基于精确序列模式的网页个性化推荐

Web Personalization based on Accurate Sequential Pattern

刘志昆 王卫平 （中国科学技术大学信息管理与决策科学 230026）

摘要：Web 个性化系统的目标是为用户提供他们想要的或需要的信息，而不必明确询问用户的需求。传统推荐系统的方法是直接收集服务器日志作为 Web 使用数据，通过分析用户的行为模式，挖掘出用户的兴趣、偏好，然后将系统认为是与用户相关的网页链接向用户推荐。本文应用新的方法——远程代理收集 Web 使用数据，为数据预处理提供了方便，并提出了精确序列模式的方法进行 Web 页面推荐，扩展了 N－Gram，从而分别提高了网页推荐的准确率和覆盖率。

关键词：序列模式 Web 个性化 Web 挖掘 Web 使用挖掘

1 引言

Web 个性化系统的目标是：为用户提供他们想要的或需要的信息，而不必明确询问用户的需求[1]。由于 Web 个性化不需要向用户询问需求，因此唯一能获得与用户有关的信息（如：动作，偏好等）只有服务器日志（Server Log），或称为 Web 使用数据（Web Usage Data）。因此，在 Web 使用数据上挖掘用户信息的技术——Web 使用挖掘[2]（Web Usage Mining，简称为 WUM）就被提了出来。WUM 的方法包括关联规则（Association Rule），序列模式（Sequential Pattern），聚类（Clustering）等。关联规则指出了一些页面经常被一起浏览尽管他们之间可能并没有直接关联，进而能够挖掘出不同兴趣的用户组；而序列模式是关联规则的扩展，除了关联规则能挖掘出的信息外，它还能挖掘出页面的时间访问顺序；页面聚类能识别出某些网页对于用户来说是可能是内容相关的；而用户聚类则识别某些用户在浏览网站时的相似性行为[3]。目前，基于关联规则的推荐系统已经比较成熟[4,5,6]，聚类分析也有过尝试[7]，Mathias Gery，Hatem Haddad[8] 用现实数据的模拟实验证明了序列模式方法比关联规则方法要好，该结果恰好与这两种方法的实际分析能力相吻合。但是，目前应用序列模式从服务器日志中分析得到的序列模式只是时间序列的网页访问顺序，并非用户浏览网页的实际顺序，如某一网页具体是从哪个网页中链接得到，这一信息并没有被序列模式捕获。因此，本文提出的精确序列模式捕获了这一信息，并且利用该精确序列模式进行 Web 的个性化推荐。

2 利用远程代理采集数据

服务器日志根据 Web 服务器参数设置的不同有着不同的字段[9]，包含了以下字段：客户端 IP 地址，HTTP 方法，URI[10] 资源，URI 查询，所用时间，协议状态，Cookie，日期，时间等。

定义 1：页面视图（Pageview），简称页面，记为 PAGE，指用户在浏览器的地址栏输入 URL 所能看到的页面，是 URI 资源与 URI 查询（URL 地址“?”后面那部分）的组合，例如：URL = http://www.example.com/example.jsp? id = user&pwd = jack，则 PAGE = $\{url_1, url_2, \cdots, url_{|PAGE|}\}$，其中 |PAGE| 为站点 Site 的所有页面总数，$1 \leqslant i \leqslant |PAGE|$。记 |.| 为集合的大小，下同。

定义 2：用户会话，记为 SESSION，指用户在与服务

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

器连接的特定时间里顺序的访问页面，该特定时间过期时长为 25.5 分钟，是由 Catledge 和 Pitkow[11] 首先测量出来的最优会话过期时长。则有会话集合 SESSION = { $SessionID_1$, $SessionID_2$, …, $SessionID_{|SESSION|}$ }。

