计算机科学2002Vol. 29№. 2

基于代理服务器的协作浏览

Collaborative Browsing Based on WWW Proxy Server

王 实[1] 高 文[1] 杜建平[2] 李锦涛[1]

(中国科学院计算技术研究所 北京100080)[1](郑州大学工学院 郑州450052)[2]

Abstract When a user accesses Internet through WWW Proxy Server, he has some kinds of interest. The Proxy Server will record his basic access information in Log. Through mining the Log, we can get the interest and evaluation of the user to the Web site visited by him. His interest and evaluation to a Web site can be represented through his access time and frequency to the Web site. If a user has some kinds of interest to some Web sites, the other Web sites that are accessed by some other users having the same interest can be recommended to him. The content of the Web sites dosn't be considered. This paper presents an approach to mine the Proxy Log, provides the evaluation about a user to a Web site, and emploies the neighborhood-based collaborative filtering approach to provide the recommendation.

Keywords Web usage mining, Collaborative filtering

1 引言

当一个工作组的用户通过 WWW 代理服务器(Proxy Server)访问 Internet 时,在 Proxy Server 的日志内会留下他们的访问记录。其基本访问方式如图1所示。

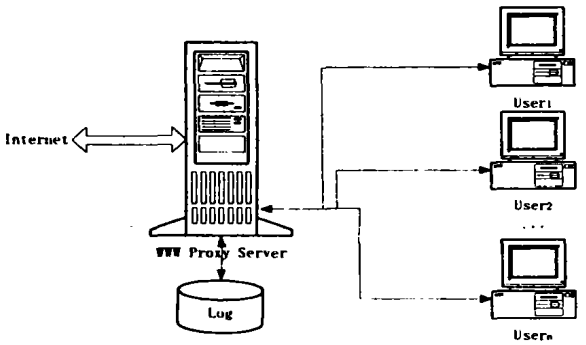

图1 一个工作组通过 WWW 代理服务器访问 Internet

当用户访问一个 Web 站点时,实际上他是带有某种兴趣来进行浏览的,因为用户之间具有不同兴趣的浏览者,他们会访问不同的 Web 站点。代理服务器会在日志中记录下他的基本访问情况。通过对其进行挖掘,我们可以得到用户对其所访问的站点的兴趣和评价。这种兴趣和评价通过其对该站点的访问时长和频度而表现出来。这样通过对 Log 的挖掘,我们可以得到工作组中每个用户对他所访问的站点的评价。这时我们就可以应用协作筛方法[1~3](Collaborative Filtering)对用户进行推荐,即如果一个用户对一些站点感兴趣,那么具有相同兴趣的另一些用户所访问的其他站点可以作为提供给他的一种推荐,这种推荐不涉及站点的内容。

本文所用到的方法属于 Web 挖掘里的 Web 访问信息挖掘(Web Usage Mining)范畴[4]。Web 访问信息挖掘是对用户访问 Web 时在服务器方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。挖掘的对象是在服务器上的包括 Server Log Data 等日志。挖掘的手段是:1)路径分析;2)关联规则和序列模式的发现;3)聚类和分类。这种挖掘方法可以从 Web 服务器那里自动发现用户存取 Web 页面的模式。得出群体用户或单个用户的访问模式和兴趣。

本文所述方法拓展了 Web 访问信息挖掘的范畴,将对访问信息的挖掘与协作筛方法结合起来用于对用户的推荐。

文[4]给出 Web 挖掘的定义,并且给出一个关于 Web 访问信息挖掘的系统 WEBMINER。文中提到的思路是通过对 Web 站点的日志进行处理,将数据组织成传统的数据挖掘方法能够处理的事务数据形式,然后利用传统的数据挖掘方法进行处理。本文的方法对代理服务器的日志进行与之不同的初步预处理,即预处理过程的目的是发现用户对一个 Web 站点的评价。然后通过基于近邻的协作筛技术进行推荐。

协作筛方法作为一种得到广泛研究和应用的方法正在许多电子商务站点得到应用。例如 Amazon.com,CDNow.com 等,其通过共享用户之间的判断而进行推荐。本文中的应用协作筛方法基于近邻的方法进行预测和推荐。

文[5]提出一种新的通过挖掘 Log,以向机器学习方法提供分类属性,而实现自适应的个性化的 Web 浏览的方法。本文所定义的用户评价与其所定义的用户分类有相似之处,但本文主要将用户评价应用于协作筛方法中,用于自动提供用户的评价而不需要人工注释,用于实现基于 Proxy Server 的协作浏览。

