基于 Web-Log Mining 的 *N* 元预测模型*

苏 中[1,2], 马少平[1,2], 杨 强[3], 张宏江[4]

[1](清华大学 计算机科学与技术系,北京 100084);
[2](清华大学 智能技术与系统国家重点实验室,北京 100084);
[3](Simon Fraser 大学,加拿大);
[4](微软中国研究院,北京 100080)
E-mail: suzhong_bj@hotmail.com
http://www.tsinghua.edu.cn

摘要: 随着 Web 上用户访问信息的不断增加,特别是 Web 服务器可提供大量的日志文件,使得有可能对这些大数据集进行知识挖掘,例如,对用户未来的访问进行预测.提出了一种利用服务器日志文件,运用 *N* 元(*N*-gram)预测模型对用户未来可能进行的 Web 访问请求进行预测.这种模型会选择性地对用户可预测的请求进行预测,从而大大提高了预测精度.实验证明,在自然语言中普遍适用的 *N* 元预测模型同样适用于网页预测.同时,采用了一种有效的简化手段,大大压缩了模型的大小,使得 5 元模型和传统的 2 元模型大小基本相同,而预测精度提高了 1 倍.该结果可以广泛地运用到 Web 上,包括网页的预发送、预取、推荐以及 Web 上的 caching 机制.试验是建立在真实的 Web 日志上的,该算法无论在预测精度上还是在可适用度上都优于以往的算法.

关 键 词: Web mining;数据挖掘;预测
中图法分类号: TP393 文献标识码: A

Internet 是一个全球的、分布的、动态的信息仓库,它存储着大量的数字化信息.在今天,它已经成为大众获得日常信息的重要来源.可是,由于庞大的信息量,对于每一个用户来说,如何能够及时地发现有用的信息则变得越来越困难.一种有效的解决方案是通过预测用户未来的网页请求来对该用户进行预发送、预取或者给该用户推荐他有可能感兴趣的网页.由于 Web 服务器日志文件中记录了该服务器被外部访问的所有过程信息,通过对这些过程信息的分析,可以客观地反映服务器的内部结构、组成、内容、访问频度等有关该服务器的重要信息.同时,在任何一个服务器上都可以很方便地得到它的日志文件,数据的来源很方便,所以对它进行分析是可行的,而且也是有效的.

受自然语言中 *n* 元语言模型的启发[1],本文提出了一种 *n* 元的概率预测模型.试验表明,这种模型同样适用于 Web 预测,而且模型的元数越高,其预测精度就越高.通过对大量的真实服务器日志文件的统计,我们发现了一种有效的模型简化手段,大大降低了模型的复杂程度,而对预测精度几乎没有影响,很多时候精度还可以提高.同时,我们比较了不同元数的模型的预测精度和可适用度(将在第 2 节中定义),运用 4 元以上的混合预测模型可以得到精度和可适用度综合指标最优的预测结果.与其他已有方法相比,本文所述方法无须知道用户的喜好信息,只需要服务器的访问日志文件.而对用户来说,不需要增加任何额外的使用负担.

本文第 1 节介绍一些相关的工作,第 2 节对模型的构造算法进行描述,第 3 节讲述模型的预测算法,第 4 节

* 收稿日期: 2000-04-03; 修改日期: 2000-07-20
基金项目: 国家重点基础研究发展规划 973 资助项目(G1998030509)
作者简介: 苏中(1976－),男,上海人,博士生,主要研究领域为基于内容图像检索,模式识别,网络数据挖掘;马少平(1961－),男,河北唐山人,博士,教授,博士生导师,主要研究领域为模式识别,信息检索,网络数据挖掘;杨强(1961－),男,北京人,博士,教授,主要研究领域为机器学习,数据挖掘,知识系统;张宏江(1960－),男,黑龙江哈尔滨人,博士,研究员,主要研究领域为视频和图像内容分析与检索,计算机视觉,信息系统.

是实验描述,最后一节是本文内容的总结和对未来工作的一些探讨.