直接从服务器日志采集的 Web 使用数据并不精确，它存在一些弊端：

（1）当不同的用户使用同一 IP 上网，服务器日志没有办法区别不同的用户会话，因为服务器日志是从 IP 地址识别用户的；

（2）服务器日志中存在非页面的请求信息，如图片，音频，视频，Flash 文件等，给数据预处理带来不便；

（3）不能自动的过虑掉除 GET 外的其他 HTTP 方法。因此，我们用 C Shahabi[12] 的远程代理来采集数据。

本文修改了 Shahabi 的远程代理——RemoteAgent，同样用 Java Applet 实现，功能如下：

（1）当用户初次访问 Site 的页面时下载该代理，下载完成后触发 Applet_Initial 事件，该事件与服务器端代理建立连接，创建一个会话，并从服务器端代理获得一个会话 ID，该会话 ID 由服务器端代理分配，标识唯一，会话过期时长为 25.5 分钟[11]；

（2）接下来用户每次访问 Site 的新页面时都触发 Applet_Initial 事件，该事件检查会话是否结束（即会话过期），如果结束则重新获得会话 ID；

（3）当用户关闭该网页或转向请求其他网页时，触发 Applet_Destroyed 事件，该事件向服务器端代理发送字符串：

会话 ID + 引用站点 URL + 该 Applet 所在页面的 URL，字段之间用空格隔开。服务器端代理则接收该字符串并记录下来，并形成 Web 使用数据。本文只注重用户的访问顺序，而不注重用户在页面上的停留时间，原因有二：

（1）由于现今的个人电脑运行速度极快，内存大增，用户若同时打开多个浏览器访问不同的站点，可能忽略了站点 Site 的页面，而延长了在该页面的停留时间，由此得到页面停留时间显然是不准确的；

（2）即使用户只访问站点 Site，但是当他打开多个浏览器进行访问时，也会造成在某些页面的停留时间过长，同样，由此得到的页面停留时间也是不准确的。

到此，Web 使用数据就收集好了。记 LOG 为 Web 使用数据，则有：

LOG = { log_1, log_2, …, $log_{|LOG|}$ }，对于每一条日志 $log_i \in LOG, 1 \leqslant i \leqslant |Log|$，$log_i$ 包含三个属性：SessionID，Referrer，URL，有 log_i. SessionID ∈ SESSION，log_i. URL ∈ PAGE，log_i. Referer 绝大部分隶属于 PAGE，但是有一部分并不是，这部分就成为数据预处理的主要依据。而对于每个会话都包含一定数目的日志条目：

$SessionID_i = \{ log_1^i, log_2^i, \cdots, log_{|SessionID_i|}^i \}$，其中 $1 \leqslant j \leqslant |SESSION|$。

3 数据预处理

由于本文的数据是从 RemoteAgent 采集来的，它比通常的服务器日志要“干净”得多，因此，数据预处理只需从下面几方面入手：

（1）对于 LOG，在原来顺序不变的情况下按 SessionID 排列；

（2）对于每个 SessionID 中的日志 log_i，如果 log_i. Referer 为空，表示该用户打开浏览器后直接就连接上了 Site 站点，它对挖掘序列模式没有贡献，则删除该条日志；

（3）对于每个 SessionID 中的日志 log_i，如果 log_i. Referer ∉ PAGE 或 log_i. Referer = Site. Homepage，表示用户是从其他站点链接到站点 Site 或用户的访问是从 Site 站点的主页开始的，它对挖掘序列模式同样没有用处，则删除该条日志；

（4）对于每个 SessionID，如果挖掘出来的精确序列长度小于 3，则删除掉整个 SessionID 包含的所有日志条目，因为可以这样认为：用户显然对该站点 Site 没有深入了解，匆匆点击了两下就离开了 Site，因此用户所提供的信息量太少，不足以发现用户的兴趣。该步骤与精确序列模式挖掘同时进行。

4 精确序列模式挖掘

首先，提取精确序列。见示例 1，精确序列与以往的序列之间的区别。

示例 1：提取精确序列，数据见表 1。

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

表 1 两种不同的 Web 使用数据中的一个会话记录

远程代理	服务器日志
100 – a	222.222.222.xxx GET a 17:40:12
100 a b	222.222.222.xxx GET b 17:40:51
100 a c	222.222.222.xxx GET c 17:41:23
100 b d	222.222.222.xxx GET d 17:42:35
100 d e	222.222.222.xxx GET e 17:43:57
100 c f	222.222.222.xxx GET f 17:45:36
100 d g	222.222.222.xxx GET g 17:47:08
100 g h	222.222.222.xxx GET h 17:48:59

