# 采用反馈机制的自适应Web推荐系统

王 悦,周国祥,朱子荣
(合肥工业大学 计算机与信息学院,安徽 合肥 230009)

**摘 要**:Web服务的初衷是能让用户能够快捷方便地找到自己需要的资源,但是当前过量的网络资源反而增加了用户的麻烦。于是出现了结合用户对Web使用的数据挖掘应用的Web推荐系统。在传统Web推荐系统的基础上引入了反馈机制。使推荐系统在线给用户提供推荐功能的同时,对自身的推荐机制改进,更体现了个性化服务的灵活性。

**关键词**:Web使用挖掘;Web推荐系统;反馈机制
**中图分类号**:TP393 **文献标识码**:A **文章编号**:1673-629X(2007)07-0133-04

## Adaptive Web Recommendation System Based on Feedback Mechanism

WANG Yue, ZHOU Guo-xiang, ZHU Zi-rong
(School of Computer and Information, Hefei University of Technology, Hefei 230009, China)

**Abstract**: The original intention is to make user can find the resources which he wanted more conveniently and quickly, but huge Web data is overwhelming users now. Web recommendation system is designed to face such problems, using data mining technology. This paper adds a feedback mechanism into Web recommendation system; make the system can update its recommendation while doing recommendation.

**Key words**: Web usage mining; Web recommendation system; feedback mechanism

## 0 引 言

随着互联网的迅速发展,Web上的资源呈爆炸性增长,Web服务提供商也越来越多。在用户有了更多选择的情况下,Web运营商只能通过提高自身的服务来吸引用户。现在的用户不再只满足于仅仅把所有的信息资源顺序列出,他们还希望Web站点能以更个性化的方式把信息推荐给他们。Web推荐系统[1]根据这种需求而产生,主要思想就是通过用户历史访问路径以及当前访问状态,运用数据挖掘的手段,预测出用户下一步准备访问的页面并推荐出来,给用户提供方便,以及人性化的感觉。

现有的Web推荐方法主要分为3类:

(1)采用Web内容过滤(Content Based Filtering)的方式,通过比较资源与用户描述文件来推荐资源,关键问题是相似度统计。这块比较流行的做法是使用矢量空间模型,通过统计余弦度量实现。缺点是难以区分资源内容的品质和风格,且不能为用户发现新感兴趣的资源;

(2)采用协同过滤(Collaborative Filtering)[2]的方式,根据用户初始输入的User profile来进行分类,提供个性化服务,这种做法由于不能动态发现用户的访问模式,有局限性;

(3)采用Web使用挖掘(又称Web日志挖掘)[3]的方式,这种做法通过挖掘Web使用日志以便于了解用户读取网页的模式。

早期,学者们尝试了各种统计方法来挖掘大量数据中的用户模式:文献[4]中,提出了使用Apriori算法来找出用户访问序列中满足条件的频繁项。后来文献[5]对算法进行了改进;文献[6]采用改进的聚类算法PageGather来对用户模式进行分类,归类的是用户没有访问到的页面,思路较新颖;在看到前面两种做法的优缺点后,出现了很多结合两者优点的混合系统,文献[7]中介绍的就是这类系统。

通过对前面方法的分析,发现当前Web推荐系统的算法处理复杂,基本上只考虑脱机处理以及对比较早的历史记录处理的情况;而用户对推荐后作出的反应很少考虑,更没有考虑站点新加入的页面用户是否感兴趣。因此文中结合Web使用挖掘。在基于文献[8]中动态自适应推荐系统,和Fish算法[9]的基础上,提出了一种带反馈Web推荐系统解决方案,以解决当前推荐系统中很少考虑到的对用户得到推荐后反馈情

**收稿日期**:2006-07-13
**基金项目**:安徽省自然科学基金资助项目(050420202)
**作者简介**:王 悦(1981-),男,四川成都人,硕士研究生,研究方向为数据挖掘与决策支持系统;周国祥,副教授,硕士生导师,研究方向为决策支持系统。

况,以及实时性问题。

## 1 用户访问模式的挖掘

### 1.1 对 Web log 的预处理

经过 Web 站点的运营以后,Web 服务器会自动记录下很多用户对站点资源访问的原始资料。Web 使用挖掘就是建立在对这些用户使用记录进行处理的基础上。但是这些原始数据不能直接处理,因为其中包含很多错误信息、缺失信息以及一些与挖掘用户事务模式无关的文件访问,而使用挖掘只需要日志中用户访问页面的记录,如网页中包含图片(JPEG,GIF)、脚本等文件则应该作为干扰项去掉,这一步类似 DW 装载数据时的数据清洗。