1 相关工作

由于 Web 的成长速度和相关信息的膨胀,对 Web 用户行为的预测逐渐成为研究的热点.Web 推荐系统,就是基于数据挖掘和机器学习的方法,对用户可能感兴趣的网页进行推荐.这方面已有一些工作.这些工作基本上集中在建立一个根据当前所在页面来预测用户下一个 Web 请求的方面.一些 Web 推荐系统和预发送系统就是基于这种思路而建立的.所谓 Web 推荐系统,就是根据用户的爱好来对用户可能访问的网页进行预测,并提供给用户进行选择,例如,WebWatcher[2]和 Letzia[3]系统.而 Web 预发送系统则更进一步,它直接将网页内容发送给客户端,在预测正确的前提下,将大大缩短用户的等待时间,也能够平衡网络的负载.

预测模型可以是 2 元的,也可以是多元的.2 元模型是基于当前用户所在页面对未来用户的访问请求进行预测.这种模型只利用了用户整个会话过程中很少的信息,所以其预测精度不高,但模型简单,适合于实际使用.例如,文献[4,5]提出的预测模型,在自信度大于 50%的情况下,其预测精度只有约 30%.多元模型利用了用户访问的一定长度的路径信息,因此有可能得到更高的准确度.但通常模型复杂、庞大,准确率没有很大的提高,因而实用性能不好.本文的目标是建立一种实用、有效的多元预测模型,它可以避免上述缺点,从而使预测精度大大提高.

图 1 给出的是一个日志文件的片断,看看在日志文件里有哪些信息.从中我们可以看出,日志文件由一些项组成,每一项记录了用户 IP、申请时间、申请 URL(uniform resource locator)、协议以及申请文档大小.

```
    uplherc.upl.com       -       -       [01/Aug/1995:00:08:52       -0400]       "GET
/shuttle/resources/orbiters/endeavour-logo.gif HTTP/1.0" 200 5052
    pm9.j51.com - - [01/Aug/1995:00:08:52 -0400] "GET /images/WORLD-logosmall.gif HTTP/1.0"
200 669
    139.230.35.135 - - [01/Aug/1995:00:08:52 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0"
200 786
    uplherc.upl.com       -       -       [01/Aug/1995:00:08:52       -0400]       "GET
/shuttle/resources/orbiters/endeavour-logo.html HTTP/1.0" 200 5052
    pm9.j51.com - - [01/Aug/1995:00:08:52 -0400] "GET /images/WORLD-logosmall.html HTTP/1.0"
200 669
    139.230.35.135 - - [01/Aug/1995:00:08:52 -0400] "GET /images/NASA-logosmall.html HTTP/1.0"
200 786
```

Fig.1 A sample server log from NASA
图 1 NASA 网站日志文件中的样本

2 模型构造

我们将原始的 Web 服务器日志文件 *L* 转化为一系列分用户的会话过程(session).所谓一个会话过程,就是同一个用户的 IP 在一段时间内的连续 Web 请求.为了简便起见,我们将 Web 服务器上不同的 URL 看成不同的字母.通常,服务器每天的日志文件包含大量的会话过程,每个会话过程是由一系列的 Web 申请构成的.

我们的算法是基于对 Web 申请频率的统计,建立一个 *n* 元的预测模型.我们将长度为 *k* 的子串称作一个 *k* 元项.我们以 *n*–1 元项为索引建立一张查找表,查找表中记录着在训练集合的会话过程中,出现在 *n*–1 元项后面的 *m* 个不同 Web 申请的出现次数,并以此求得其条件概率.整个算法的流程就是通过对所有会话过程的训练集合扫描一遍来得到这样一张查找表.下面给出算法描述.

算法 **1**. Algorithm PathModelConstruction (*n*: length of *n*-gram–1; *m*: predictive steps; *L*: log file)

```
Begin
Filter the log file L then extract all the sessions from the L.
Initialize the hash table T that stores the occurrence of the document in m steps after n-grams.
Initialize the hash table H that stores results of this model.
  For I=1 to Total Session number
    For J=1 to Current Session Length
      If (J>n) Then
```

P=previous *n* requests from the current position
C=Set of distinct pages that are the next *m* requests after the current position
For each item of *C Ci*, Do
T[*P*,*Ci*]++; Update *P*(*Ci*|*P*)
If *P*(*Ci*|*P*)>ε AND *P*(*Ci*|*P*)>$H_{n\_m}(P).p$
$H_{n\_m}(Ci).p$=*P*(*Ci*|*P*),$H_{n\_m}(P)$=*Ci*
End If
End Do
End If
End For
End For
Return H_{n_m}
End

下面来描述我们整个实验的步骤:

(1) 在日志文件中清除由搜索引擎的 Crawler 以及 Proxy 发出的 Web 申请,并将其余数据装入数据库.