一般的序列从服务器日志中挖掘出来的是 a→b→c→d→e→f→g；而本文要挖掘的是精确的序列，即每个页面具体是由哪个页面链接而来的，其精确的序列如图 1。

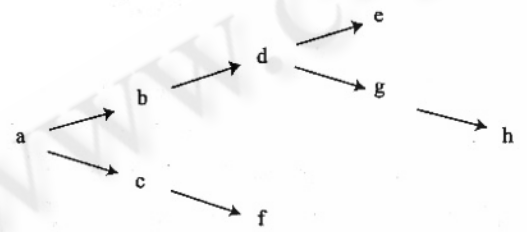

图 1 精确的序列模式

从中可以看出，精确序列从一个会话中挖掘出来的序列可能不只一个，可能是多个：a→b→d→e、a→b→d→g→h、a→c→f，简记为：abde、abdgh、acf。由这样的序列形成的集合记为 S。

其次，我们将从精确序列中提出精确序列模式，序列模式是一种 N – Gram[13]，它由用户在与服务器会话期间形成的日志中挖掘而来的。

定义 3：精确序列模式（ASP），给定前缀长度为 |A| 的 URL 序列，后缀长度为 |C| 的 URL 序列，及前缀序列和后续序列距离为 n，定义如下：

$$A \xrightarrow{n} C : freq$$

$$A := url_1, \cdots, url_{|A|}$$

$$C := url_k, \cdots url_{k+|C|-1}$$

$$k = |A| + n, n \in N^+$$

精确序列模式表达了这样一个关系：某个用户最后 |A| 次顺序点击了页面集合 A，则再经过 n 次点击，集合为 C 的页面将被用户请求浏览。由相同的前缀长度 |A| 、后缀长度 |C| 及前后缀距离 n 组成的序列模式集合记为 ASP(|A|,n,|C|)，对于 ASP(|A|,n,|C|) 中的每个模式记为：

$$Pat(A,|A|,n,|C|) => C : freq$$

其中 freq 为该模式的频度。为了解决部分推荐系统覆盖率[6]偏小的问题，在精确序列模式挖掘的过程中，本文扩展了 N – Gram，提取了 |C| 的长度为 1 和 2 的结合，且长度 2 优先。即，若精确序列能提取后缀长度为 2 的子序列就提取出该模式，否则提取出后缀长度为 1 的序列模式。将这种精确序列模式记为：

$$Pat(A,|A|,\underline{n}) => C : freq$$

通常省略掉参数 |A|，n，记为

$$Pat(A) => C : freq$$

在集合 ASP(|A|,n) 抽取具有相同前缀的 A 的模式，形成一个个小模式集合：

$$Pat(A) => \{C_1 : freq_1; C_2 : freq_2; \cdots, C_n : freq_n\}$$

易见，此时的 Pat(A) 是附有频度的页面集合，是页面集合 PAGE 的一个子集，即：

$$Pat(A) = \{p_l : \sum_{l=1}^{n} freq_l * \delta(p_l)\}$$

$$\delta(p_l) = \begin{cases} 1, if\ p_l \in C_l \\ 0, if\ p_l \notin C_l \end{cases}$$

本文后面如果没有特别说明，Pat(.) 指的就是这样的页面集合。

示例 2：精确序列模式提取，设可用精确序列为 S = {abde,acbde,abe,abed}，则有 ASP(1,1) = {(a→bd:1),(b→de:2),(d→e:2),(a→cb:1),(c→bd:1),(a→be:2),(b→ed:1),(b→e:1),(e→d:1)}