### 1.2 用户事务的识别

经过清洗后的数据就能够用来发掘用户模式了。由于目前的数据只是如(日期,客户端 IP,字节,服务器,请求,状态,服务名,耗用时间,用户代理,Cookie,参照页)的多元组,因此还需要识别出其中用户的访问事务才能更好地发掘出用户模式。

定义 1:用户访问事务 $t = \{\text{tranID}, \text{UserID}, P\}$;$P = \{\text{url1}, \text{time1}; \text{url2}, \text{time2}, \cdots\}$。其中 tranID 表示事务的 ID 号,UserID 表示用户的 ID 值(缺省情况下以用户 IP 替代),而 $P$ 为用户在该事务范围内访问过所有的页面(链接) 以及时间点。

通过对划分在固定时间段内用户访问进行统计,把经过预处理的 Web log 划分为用户访问事务的集合。

### 1.3 用户模式的发掘

除了 Apriori 算法,模式发掘方式上已有多种新兴的算法,包括对 Apriori 算法的改进。但从实现角度等原因,这里沿用经典的 Apriori 算法[4]。

下面是该算法中一些基本定义,由于文中只涉及用 Apriori 算法生成频繁项集,故没对关联规则产生描述。

定义 2:对 $\exists R \in P$,对事务 $t$ 支持度为 $\text{support}(R) = P(R \cup t)$,$R$ 对事务集 $T$ 的支持度为 $\text{support}(R) = \sum_{t \in T} P(R \cup t)$。

定义 3:频繁项集 $F$ 的项数 $r$ 为平凡项集中,包含在所出现频繁项里子项的个数。

定义 4:频繁项集 $F$ 为任一满足对事务集 $T$ 的支持度超过最小支持值的用户访问集合。其中当 $|F| = k$ 时,$k$ 项频繁集 $F$。

根据算法迭代求集,得到每个用户的最大频繁项集。即每个用户在一段时限里访问最多的页面项,一般以天为时间单位计算。

## 2 反馈推荐系统的基本思路

### 2.1 RF(Recommendation Feedback)算法思想

RF 算法的设计主要考虑了为用户推荐页面项,用户点击,再为用户推荐项是一个循环过程。能迅速发现用户对推荐的喜欢程度并把它反馈到推荐项中去,充分地体现 Web 推荐系统人性化。具体做法分 4 步:用户分类、生成初始推荐集、生成推荐、权值调整。

### 2.2 具体实现的方法

用户分类:根据 User profile 中对用户进行分类,用户在初始注册时会填入一些系统要求的类别信息,根据这些信息把所有的用户分类开。

生成初始推荐集:采用 Apriori 算法求出同一类别用户所有的频繁项集,再通过对这些频繁项集的统计,得到支持度达到最小支持度(min_support)的页面项。把这些项都归入一个集合作为初始推荐集 $C$。注意:由于 $C$ 比较庞大,一般是站点在脱机情况时处理生成。

定义 5:初始推荐集 $C = \{p_i \mid \text{support}(p_i) > \text{min\_support}, \forall p_i(P1 \cup P2 \cup P3\cdots)\}$,其中 $p_i$ 为页面项,$P_i$ 为用户 $i$ 的最大频繁项。

初始推荐集生成的具体操作为:假设当前服务器上有如表 1 中所示的事务集,其中 Tans_name 表示事务的名称,后面的 Pages 代表本事务中所有能访问到的页面。

表 1 访问事务集

| Tans_name | Pages |
|---|---|
| t001 | ACDEG |
| t002 | ABCEF |
| t003 | ADEF |
| t004 | BCDEFG |
| t005 | ABCDG |
| t006 | BCEF |

采用 Apriori 算法对其分析得出所有的频繁项集之前,需要对 Apriori 算法的参数即最小支持度(min_support)进行设定。通常 min_support 的值设为 60%。其意义为:设当前集合中总共有 $n$ 个事务,则满足该最小支持度的频繁项应该首先满足以下条件,即:该项在集合中所有事务中的出现频率(次数) 要大于等于 $60\% * n$。对上表中的假设,出现频率大于等于 $60\% * 6 = 3$ 的项为频繁项。

表 2 是根据 Apriori 算法对表 1 中事务集得出的结果(按项数排列)。

生成推荐列 $L$:对初始推荐集合中所有项按支持度高低加权排序得到推荐列。选出 $n$ 项推荐给用户。下面是 $L$ 的定义:

表 2　Apriori 频繁集结果

| 1 项频繁集 | 2 项频繁集 | 3 项频繁集 | 4 项频繁集 |
|---|---|---|---|
| A(4) | AD(3) | BCE(3) | BCEF(3) |
| B(4) | BC(4) | BCF(3) | |
| C(5) | BE(3) | CDG(3) | |
| D(4) | BF(3) | CEF(3) | |
| E(5) | CD(3) | | |
| F(4) | CE(4) | | |
| G(3) | DE(3) | | |
| | DG(3) | | |
| | EF(4) | | |

定义 6:$L$ 为 $n$ 阶动态推荐列,如果有:$L = \{l_1, l_2, l_3, \cdots, l_n\}$。其中 $l_i = \{p_i, w_i \mid w_i = \text{support}(p_i) \text{ MOD } Z\}$。称 $w_i$ 为 $p_i$ 项的权值,$Z$ 为控制参数,控制权值的长度。$L$ 为一个队列,满足 FIFO 特性,队列中 $l_i$ 按权值从大到小排列,权值一样时照时序排列。$L$ 首项和末项的权值设定界限[Max,Min]。

权值调整:通过前面的准备工作后,推荐系统为 Web 站点运行时提供了一系列(不会太多)的按照用户分类的动态推荐队列。

下面针对一个用户分类的推荐队列 $L$ 说明反馈机制的实现,这是整个算法的核心。

①For (every $t$ of times)

{

②foreach $l_i$ in $L$

{

③$l_i.w_i = l_i.w_i - k$;//每循环一次,权值自动减小 $k$

}

④$L$.Out(); //从队列尾出列最小权值的项

⑤$L$.Insert (new $l$ );//从队列头插入新项,权值设置为最大

⑥$L$.AjustWeight();//调整权值

⑦$L$.Sort(); //排序

⑧$L$.recommed($n$)//把推荐列前 $n$ 项推荐给同类用户

}

这里主要运用用户对推荐链接的点击来预测用户对推荐系统的反馈情况。再根据这些信息以及 Web 站点的变化信息对推荐项进行调节,以实现 Web 推荐系统的自适应调整。算法利用了一个加权的动态队列,将各种反馈影响转变为对权值调整实现动态推荐。

注意:⑤$L$.Insert (new $l$) 为插入新项,新项的来源除了来自初始推荐集 $C$,还有 Web 站点新加入的页面项,Web 站点的动态变化也是对推荐集有影响的重要因素之一(尤其是新闻发布系统,信息更新较快)。

其中用户的反馈性主要体现在 ⑥$L$.AjustWeight(),这步完成的是对 $L$ 中所有项权值的调整。调整的依据是通过对 Web log 中的事务再统计,得到生成推荐列时间点后用户的访问情况。然后把用户访问中对推荐过的项的点击情况反馈给权值,具体做法是:假设同类的用户对推荐出的项点击了 $m$ 次,则该项权值 $w = w + m \text{ MOD } Z$,如果计算后 $w > L.\text{head}.w$,$w = L.\text{head}.w$;反之,没有点击到的推荐项的 $w = w - k$,如果计算后的 $w < L.\text{tail}.w$,$w = L.\text{tail}.w$。

以上的算法是在站点运行时,实时为用户提供反馈式的 Web 推荐服务。具体的运行间隔视服务器的性能以及在线用户的数目而定。此外每次站点脱机后,还需要从 Web log 挖掘出用户模式,建立新的初始动态队列 $L$ 为实现下次在线推荐做好准备,这一步是至关重要的。

## 3　系统分析评价

系统用户模式的识别采用了成熟的 Apriori 算法,识别准确率、可靠性较高;其 RF 算法是系统实现实时推荐的关键,算法中提出加权反馈,并实现了把用户反馈映射到系统模型中,即:对使用到的链接,使其加深(加大权值),没有使用到的链接,加速其消亡(出列)的速度。这样推荐列的准确度,以及推荐出链接的个性化程度将随着用户的使用而逐渐增强;算法还把实时反馈的思想引入到 Web 推荐系统中,即考虑新加入的用户没点击过但可能感兴趣的页面。此外在设计时,尽量减少了在实时运行时系统的大复杂度的计算,使系统在实现反馈的同时,并没有过度的增加时空开销,减小了服务器的压力。这在实时性和性能矛盾的现状下,不失为一个折中的做法。

笔者搭建了一个应用在新闻发布领域的反馈式自适应 Web 推荐系统。并采用 Mobasher 的评价测度[5]对本系统的推荐性能进行评价:该测度主要通过对推荐系统推荐覆盖率、准确率进行综合比对来评价系统的推荐性能。在试验中,发现 RF 算法在系统运行初期的推荐覆盖率与准确率较低,但经过一段时间的反馈学习,测度能比较快且平稳地达到一个固定的值;此外系统处理用户对系统推荐的页面的反馈(即用户是否点击推荐页面)以及新加入项的情况建立推荐模型也能达到不错的效果。由此证明采用 RF 算法的反馈推荐系统不但能完成普通 Web 推荐系统功能,还能实现实时性要求高的环境下的 Web 推荐。

## 4　小　结

提出了采用反馈机制的自适应 Web 推荐系统的一种解决方案,并详细描述了核心 RF 算法的实现。

解决了当前 Web 推荐系统中无法处理用户对推荐服务的反馈情况,以及 Web 站点动态更新后对整个系统的影响情况。在 Web 站点中,本身靠信息的更新吸引用户。其中以新闻发布系统的页面项时效性强,新页面的实时更新对站点的生存至关重要。如果不能即时为用户推荐最新的用户感兴趣的页面,新闻发布就失去了意义。因此文中提出的反馈式自适应系统的 RF 算法在考虑到用户反馈对推荐影响的同时,还考虑了实时性的因素。研究的下一步目标是结合 Web 内容过滤 CBF 和统计学的原理进一步探索 Web 挖掘在基于 B/S 模式下的 MIS 系统中的具体应用。

**参考文献:**


[1] Jin Xin, Mobasher B, Zhou Yanzan. A Web Recommendation System Based on Maximum Entropy[C]//Proceeding of the International Conference on Information Technology: Coding and Computing (ITCC2005). Washington DC: IEEE Computer Society Press, 2005:1 - 6.

[2] Sarwar B, Karypis G, Konstan J, et al. Item - based collaborative filtering recommendation algorithms[C]//In: Proceedings of the tenth international conference on world wide web. New York: ACM Press, 2001:285 - 295.

[3] Mobasher B, Cooley R, Srivastava J. Automatic Personalization Based on Web Usage Mining[C]//Communications of the ACM. New York: ACM Press, 2000:142 - 151.

[4] Agrawal R, Imielinski T, Swami A. Mining Association Rules between Sets of Items in Large Databases[C]//In Proc of the ACM SIGMOD Conference on Management of Data. New York: ACM Press, 1993:207 - 213.

[5] Mobasher B. Effective Personalization Based on Association Rule Discovery from Web Usage Data[C]//Proc. 3rd ACM Workshop Web Information and Data Management (WIDM 2001). New York: ACM Press, 2001:9 - 15.

[6] Perkowitz M, Etzioni O. Adaptive Web sites: automatically synthesizing Web pages[C]//In Proceedings of Fifteenth National Conference on Artificial Intelligence. USA: AAAI Press, 1998:15 - 21.

[7] 戴东波,印 鉴.结合使用挖掘和内容挖掘的 Web 推荐服务[J].计算机工程与应用,2005(18):162 - 165.

[8] Ou Jian Chih, Lee Chang - Hung, Chen Ming - Syan. Web log mining with adaptive support thresholds[C]//WWW(Special interest tracks and posters). Chiba, Japan: ACM Press, 2005: 1188 - 1189.

[9] De Bra P, Post R. Searching for Arbitrary Information in the World - Wide Web: the Fish - Search for Mosaic[C]//Second WWW Conference. Chicago: ACM Press, 1994:45 - 51.


(上接第 86 页)

PHONE 主动发起呼叫的过程类似,在此不再赘述。

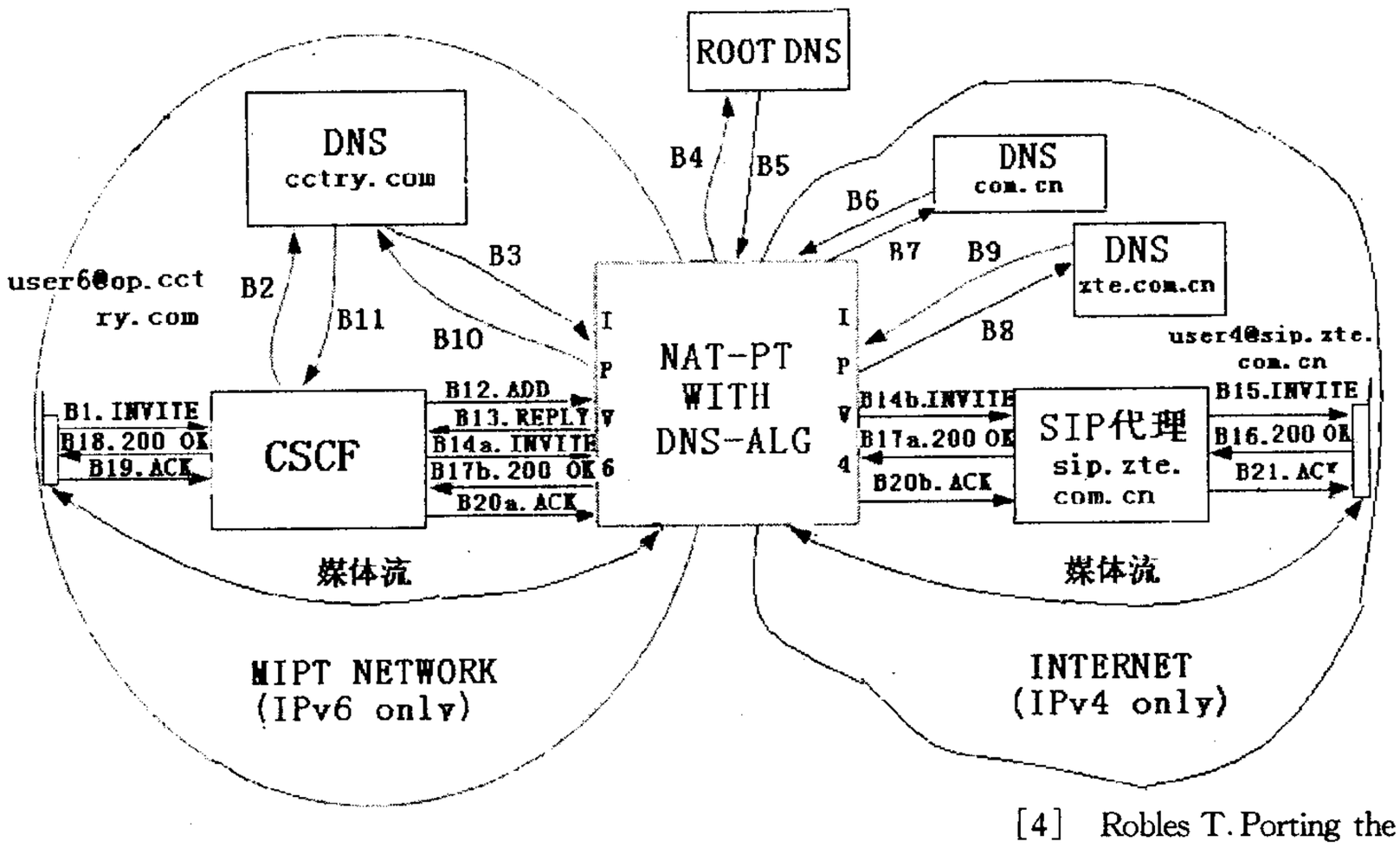


图 3 MIPT 中的 IPv6SIP PHONE 主动发起呼叫

## 5 结束语

通过一个具体的例子,对 3G 中基于 IPv6 的 SIP 电话,如何与目前正广泛使用的基于 IPv4 的 SIP 电话进行正确的通讯进行了描述。

**参考文献:**


[1] Deering S, Hinden R. Internet Protocol Version 6. 0. RFC 1883 IETF[S]. 2003.

[2] Rosenberg J, Schulzrinne H, Camarillo G. SIP: Session Initial Protocol. RFC3261[S]. 2002.

[3] Gabor B, Krisztian K. SIP sessions between IPv4 and IPv6 clients and SIP based call setup in 3GPP IMS with NAT in place [J]. Communications, IEEE Transactions, 2002, 45 (9):468 - 477.

[4] Robles T. Porting the Session Initiation Protocol to IPv6[J]. Networking, IEEE/ACM Transactions, 2003, 33(4):1121 - 1132.

[5] Wilijakka J. Analysis on IPv6 Transition in 3GPP Networks [EB/OL]. 2004 - 05. draft - ietf - v6ops - sgpp - analysis - 10. txt.

[6] Nakakima M, Kobayashi N. IPv4/IPv6 translation technology, IETF[S]. 2003.