删除日志中的图片申请,因为通常这些图片都是包含在某个页面中,对这些图片的申请是由 HTTP 协议发出的,而不是用户.

(2) 抽出该数据库中所有的对话过程.对于一个用户的申请,如果相邻两个 Web 申请的时间间隔大于某个域值 *T*,就认为它们属于不同的对话过程.通过实验观察,我们将时间域值定为两小时.

(3) 去除所有访问次数小于域值 θ 的页面访问信息.通过对大量的 Web 服务器日志文件的统计,我们有理由进行这种处理.具体内容将在第 5 节加以介绍.通过这种过滤,我们大大简化了预测模型的大小,而且不但模型适用度几乎没有变化,预测精度还得到一定程度的提高.在实验中我们将 θ 设为 5.

3 预测算法

基于第 2 节中构造的预测模型,我们可以对 Web 用户进行实时预测.假定 H_{n_m} 是我们由历史日志文件得到的预测窗口大小为 *m* 的 *n*+1 元预测模型.我们的预测算法描述如下:

算法 **2**. Algorithm *m*-step *n*-gram+ (*P*: user's current clicking sequence; *n*: minimal path length)
Begin
If Length of *P*<*m*
Then return ("No Prediction");
Else
Begin
For *I*=max (Length(*P*), max Length of prediction model) down to *n* do
If *P* is an index in hash table H_{I_m} then
Prediction=H_{I_m}[*P*];
Return (Prediction);
End If
End For
Return ("No Prediction");
End
End If
End

作为对比实验,我们还对每个独立的 *n*+1 元预测模型进行了实验,算法描述如下:

算法 **3**. Algorithm m-step n-gram (P: user's current clicking sequence)

Begin

If Length(P)>=n and sub_string(P,Length(P)–n+1,n) is an index of H_{n_m} Then

Prediction=$H_{n\_m}[P]$;

Return (Prediction);

End If

Return ("No Prediction");

End

为了对所提出的算法进行评价,我们使用了下面的评估函数.假定在日志文件中我们得到了这样一个长度超过 k 的会话过程的集合 $S(k)=\{S_1,S_2,\ldots,S_l\}$.我们以其中的一部分作为训练集合,以得到预测模型,以另一部分作为测试集合,以评价预测结果.假定 P^+是预测正确的次数,P^-是预测错误的次数.$|R|$是用户申请的总次数,我们定义以下函数来评估预测模型 H_{n_m}:

$$precision=\frac{P^+}{(P^++P^-)}, \tag{1}$$

$$applicability=\frac{P^++P^-}{|R|}. \tag{2}$$

式(1)给出的是我们定义的预测准确度,它描述了模型预测的正确率.式(2)给出的是我们定义的可适用度,它描述了模型预测能力,即模型发言次数占总申请次数的百分比.随着模型元数的增加,预测精度逐渐上升,但可适用度却逐渐下降.为了在二者之间取一个折衷,在实验中,我们还用 precision*applicability 作为综合指标对模型进行评价.

4 实验结果

首先我们介绍一下我们使用的数据文件.第 1 份数据来源于 NASA(National Aeronautics and Space Administration)的 Web 服务器.它记录了从 1995 年 8 月 1 日 00:00:00 到 8 月 31 日 23:59:59 所有的访问信息.一共有 18 688 个不同的访问 IP 对 15 429 个页面进行了 1 569 898 次 Web 访问.从中我们抽出了 171 529 次有效会话过程.另一份数据来源于 Zukerman 等人在他们的实验中使用的 Monash University.web 服务器记录的 50 天的日志文件.它总共记录了对 6 727 个不同的页面地址进行的 525 378 次用户申请,这些申请来源于 52 455 个不同用户的 IP 地址.我们从中得到了 268 125 个有效会话过程.

图 2 是我们在网站日志文件中发现的很重要的规律.横坐标轴是日志文件中记录的同一页面的访问次数.图中有两条曲线,上面一条"Page"展示的是,对于 x-轴上的某一个 X,y 轴的值就是访问页面次数小于等于 X 的所有页面占总页面数的百分比.对于同样的 X,下面一条曲线"Request"展示的是对于这些访问页面次数小于等于 X 的所有页面在日志记录的所有访问次数占总访问次数的百分比.式(3)和式(4)给出了其数学定义,其中 S 是所有访问页面的总集合,S_i 是被访问次数为 i 次的页面集合.

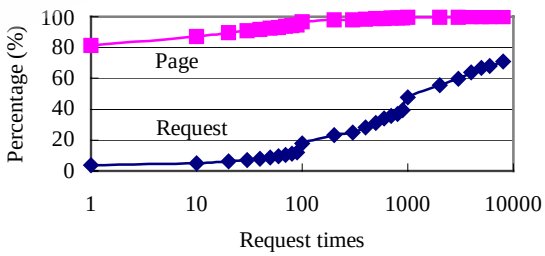

Fig.2 Page vs. request percentage

图 2 页面和访问次数的百分比

$$PageRatio(X)=\frac{\sum_{i=1}^{x}|S_i|}{|S|}, \tag{3}$$

$$RequestRatio(X)=\frac{\sum_{i=1}^{x}(|S_i|*i)}{\sum_{i=1}^{\infty}|S_i|*i}. \tag{4}$$

从图 2 中我们可以看到,大部分页面的访问次数很少,在页面的总点击次数中占很小的比例.例如,当 X=10 时,也就是被访问次数小于等于 10 次的页面,它们占总页面空间的 85%以上,而它们的总访问次数却不到 10%.由此我们可以看到,即使我们忽略了 85%左右的页面数目,也只是损失了 10%左右的访问信息.由于 N 元模型的大小和页面数目基本成正比,所以,如果我们忽略了这 85%的页面,模型大小大约只有原来的 1/6,我们也只是损失了 10%的信息.而且由于这些低访问频率的页面往往在模型中起着噪声的作用,实验结果表明,损失这 10%的信息,并没有降低我们的预测精度和可适用度.在实验中,我们将 X 定为 5,模型大小的压缩比为 30%,使其可以存放在内存中.图 3 为模型压缩时和不压缩时其预测精度的对比.从图中我们可以看到,在压缩以后,除了二元模型预测精度有所下降以外,其他多元模型的预测精度都有不同程度的提高.我们还对其可适用度做了对比,结果也基本相同.

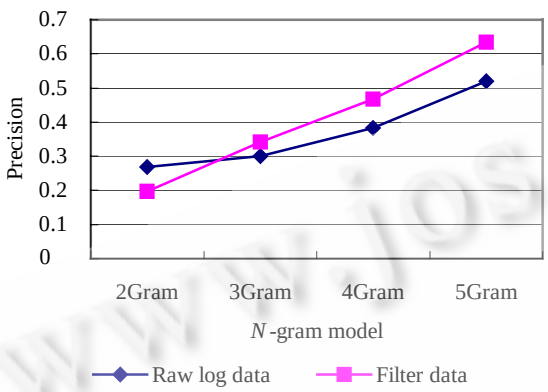

Fig.3 Comparing precision of using raw data and filtered data of different n-gram models

图 3 比较在原始数据和过滤后数据上不同 n 元模型的预测精度

对于所有的数据,我们取 4/5 为训练集,剩下的 1/5 作为测试集.我们用公式(1)和公式(2)计算每种情况下的预测精度和可适用度.图 4 给出了在预测窗口为 1 时,不同元的预测模型的预测精度和可适用度.其中 4 元$^+$是我们所使用的 m=1,4 元$^+$的预测算法.其他的则是各阶模型单独预测的结果.从这个结果我们可以清楚地看到,随着模型阶数的上升,预测精度不断上升,而可适用度则相应下降.图 5 给出的是在不同模型下,“精度*可适用度”综合指标对各模型的评价结果.可以看到,m=1,4 元$^+$比任何一种预测方法都要好.

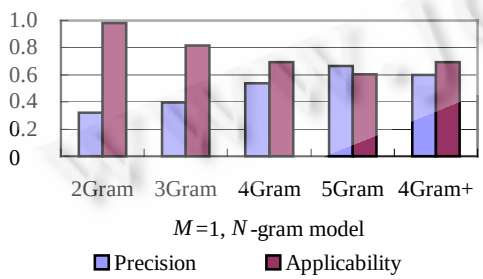

Fig.4 Precision and applicability

图 4 预测精度和可适用度

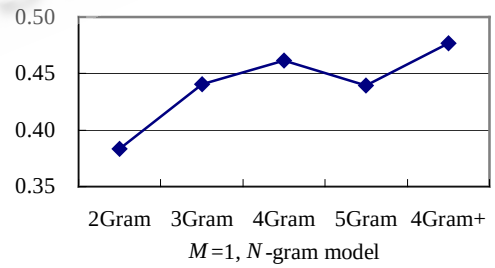

Fig.5 Precision*Applicability

图 5 预测精度*可适用度

5 小 结

本文的工作是试图将在语言模型中普遍使用而且取得成功应用的 n 元模型运用到 Web 用户的行为预测上.在大量真实的 Web 服务器日志文件上的实验表明,这种方法比其他方法的预测结果都要好.而且通过利用各

阶 *n* 元预测模型的综合生成的混合预测模型,使得综合指标得到很大的提高.本文提出的模型简化方法大大降低了模型的规模和复杂性.其结果可以广泛地运用到 Web 上,包括网页的预发送、预取、推荐以及 Web 上的 caching 机制.

致谢 感谢韩靖在初始阶段的工作.同时感谢 Susan Dumais 和 Eric Horvitz 的反馈意见和指导.我们同时感谢 David Abrecht 和 Ingrid Zukerman 给我们共享珍贵的数据文件.

References:

[1] Lee, K.F., Mahajan, S. Automatic Speech Recognition: the Development of the SPHINX System. Dordrecht, Netherlands: Kluwer, 1989

[2] Joachims, T., Freitag, D., Mitchell, T. WebWatch: a tour guild for the World Wide Web. In: Proceedings of the 15th International Joint Conference on Artificial Intelligence, IJCAI'97. 1997. 770~775.

[3] Lieberman, H. Letizia: an agent that assists Web browsing. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI'95. 1995. 924~929.

[4] Albrecht, D.W., Zukerman, I., Nicholson, A.E. Pre-Sending documents on the WWW: a comparative study. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence, IJCAI'99. 1999.

[5] Zukerman, I., Albrecht. W., Nicholson, A. Predicting user's request on the WWW. In: Proceedings of the 7th International Conference on User Modeling, UM'99. 1999.

An *N*-Gram Prediction Model Based on Web-Log Mining*

SU Zhong[1,2], MA Shao-ping[1,2], YANG Qiang[3], ZHANG Hong-jiang[4]

[1](*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*);

[2](*State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing* 100084, *China*);

[3](*Simon Fraser University, Canada*);

[4](*Microsoft Research China, Beijing* 100080, *China*)

E-mail: suzhong_bj@hotmail.com

http://www.tsinghua.edu.cn

Abstract: As an increasing number of users access information on the Web, there is a great opportunity to learn about the users' probable actions in the future from the server logs. In this paper, an *n*-gram based model is presented to utilize path profiles of users from very large data sets to predict the users' future requests. Since this is a prediction system, the recall cannot be measured in a traditional sense. Therefore, the notion of applicability is presented to give a measure of the ability to predict the next document. The new model is based on a simple extension of existing point-based models for such predictions, but the results show that by sacrificing the applicability somewhat one can gain a great deal in prediction precision. The result can potentially be applied to a wide range of applications on the Web, including pre-sending, pre-fetching, enhancement of recommendation systems as well as Web caching policies. The tests are based on three realistic Web logs. The new algorithm shows a marked improvement in precision and applicability over previous approaches.

Key words: Web mining; data mining; prediction

* Received April 3, 2000; accepted July 20, 2000
Supported by he National Grand Fundamental Research 973 Program of China under Grant No.G1998030509