ASP(1,2) = {(a→de:1),(b→e:2),(a→bd:1),(c→de:1),(a→e:1),(a→ed:1),(b→d:1)}

ASP(2,1) = {(ab→de:1),(bd→e:2),(ac→bd:1),(cb→de:1),(ab→e:1),(ab→ed:1),(be→d:1)}

ASP(2,2) = {(ab→e:1),(ac→de:1),(cb→e:1),(ab→d:1)}

最后，对所得的模式集合进行剪枝提取频繁序列模式。设一个最小的阀值 min_freq，将所有频度 freq 小于该阀值的模式全部剪掉。

5 网页推荐算法

网页推荐的目的是减少用户在网站上的搜索页面

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

的时间，即减少用户为请求最终页面而访问的中间过渡页面，根据用户的当前行为模式分析出用户的偏好并向其推荐相关的页面。因此，首先要获得并分析用户的行为，才能向用户推荐相关页面。目前向用户推荐页面使用最多的是 Top－N 方法，即向用户推荐相关度最大的 N 个页面，动态地把它们加入到用户当前的访问页面。为了减少用户访问的页面数，采用[13]的 ASP(2,2)，即，只要获得用户前连续两次请求就可对其推荐相关网页。算法如下：

(1) 初次获得用户请求的两个连续行为序列 Seq；

(2) 记录所得的行为序列 Seq，此时用户所访问的页面集合记为 Visited；

(3) 从精确序列模式集合 ASP(2,2)中找出模式 Pat(Seq)的集合 Pattern；

(4) 将集合 Pattern 除去已访问的页面 Visited 后，将频度为前 Top－N 的页面 Top_N_Pages 向用户推荐；

(5) 用户再次请求页面时重复步骤(2)(3)，将得到的新集合 Pattern'，并上 Top_N_Pages，再除去新的页面集合 Visited'后，再将频度为前 Top－N 的页面集合向用户推荐；

(6) 不断的重复步骤(5)，直到用户与 Site 失去连接。

6 结论

本文在传统 Web 推荐系统的基础上，提出了一种新的方法进行 Web 页面的推荐——精确序列模式，提高网页推荐的精确率，并扩展了 N－Gram 来增加网页推荐的覆盖率，从而提高了整个推荐系统的性能。下一步工作，将在实际的网站中应用该推荐模型，以此验证推荐模型的结论，并进一步完善该推荐模型。

参考文献

1 MULVENNA M D, ANAND S S, BUCHNER A G.. Personalization on the net using web mining. Commun. ACM, 2000, 43: Pages 123 - 125.

2 R Cooley, J Srivastava, B Mobasher. Web mining: Information and pattern discovery on the world wide web. In 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI' 97), November 1997.

3 M Eirinaki, M Vazirgiannis. Web Mining for Web Personalization. ACM Transactions on Internet Technology, February 2003. Vol. 3, No. 1: Pages 1 - 27.

4 Bamshad Mobasher, Honghua Dai, Tao Luo, Miki Nakagawa. Effective Personalization Based on Association Rule Discovery from Web Usage Data. WIDM 2001, GA, USA.

5 Xiaobin Fu, Jay Budzik, Kristian J Hammond. Mining Navigation History for Recommendation. Proceedings of the 5th international conference on Intelligent user interfaces. 2000.

6 AM Ahmad, MHA Hijazi. Web Page Recommendation Model for Web Personalization. LNAI 3214, 2004. Pages 587 - 593.

7 KA Smith, Alan Ng. Web page clustering using a self－organizing map of user navigation patterns. Decision Support Systems 35 (2003) Pages:245－256.

8 Mathias Gery, Hatem Haddad. Evaluation of Web Usage Mining Approaches for User's Next Request Prediction. WIDM'03, 2003, New Orleans, Louisiana, USA.

9 http://www.w3.org/TR/WD－logfile.html

10 http://www.w3.org/Addressing/

11 Catledge L, Pitkow. Characterizing browsing behaviors on the world wide web. Computer network and ISDN Systems 27 pages 1065－1073,1995.

12 Cyrus Shahabi, Farnoush Banaei－Kashani, Jabed Faruque. A reliable, efficient, and scalable system for web usage data acquisition. WebKDD'01 Workshop, ACM－SIGKDD 2001, San Francisco, CA.

13 E Frias－Martinez, V Karamcheti. Sequential pattern mining from web usage data. In WEBKDD Workshop: Web Mining for Usage Patterns and User Profiles, July 2002.

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn