文章编号:1671－251X(2012)08－0100－05
郭孝园,何臻. 基于 Web 日志的煤矿企业网站个性化推荐服务研究[J]. 工矿自动化,2012(8):100-104.

基于 Web 日志的煤矿企业网站个性化推荐服务研究

郭孝园[1], 何臻[2]

(1. 中煤科工集团常州自动化研究院, 江苏 常州 213015; 2. 中国矿业大学信电学院, 江苏 徐州 221008)

摘要:为了解决煤矿企业网站用户查找信息难的问题,提出了一种基于 Web 日志的煤矿企业网站个性化推荐服务模型。该模型应用关联规则对新用户进行页面推荐,应用聚类算法对老用户进行页面推荐;并结合点击网页的次数、网页的浏览时间、雅可系数与最长公共路径系数来度量用户兴趣度的方法,可为用户准确地推荐其感兴趣的页面。测试结果表明,该模型能够有效地对网页资源进行分类并进行个性化推荐。

关键词:企业网站; Web 日志; 网页分类; 个性化推荐; 个性化服务

中图分类号:TD67 **文献标识码**:B **网络出版时间**:2012－08－03 13:58

网络出版地址:http://www.cnki.net/kcms/detail/32.1627.TP.20120803.1358.025.html

Research of Personalized Recommendation Service of Coal Enterprises Web Site Based on Web Log

GUO Xiao-yuan[1], HE Zhen[2]

(1. Changzhou Automation Research Institute of CCTEG., Changzhou 213015, China.
2. School of Information and Electrical Engineering of CUMT., Xuzhou 221008, China)

Abstract: In order to solve problem of difficult finding information for users of coal mine enterprise Web site, the paper put forward a personalized recommendation service model of coal enterprise Web site based on Web log. The model uses association rules to recommend page for new users, and uses clustering algorithm to recommend page for old users. And it measures interesting degree of users with click number of Web page, browsing time of Web page, Jacobian coefficient and the longest public path coefficient of a user to accurately recommend interested pages for the user. The test results showed that the model can classify Web resources and provide personalized recommendation.

Key words: enterprise web site, Web log, classification of Web pages, personalized recommendation, personalized service

0 引言

企业网站不仅是企业的一个信息发布平台,同时也是一个企业向外展示的窗口。最近几年煤矿企业发展迅速,大型煤矿企业的业务早已不仅仅局限于煤炭行业。例如,神华集团、河南煤业化工集团的业务涉及煤炭、化工、有色金属、装备制造、物流贸易、建筑矿建、现代服务业等板块。公司的网站越来越复杂,提供信息越来越丰富。然而,网站在向户提供丰富的信息的同时,用户查找自己感兴趣的信息却越来越困难,产生了“信息迷航”问题。

为了解决“信息迷航”问题,研究者们提出了 Web 站点的个性化服务,即在用户浏览页面之前,事先为用户定制其感兴趣的内容,并且站点尽可能地自动调整自己,包括内容和链接结构,来迎合用户的浏览兴趣。

传统的企业网站为了给用户提供个性化服务,多采取问卷调查的方式。然而这种方式不仅触及到

收稿日期:2012－06－12

作者简介:郭孝园(1982－),男,江苏徐州人,助理工程师,硕士,现主要从事自动化设备的销售、管理工作。E-mail: guoxiaoyuan1982@126.com

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

用户的隐私，使得部分用户不愿提交个人兴趣信息，而且影响了浏览速度和使用体验。

1995 年，卡内基梅隆大学 Robert Armstrong 等人提出了个性化导航系统 WebWatcher；麻省理工学院 Henry Lieberman 提出个性化导航智能体 Letizia；斯坦福大学的 Marko Balabanovic 等推出了个性化推荐系统 LIRA。这 3 个系统标志着个性化服务的开始[1]。最近几年，已有一些商业化的 Web 日志挖掘软件，如 WebTrends、AWStats、逆火 Web 日志分析器等。但是这些系统的功能还很有限，性能有很大的局限性，离大规模的商业应用还有一定的距离。因此，基于 Web 日志的个性化服务研究还有很大的研究空间。