文中给出本文用到的基于近邻的协作筛方法。讨论了如何挖掘 Proxy 日志以自动提供用户评价的方法。给出实验比较过程。最后给出结论以及将来的工作。

2 协作筛方法

自动协作筛方法正在迅速变成一种重要的技术用于减少信息的过载。它收集用户在一个给定领域对不同项的判断,匹配出那些具有共同兴趣和口味的用户,这些用户共享他们彼此之间的分析和判断。协作筛方法能够提供有用的个性化推荐。

在协作筛方法中得到最广泛应用的方法是基于近邻的方法[1,2]。在这种方法中,与当前用户具有相似评价的其他一些用户被作为当前用户的近邻,然后近邻的评价被联合起来以用于对当前用户提供推荐。其他的一些方法包括贝叶斯网络

方法[3]、神经网络分类器方法[6]、导出规则学习方法[7]。一个基本的基于近邻的协作筛方法的目的如表1所示。

表1 对站点的推荐

	IP_1	IP_2	…	IP_m
$User_1$	5	2	…	4
$User_2$	2		…	3
…	2	2	…	2
$User_n$	5	1	…	?

已知 $User_1$ 对 IP_1 的评价为5分，对 IP_2 的评价为2分，…，对 IP_m 的评价为4分；$User_2$ 对 IP_1 的评价为2分，对 IP_2 没有评价，…，对 IP_m 的评价为3分；…；那么当新用户 $User_n$ 对 IP_1 的评价为5分，对 IP_2 的评价为1分，…时，其对 IP_m 的评价为多少？

在基于近邻的协作筛方法中，其算法主要分为如下三步：①计算当前用户和其他所有用户中每一个用户的相似性；②挑选一些具有最大相似性的用户作为近邻以用于预测或推荐；③根据这些近邻进行预测或推荐。

2.1 计算相似性

已知关系表 $r_{n\times m}$，它的每一个列为一个 IP 地址(共 m 个 IP 地址)，它的行为一个用户对这些 IP 地址的评价(共有 n 个用户的评价)，那么用户 a 和用户 u 的相似性为：

1)采用 Pearson 相关性系数来计算用户 a 和用户 u 的相似性：

$$W_{a,u}=\frac{\sum_{i=1}^{m}(r_{a,i}-\bar{r}_a)\times(r_{u,i}-\bar{r}_u)}{\sigma_a\times\sigma_u} \tag{1}$$

Pearson 相关性系数从一个线性回归模型导出，依赖一系列假定，即：关系必须是线性的，误差必须是独立的并且有一个具有均值为0的概率分布等。当这些假定不满足时，其效果不好。

2)采用 Spearman 阶相关性系数来计算用户 a 和用户 u 的相似性：

$$W_{a,u}=\frac{\sum_{i=1}^{m}(rank_{a,i}-\overline{rank}_a)\times(rank_{u,i}-\overline{rank}_u)}{\sigma_a\times\sigma_u} \tag{2}$$

Spearman 阶相关性系数与 Pearson 很相似，但它不依赖模型的假定。它计算阶之间的相关性而不直接计算值之间的相关性。

当评价尺度为一些离散的阶时(例如：整数1-5，1-7，或1-20)采用 Spearman 阶相关性系数来计算相似性。如果评价尺度是连续的，那么考虑用 Pearson 相关性系数。

2.2 选择近邻

当计算完一个用户和其它所有用户的相似性之后，要选择一个用户子集作为近邻以用于推荐，用户子集的规模要从精度和计算时间两方面考虑。

如果系统要选择近邻，那么有两种选择近邻的方法：

1)*阈值法*[1] 设置一个阈值限制，选择那些与当前用户的相似性大于该阈值的用户作为当前用户的近邻。设置一个高的阈值将得到好的相似性，但对许多用户来说，高的相似性很难得到，结果将导致很少的近邻，这样预测就不能覆盖到很多项。

2)*最好的 k 个近邻法*[2] 从大到小选择 k 个与当前用户的相似性最好的邻居作为近邻。这种方式能够保证预测的覆盖程度，但如果 k 值选择过大，那么对那些具有很高相关性的用户将引入过多的噪音。如果 k 值选择过小，那么对那些不具有很高相关性的用户将带来不好的预测。

2.3 进行预测

一旦选择好近邻，那么就要从这些邻居的评价中联合计算出一个预测。有两种较好的预测方法(设有 k 个近邻)：

1)Deviation-from-mean[8] 通过从近邻的均值中计算偏差的加权平均：

$$p_{a,i}=\bar{r}_a+\frac{\sum_{u=1}^{k}(r_{u,i}-\bar{r}_u)\times w_{a,u}}{\sum_{u=1}^{k}w_{a,u}} \tag{3}$$

