于二叉树的加密算法[J].实验科学与技术,2006(2):81-85.

[4] 张乐星.一种新的网络数据加密方案[J].计算机时代,2004(3):10-11.

[5] 唐明华.用改进的广度优先搜索算法计算点的出行范围[J].茂名学院学报,2006,16(3):35-38.

(上接第 142 页)

参考文献:

[1] Abraham, Ajith. Business Intelligence from Web Usage Mining[J]. Journal of Information & Knowledge Management, 2003,2(4):375-390.

[2] 费爱国,工新辉.一种基于 Web 日志文件的信息挖掘方法[J].计算机应用,2004,24(6):57-59.

[3] 谢 芳,王 波.基于关联规则个性化推荐的改进算法[J].计算机应用,2006,26:149-151.

[4] 胡可云,陆玉昌,石 纯.基于概念格的分类和关联规则的集成挖掘方法[J].软件学报,2000,11(11):1478-1484.

[5] 石 晶,龚震宇,裘杭萍.基于 Web 使用挖掘的个性化服务系统[J].电子科学大学学报,2002,31(4):399-403.