本文提出了一种应用于煤矿企业的基于 Web 日志的个性化推荐服务模型，并以河南煤业化工集团的服务器日志数据为实验数据进行了测试。

1　推荐服务模型

基于 Web 日志的个性化推荐服务模型由数据预处理、离线分析和在线推荐 3 个部分组成。为了准确地反应用户兴趣度，尽可能多地考虑多种衡量用户兴趣度的指标，本文提出了一种综合考虑点击网页的次数、网页的浏览时间、雅可系数与最长公共路径系数来度量用户兴趣度的方法。个性化推荐服务模型架构如图 1 所示。

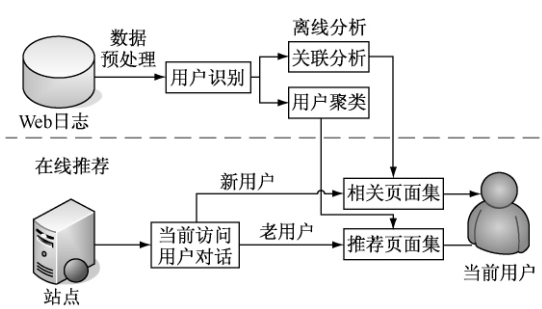

图 1　个性化推荐服务模型组成

模型首先对 Web 服务器日志进行预处理，利用启发规则进行用户识别。然后利用关联规则对网页资源进行分类生成相关页面集。对用户进行聚类，并生成每类用户的推荐页面集。当用户访问网站时，模型先判断用户是新用户还是老用户，如果是老用户则判断该用户是哪一类用户，然后将该类用户共同感兴趣的推荐页面集推荐给该用户；如果是新用户则将利用关联规则生成的相关页面集推荐给用户。

1.1　数据预处理

原始的 Web 日志具有冗余性、不完整性、杂乱性、模糊性的特点，不适合直接用于 Web 挖掘，因此，要对数据进行预处理。数据预处理的目的是把 Web 日志转化为适合进行数据挖掘的可靠的、精确的数据。数据预处理是个性化推荐的基础，数据预处理的结果影响着个性化推荐的准确性。

数据预处理过程主要包括数据清洗、用户识别、会话识别、路径补充 4 个步骤[2]，如图 2 所示。

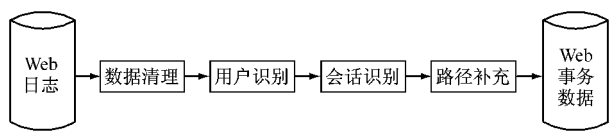

图 2　Web 日志预处理步骤

1.1.1　数据清洗

Web 日志中有很多记录，每条记录有很多属性，并非所有的记录、所有的属性都对 Web 日志挖掘有用。数据清洗就是删除一些无用的记录和一些无用的属性。数据清洗包括以下几步[3]：删除错误请求项、删除由 Spider（网络蜘蛛）或 Agent（代理服务器）产生的请求项、删除网页附属资源及删除无关列。

（1）删除错误请求项

错误请求项是带有“error”或“failure”状态的项，这些请求对挖掘过程是无用的。可通过检查状态码（sc-status）去除这种请求。当 sc-status＝200 时，表示一切正常，对 GET 和 POST 请求的应答文档跟在后面。一般地说，只有以 2 开头的状态码表示成功，其它状态码都表示错误或失败。在 Web 日志清理时，应该删除状态码不是以 2 开头的请求项。

用户请求方法（cs-method）字段通常有 GET、POST、HEAD。其中只有 GET 是指请求指定的页面信息，并返回实体页面。应该删除 cs-method 字段不是‘GET’的记录。SQL 查询语句如下：

```
delete from source where sc_status not like '2__'
delete from source cs_method not like 'GET'
```

（2）删除由 Spider 或 Agent 产生的请求项

Spider 是搜索引擎用来扫描网络上超链接的一种软件。例如用户在使用百度进行搜索时，就会调用 Spider 扫描网站。这种请求记录对日志分析没有用处，应该删除。由于 Agent 的同一个 IP 通常有多个用户，这样的请求难以进行用户识别，也应删除。SQL 查询语句如下：

```
delete from source where User_Agent like '%crawl%'
delete from source where User_Agent like '%spider%'
```

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

```
delete from source where User_Agent like '%slurp%'
delete from source where User_Agent like '%bot%'
delete from source where User_Agent like '%robot%'
```

（3）删除网页附属资源

页面中有时包含图形、声音、动画以及广告等文件，然而用户一般不会主动式地请求这些附属文件，当用户访问网页时，这些文件也会自动下载并记录在日志中。因此，应删除扩展名为.bmp、.js、.css等记录。另外，扩展名为.cgi的脚本文件的索引日志记录也应该被删除。SQL查询语句如下：

```
delete from source where cs_uri_stem like '%.gif'
delete from source where cs_uri_stem like '%.jpg'
delete from source where cs_uri_stem like '%.swf'
delete from source where cs_uri_stem like '%.css'
delete from source where cs_uri_stem like '%.js'
delete from source where cs_uri_stem like '%.bmp'
delete from source where cs_uri_stem like '%.ico'
delete from source where cs_uri_stem like '%.png'
delete from source where cs_uri_stem like '%.jpeg'
delete from source where cs_uri_stem like '%.cgi'
```

（4）删除无关列

Web日志有很多数据属性，但只有date、time、cs-method、cs-uri-stem、cs-uri-query、c-ip、cs(Referer)、sc-bytes和time-taken与Web日志挖掘有关，其它列应该删除。

1.1.2 用户识别

识别出每个访问网站的用户，这一任务因为公司防火墙和代理服务器、网吧、机房的存在而变得复杂。可利用如下的启发式规则来识别用户：

规则1：不同的IP地址代表着不同的用户。

规则2：当IP地址相同时，不同的操作系统代表不同的用户。

规则3：当IP地址相同时，不同的浏览器代表不同的用户。

规则4：在IP地址相同，用户使用的操作系统和浏览器也相同的情况下，则可以根据网站的拓扑结构图对用户进行识别。如果用户请求的某个页面不能从已访问的任何页面到达，则判断这是一个新用户。

1.1.3 会话识别

会话识别[4]就是将用户的访问日志记录分为单个的会话。本文根据相邻请求的时间差和页面的信息量来识别会话，计算公式如下：

$$\frac{Time_i - Time_{i-1}}{Bytes} < \alpha \tag{1}$$

式中：$Time_i - Time_{i-1}$表示相邻请求时间间隔；$Bytes$代表页面信息量以字节为单位；α为一阈值。

1.1.4 路径补充

将日志文件中没有的，但又反映了用户行为的数据补充进来，这就是路径补充所做的工作，解决的方法类似于用户识别中的方法。如果当前请求的页与用户上一次请求的页之间没有超文本链接，那么用户很可能使用了浏览器上的“BACK”按钮调用缓存在本机中的页面。

原始数据来自于河南煤业化工集团公司服务器，时间段为2011－08－16—29共1 932 696行日志记录。经数据预处理：剩下81 488行记录，识别出15 448个用户，去除访问次数少于5次的用户，识别出1 635个页面。

1.2 离线分析

离线分析主要是分析用户浏览模式，是个性化推荐服务的核心。离线分析由2个部分组成：关联分析和用户聚类。关联分析是为了得到网页的相关页面集，用于新用户的页面推荐集[5]；用户聚类是将用户按兴趣度聚为各类，然后在此基础上得到各类用户的推荐页面集。

1.2.1 关联分析

关联分析主要由于下几个部分组成：

（1）网站角色设置。在进行Web挖掘之前需要先对网站的角色进行划分，角色对应的是网站的具体资源，也就是那些URL的信息。而网站的设计是一个树形结构，一个角色不止对应一个URL，可以多个URL对应一个角色，但是不允许一个URL对应多个角色。把URL资源存入表$DB_{URL-Role}$中，如表1所示。

表1 表$DB_{URL-Role}$结构

列名	属性	允许为空	含义
Aid	int	否	唯一标识符，自动增长
URL	varchar(200)	否	网页网址
actor	varchar(200)	否	角色名称

（2）用户的频繁角色集合。在频繁模式挖掘中，需要挖掘用户进入网站以后扮演的角色，也就是用户进入网站访问的最终目的。最终的频繁项目的访问路径也可以通过角色树推断出来。例如在A→B→C→D(化工)这个访问过程中，用户的主要目的是访问D，即担任“化工”，这个角色。但是如果挖掘出来的频繁大项目集中发现用户还担任了访问“有色金属”这个角色，可将这些频繁角色都记入表

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$DB_{User-Role}$ 中，如表 2 所示。

表 2　表 $DB_{User-Role}$ 结构

列名	属性	允许为空	含义
Bid	int	否	唯一标识符，自动增长
Uid	int	否	用户 ID
actors	varchar(200)	否	角色集

（3）角色之间关联的频繁模式集合。角色之间的关联挖掘，主要是利用对事务的日志文件进行挖掘，输出访问网站的角色之间的关联规则。通过发现用户在网站访问的行为，以发现页面之间的相关性。挖掘到的结果存入到表 $DB_{Role-Role}$ 中。例如，通过挖掘发现“煤炭”和“化工”是 2 个频繁角色。当用户访问煤炭相关页面时，可以将化工相关页面推荐给他。

1.2.2　用户聚类

该部分主要作用是将用户聚类，然后在用户聚类的基础上得到同类用户共同感兴趣的页面集。

为了对用户聚类，要考察用户对网页的兴趣度。本文认为，用户对网页的兴趣度主要由以下几方面构成：点击网页的次数、网页的浏览时间、雅可系数与最长公共路径系数。

首先，构建用户浏览兴趣多重模糊矩阵：

$$\boldsymbol{M}=\begin{pmatrix}(h_{11},t_{11}) & (h_{12},t_{12}) & \cdots & (h_{1m},t_{1m})\\ (h_{21},t_{21}) & (h_{22},t_{22}) & \cdots & (h_{2m},t_{2m})\\ \vdots & \vdots & \vdots & \vdots\\ (h_{n1},t_{n1}) & (h_{n2},t_{n2}) & \cdots & (h_{nm},t_{nm})\end{pmatrix} \tag{2}$$

式中：h_{il} 表示用户 u_i 点击页面 P_l $(1\leqslant l\leqslant m)$ 的频率；t_{il} 表示用户 u_i 访问页面 P_l 的时间比率。

计算 h_{il}、t_{il} 的方法如下：

$$h_{il}=\frac{N(u_i,P_l)}{\sum_{l=1}^{m}N(u_i,P_l)} \tag{3}$$

$$t_{il}=\frac{T(u_i,P_l)}{\sum_{l=1}^{m}T(u_i,P_l)} \tag{4}$$

式中：$N(u_i,P_l)$ 表示用户 u_i 点击页面 P_l 的次数；$\sum_{l=1}^{m}N(u_i,P_l)$ 表示日志记录中用户 u_i 的总点击数；$T(u_i,P_l)$ 表示用户 u_i 访问页面 P_l 的时间；$\sum_{l=1}^{m}T(u_i,P_l)$ 表示日志记录中用户 u_i 访问页面 P_l 的总时间。

采用最大最小法计算出基于次数的相似度系数 r_{ij}^{h} 和基于时间的相似度系数 r_{ij}^{t}。

然后，再考虑雅可系数与最长公共路径系数。

设 m 个用户访问路径集合 $U=\{C_1,C_2,\cdots,C_m\}$，其中每一条访问路径 $C_i=\{P_1,P_2,\cdots,P_n\}$，$P_i$ $(1\leqslant i\leqslant n)$ 代表一个被访问过的页面。

定义：$|C|$ 表示路径的长度，即会话路径中的页面个数。

雅可比系数不考虑事务数据先后次序关系，一般用于描述事务数据的相似性，其公式[6]：

$$S'_{ij}=\left|\frac{C_i\cap C_j}{C_i\cup C_j}\right| \tag{5}$$

式中：$|C_i\cap C_j|$ 表示 2 条会话路径中相同页面的个数；$|C_i\cup C_j|$ 表示 2 条会话路径中页面的总个数。

最长公共路径系数 CM 是考虑页面顺序的，其公式：

$$S''_{ij}=\frac{|comm(C_i,C_j)|}{\max(|C_i|,|C_j|)} \tag{6}$$

式中：$|comm(C_i,C_j)|$ 表示最长的公共路径长度；$\max(|C_i|,|C_j|)$ 表示会话的最长路径长度。

定义 2 个不同用户 u_i 和 u_j 的兴趣相似度度量公式：

$$S_{ij}=\alpha S'_{ij}+\beta S''_{ij}+\theta r_{ij}^{h}+\omega r_{ij}^{t} \tag{7}$$

式中：α、β、θ、ω 为调节系数，可根据实际情况，实验得到。

得到用户兴趣相似度后，使用 Single Link 算法[7]对用户事务模式聚类。聚类算法步骤：

步骤 1：先将所有数据对象都各自算作一类；

步骤 2：选择相似度最大的 2 个类并将其合并；

步骤 3：按照 2 个类之间的相似度为 2 个类中元素之间相似度最大者的方法，计算合并后的类－当前类的相似度；

步骤 4：循环执行步骤 2 和步骤 3，直到类的个数为 1 或满足一个阈值 η。

用户聚类后，分析各类用户访问次数最多的 5 个页面，从而得到个性化推荐页面集。

1.3　在线推荐

当用户访问网站时，系统首先判断该用户是新用户还是老用户。如果是新用户则在用户访问网站中一个页面的时候将根据离线分析模块中利用关联分析得到的相关页面集推荐给用户。如果是老用户，先判断他属于哪一类用户，然后将个性化推荐页面集中，将该类用户感兴趣的页面推荐给他。本文利用 C 语言和 SQL Server 数据库实现了一个个性

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

化推荐服务模型，列出了每类用户的推荐页面集，效果如图 3 所示。

序号	类别	URL	网页标题
1	1	http://www/hnccgc.com/xwzx/get_news.php?id=7278	安化建全国首个煤基聚酯材料园
2	1	http://www/hnccgc.com/xwzx/get_news.php?id=6675	2011年世界500强企业名单发布，河南煤化集团位列第446位
3	1	http://www/hnccgc.com/xwzx/get_news.php?id=7281	国家能源局领导到众维煤业调研，盛赞企业安全高效发展业绩
4	1	http://www/hnccgc.com/xwzx/get_news.php?id=7280	加深交流，优化服务——省地税局领导到鹤煤调研
5	1	http://www/hnccgc.com/xwzx/get_news.php?id=7240	山西省省长王君到新洛轴考察
6	2	http://www/hnccgc.com/cybk/cybk_sub01.html	煤炭
7	2	http://www/hnccgc.com/cybk/cybk_sub04.html	装备制造
8	2	http://www/hnccgc.com/cybk/cybk_sub07.html	现代服务业
9	2	http://www/hnccgc.com/cybk/cybk_sub05.html	物流贸易
10	2	http://www/hnccgc.com/cybk/cybk_sub02.html	化工
11	3	http://www/hnccgc.com/cpyfw/cpyfw_01.html	煤炭产品
12	3	http://www/hnccgc.com/cpyfw/cpyfw_02.html	化工产品
13	3	http://www/hnccgc.com/xwzx/get_news.php?id=7276	安阳龙宇新材料产业基地暨一期PET项目奠基
14	3	http://www/hnccgc.com/xwzx/get_news.php?id=7238	集团公司强调：积极应对市场变化，加快调整产品结构
15	3	http://www/hnccgc.com/xwzx/get_news.php?id=7220	集团公司21项成果获管理创新成果奖
16	4	http://www/hnccgc.com/kjcx/get_news.php?id=7272	“新型长链硅烷偶联剂”项目顺利通过中期检查
17	4	http://www/hnccgc.com/kjcx/get_news.php?id=7156	集团研究院一项目达国际先进水平
18	4	http://www/hnccgc.com/kjcx/get_news.php?id=7117	MH2宇航级碳纤维试验用丝送货，集团公司在高端新材料领域再上新台阶
19	4	http://www/hnccgc.com/kjcx/get_news.php?id=7020	研究院一示范项目通过结题验收，煤矿防突再添技术保障
20	4	http://www/hnccgc.com/kjcx/get_news.php?id=7074	集团公司下发执行科技创新市场考核验收办法
21	5	http://www/hnccgc.com/zyzx/get_news.php?id=7073	超固一矿一线优秀技工可享受副科级待遇
22	5	http://www/hnccgc.com/zyzx/get_news.php?id=6146	纵横煤海的铮铮铁汉——记集团公司劳动模范、鹤煤二矿综采队队长李秀川
23	5	http://www/hnccgc.com/zyzx/get_news.php?id=5563	鹤煤八矿班组长工作有“奔头”
24	5	http://www/hnccgc.com/zyzx/get_news.php?id=6023	“采煤状元”的昨天和今天
25	5	http://www/hnccgc.com/zyzx/get_news.php?id=6009	千米岩巷“智囊团”——记成交煤矿开拓一队QC小组