2)Z-scores[2] 是 Deviation-from-mean 的一种扩充，其考虑到在用户的评价分布之间不同的散布，把用户的评价转换到 Z-scores，然后计算 Z-scores 的加权平均：

$$p_{a,i}=\bar{r}_a+\sigma_a\times\frac{\sum_{u=1}^{k}\frac{(r_{u,i}-\bar{r}_u)}{\sigma_u}\times w_{a,u}}{\sum_{u=1}^{k}w_{a,u}} \tag{4}$$

本文中我们应用这两种方法进行预测。一个预测引擎根据上述算法进行预测，它根据当前用户在 Log 中的访问情况，得到其所访问的 IP 站点的评价表，然后预测引擎根据上述算法推荐给该用户一些最好的站点。

3 自动提供用户评价

协作筛方法的一个主要问题是需要人为地提供评价。在基于 Proxy 的环境下，这种评价可以通过 Web Usage Mining 自动获得。其思想如下：如果用户频繁地长时间地访问一个站点，那么一定说明他对该站点感兴趣。访问的频率越高，访问的时间越新、越长，那么他对该站点的兴趣越大，也就说明他对该站点的评价越高。

Proxy 的日志中每一项的格式遵循 W3C 标准[9]：

表2 日志格式

Field	Description
Date	Date, time, and timezone of request
Client IP	Remote host IP and / or DNS entry
User name	Remote log name of the user
Bytes	Bytes transferred (sent and received)
Server	Server name, IP address and port
Request	URI query and stem
Status	http status code returned to the client
Service name	Requested service name
Time taken	Time taken for transaction to complete
Protocol version	Version of used transfer protocol
User agent	Service provider
Cookie	Cookie ID
Referrer	Previous page
…	…

例如：zjie [13/Aug/1999:11:25:10+0800] 159.226.42.118:-204.71.200.74:www.yahoo.com http://www.yahoo.com/ 200 10093 6.90 548 0.00

定义1 一个用户访问一个 IP 地址的次数 Count(u,IP)：在一个固定的时间段 T 内，为了去掉那些用户在网上冲浪时偶然访问的站点，要求 Count(u,IP)≥2。

定义2 一个用户访问一个 IP 地址的时长 TimeLength(u,IP)：

$$TimeLength(u,IP)=\frac{TotalTimeLength(u,IP)/Size(IP)}{\max_{IP\in\{visited\ IP\ by\ u\}}(TotalTimeLength(u,IP)/Size(IP))} \tag{5}$$

定义3 一个用户对一个 IP 地址的访问新度 new(u,IP)：

$$new(u,IP)=\frac{\sum_{i=1}^{Count(u,IP)}(Time_i(u,IP)-Time(StartLog))}{Count(u,IP)}/T \quad (6)$$

其中 $Time_i(u,IP)$表示用户 u 在一个固定的时间段 T 内第 i 次访问 IP 的时刻，$Time(StartLog)$为开始记录 Log 的时刻。

定义4 一个用户对一个 IP 地址的评价 $Evaluation(u,IP)$：

$$Evaluation(u,IP)=(\log_2 Count(u,IP))\times(1+TimeLength(u,IP)+new(u,IP)) \quad (7)$$

对 Log 进行处理，以找到在一个固定的时间段 T 内每一个用户的评价。寻找用户评价的算法为：

1)对日志进行预处理。

2)根据每一个用户的内部 IP 或用户名，划分日志。即在 Log 中找到每一个访问者的访问记录集。

3)在每一访问者的访问记录集中根据他访问的 IP 地址进行划分，形成该用户对其中每一个 IP 地址的访问子集，要求该 IP 地址至少被访问两次，不足两次的 IP 地址被认为没有被该用户访问。根据(7)式计算该用户对该 IP 地址的评价值。

4)这样，最终我们可以得到每一个用户对其访问的每一个 IP 的评价值。

通过挖掘 Log，在找到每一个用户对其访问的每一个 IP 的评价值后，我们就可以形成用户评价表。用户评价表是一个用户对 IP 地址的矩阵，其中每一个元素为用户对一个特定 IP 地址的评价，那么预测或推荐的目的就是预测特定的空元素的值。因为一个用户只会评价很少百分比的 IP 地址，那么这个矩阵是非常稀疏的。表1是一个例子。

在这个矩阵中，如果用户没有访问一个 IP，或访问的次数不到两次，那么该用户对该 IP 的评价为空。形成这样一张二维表后，就可以针对其应用协作筛方法进行推荐。

4 实验比较

实验的评价标准为：

1)*覆盖度* 其定义为协作筛预测系统所能提供预测的 IP 个数与所有 IP 地址的个数的比值。

2)*精度* 通过比较预测值和用户已经做出的评价值来测量协作筛预测系统的精度。常用的方法为平均绝对差[1]方法：对一个用户 u 来说，如果$\{r_{u,1},\cdots,r_{u,m}\}$是他给出的真实评价值，$\{p_{u,1},\cdots,p_{u,m}\}$是推荐系统给出的预测值，那么平均绝对差为：

$$|E|=\frac{\sum_{i=1}^{m}|p_{u,i}-r_{u,i}|}{m} \quad (8)$$

我们选取了一个具有223名用户的 WWW Proxy Server，取得其6个月的 Log。Log 为179.2兆，共有142,0529项，经过预处理其中包含的站点个数为9511个。

选择计算相似性方法实验，其采用最好的10个近邻，Deviation-from-mean 预测方法：

表3 选择计算相似性方法实验

计算相似性方法	精度
Spearman	0.78832
Pearson	0.78965

选择近邻实验，其采用 Pearson 相关性，Deviation-from-mean 预测方法：

表4 选择近邻实验

选择近邻标准	覆盖度	精度
全体用户	90.4	0.82306
阈值为0.1	89.1	0.82105
阈值为0.2	80.7	0.84743
阈值为0.3	60.2	0.81207
阈值为0.4	39.1	0.83115
阈值为0.5	19.4	0.81943
最好的10个近邻	90.2	0.78965
最好的20个近邻	90.27	0.79318

选择预测方法实验，其采用 Pearson 相关性，最好的10个近邻：

表5 选择预测方法实验

预测方法	精度
Deviation-from-mean	0.78965
Z-scores	0.78437

上述实验说明：针对我们的问题，在计算相似性方法中 Spearman 方法和 Pearson 方法的差异很小，可以替换使用；在近邻选择方法中，最好的10个近邻是最佳结果。在预测方法中 Z-scores 方法有最好的结果，但与 Deviation-from-mean 的差别并不大。这些实验说明根据 Log 进行挖掘而得到用户的评价，随后进行协作筛推荐的方法是可行的。

结论以及将来的工作 本文中我们提出一种新的方法，旨在挖掘 WWW 代理服务器的 Log，以自动得到协作筛方法所需的用户评价，随后利用协作筛方法进行推荐。其本质上是 Web 访问信息挖掘(Web Usage Mining)中的一种推荐方法。在这种方法中，首先我们根据用户的访问历史记录定义用户的评价，该评价的定义是根据用户访问时间和频度而进行综合考虑的结果，反映的是用户对一个 Web 站点的兴趣。在得到用户对 Web 站点的评价后，我们进一步采用基于近邻的协作筛方法进行预测，以提供给用户个性化的推荐。实验说明这种方法是可行的。

我们方法的特点是：1)从 Log 中挖掘是用户的访问兴趣以形成用户对 Web 站点的评价；2)周期性、离线地进行挖掘；3)挖掘的对象是特定用户的特性，预测和推荐的结果针对特定的用户；4)不需要打扰用户以输入自己的评价，方法自动进行评价。

我们进一步的工作要将这种方法和用户访问的页面内容结合起来，以更好地对用户进行推荐。

参考文献

1 Shardanand U, Maes P. Social information filtering: Algorithms for automating "word of mouth". In: Proc. of ACM CHI'95 Conf. on Human Factors in Computing Systems, 1995. 210～217

2 Herlocker J L, et al. An Algorithmic Framework for Performing Collaborative Filtering. SIGIR, 1999. 230～237

3 Breese J, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In UAI, 1998

4 Cooley R, et al. Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems, 1999, 1(1)

5 Chan P K. A non-invasive learning approach to building web user profiles. In: Proc. WEBKDD99, 1999

6 Billsus D, Pazzani M J. Learning collaborative information filters. In: Proc. of the 1998 Workshop on Recommender Systems. AAAI Press, August 1998

7 Basu C, et al. Recommendation as classification: using social and content-based information in recommendation. Same to[6]

8 Resnick P, et al. GroupLens: An open architecture for collaborative filtering of netnews. In: Proc. of ACM CSCW'94 Conf. on Computer Supported Cooperative Work, 1994. 175～186

9 Luotonen A. The common log file format. http://www.w3.org/pub/WWW/, 1995