图 3 个性化页面推荐模型效果

2 结语

提出了一个应用于煤矿企业的基于 Web 日志的个性化推荐服务模型，模型应用关联规则对新用户进行页面推荐，应用聚类算法对老用户进行页面推荐；为了准确反应用户兴趣度，提出了结合点击网页的次数、网页的浏览时间、雅可系数与最长公共路径系数来度量用户兴趣度的方法。最后利用河南煤业化工集团的服务器日志数据为实验数据进行了实验。实验结果表明，该模型能够有效对网页进行分类并能够根据客户信息及时进行个性化推荐。

在已完成的工作上，笔者认为还有以下工作需要完善：

（1）用户识别的方法需要改进。本文采用的用户识别方法虽然操作简单，但是准确度较差。用户识别一直是一个研究难点，从目前的研究文献来看，对于非注册网站，还没有有效且不侵犯用户隐私的解决方法。

（2）研究评价个性化推荐效果的指标。本文给出的个性化服务系统原型的推荐效果只能依靠问卷调查的方法获取，不能通过分析日志记录等方式自动获取。用户反馈不及时，而且给用户带来了额外的麻烦。

参考文献：

[1] 于亚秀. 基于 web 日志挖掘的个性化服务研究[D]. 上海：华东师范大学，2009.

[2] MUNK M，JOZEF K，ŠVEC P. Data Preprocessing Evaluation for Web Log Mining：Reconstruction of Activities of a Web Visitor[J]. Procedia Computer Science，2010(1)：2273-2280.

[3] 何臻. 基于 Web 日志的个性化推荐服务研究[D]. 徐州：中国矿业大学，2011.

[4] LIU Haibin，VLADO K. Combined Mining of Web Server Logs and Web Contents for Classifying User Navigation Patterns and Predicting Users′ Future Requests[J]. Data & Knowledge Engineering，2007，61：304-330.

[5] 琚生根，陈黎，骆学春，等. 基于角色的个性化 WEB 日志挖掘模型研究[J]. 四川大学学报：自然科学版，2008，45(2)：271-275.

[6] 赵晶晶. Web 日志挖掘在网站个性化服务中的应用研究[D]. 大连：大连海事大学，2009.

[7] 石剑飞，闫怀志，牛占云. 基于凝聚的层次聚类算法的改进[J]. 北京理工大学学报，2008，48(1)：66-69.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net