基于主观 Bayes 方法对 Web 使用挖掘的研究

方贤进[1,2],李龙澍[1,2],钟 娟[3]

(1.安徽大学 教育部智能计算与信号处理重点实验室,安徽 合肥 230039;
2.安徽大学 计算机科学与技术学院,安徽 合肥 230039;
3.安徽理工大学 计算机科学与技术系,安徽 淮南 232001)

摘 要:为了更加合理地组织 Web 服务器的结构,使用户能及时快速地浏览到自己所需的网页信息,借鉴专家系统的不确定性推理方法——主观 Bayes 方法 ,提出了网页链接的可信度思想,并给出了网页链接的可信度因子模型。该模型可以定期、定时地根据 Internet 用户浏览的 Web 日志记录,动态地改善 Web 服务器的结构,从而实现基于用户浏览兴趣的网页链接结构的改进。

关键词:可信度;主观 Bayes 方法;Web 挖掘;Web 使用挖掘

中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2007)06-0056-04

Research on Web Usage Mining by Subjective Bayesian Approach

FANG Xian-jin[1,2], LI Long-shu[1,2], ZHONG Juan[3]

(1. Ministry of Education Key Lab. of Intelligent Computing & Signal Processing of Anhui Univ., Hefei 230039, China;
2. Computer Science & Technology School of Anhui Univ., Hefei 230039, China;
3. Computer Science and Technology Department of Anhui Univ. of Science and Technology, Huainan 232001, China)

Abstract: In order to reasonably optimize structure of website and enable Internet users to browse their interested Web page quickly in time, according to the idea of subjective Bayesian approach used in uncertainty reasoning in expert system fields, this paper presents the thought of Web link credibility, and gives the Web link estimation factor model, which may periodically and dynamically improve the structure of the Web server according to the Web log of Internet users browsed the Web pages, implements the improvement of Web link structure based on users' browse interests.

Key words: credibility; subjective Bayesian approach; Web mining; Web usage mining

0 引 言

人们通过 Web 接触到了大量的数据和信息,但由于 Web 页面的复杂,而且是无结构的、动态的,导致人们难以迅速、方便地在 Web 上找到所需要的数据和信息。

尽管各种搜索引擎在一定程度上解决了人们对信息的需求,但远没有达到令人满意的程度。文中将传统的数据挖掘技术与专家系统中的主观 Bayes 方法相结合,进行 Web 使用挖掘。

收稿日期:2006-09-15

基金项目:安徽省自然科学基金项目(050420204)

作者简介:方贤进(1970-),男,安徽舒城人,博士研究生,研究方向为人工智能;李龙澍,教授,博士生导师,研究方向为智能软件、软件体系结构、知识工程等。

1 Web 数据挖掘及网页兴趣度的不确定性因素

1.1 Web 数据挖掘的概念

数据挖掘:是指从数据中提取模式的过程,简单理解为从数据中挖掘有用的信息。

Web 数据挖掘:即 Web 挖掘,是数据挖掘技术在 Web 环境下的应用,是集 Web 技术、数据挖掘、计算机技术、信息科学等于一体的一项技术[1]。

1.2 Web 数据挖掘分类

一般 Web 数据挖掘可分为 Web 内容挖掘、Web 结构挖掘和 Web 使用挖掘三类[2]。其中 Web 使用挖掘是从服务器端记录的用户访问日志或从用户的浏览信息中抽取感兴趣的模式[3,4],通过分析这些数据可以帮助理解用户的行为,从而改进站点的结构或为用户提供个性化的服务。

1.3 网页兴趣度的不确定性因素

由于对同一网页,在不同的时间段内访问的用户的多少会不同,即在不同的时间段对同一网页访问的用户数量带有不确定性,不同的用户对此网页的兴趣度会发生变化;同时,在不同的时间段,对同一网页的浏览频率也会不同,也就是对此网页的兴趣度会发生变化,带有不确定性;还有用户对某一网页的浏览时长在不同的时间段也可能不同,甚至此网页以后不再有参考价值,即也带有不确定性。

鉴于以上种种不确定性因素,又是影响站点结构的主要因素,笔者采用专家系统中的主观 Bayes 方法[5],对影响网页兴趣度的因素利用可信度思想分别进行不确定性的网页兴趣度的可信度度量,在此基础上再综合进行网页兴趣度可信度度量,从而使站点的结构更加合理,更具有公共性服务。

2 网页兴趣度因子模型

2.1 网页兴趣度可信度因子

定义 1　可信度:是人们在实际生活中根据自己的经验或观察对某一事件或现象为真的相信程度,用 B 表示。网页兴趣度评估机制的可信度指网页兴趣度评估机制自身可靠的可信程度,用 $B(m)$ 表示。m 指对应的网页兴趣度评估机制。如:网页浏览人数评估机制、网页浏览频率评估机制、网页浏览时长评估机制等。

定义 2　网页兴趣度可信度:指某一网页经过网页兴趣度评估机制后得到的可信度值[6],用 $SB(u)$ 表示,u 代表某个网页。如:用 $SB(u)$ 表示某网页在网页浏览人数评估机制下的可信度。

网页兴趣度规则的可信度在此采用网页兴趣度可信度因子表示,于是有:

定义 3　网页兴趣度可信度因子:指经过网页兴趣度评估后,网页可信度 t 在网页兴趣度机制 m(如网页浏览人数机制)下的主观信任度的一种修改量,反映了网页兴趣度评估专家对网页评估后增加或减少可信度的程度,用 $\mathrm{SBF}(t,m)$ 表示。它以概率的形式对网页评估的条件概率相对于网页的先验概率改变的比例表示[7],如公式(1):

$$\mathrm{SBF}(t,m)=\begin{cases}\dfrac{P(t\mid m)-P(t)}{1-P(t)}, & \text{若 } P(t\mid m)\geqslant P(t)\\[2ex] \dfrac{P(t\mid m)-P(t)}{P(t)}, & \text{若 } P(t\mid m)<P(t)\end{cases}\tag{1}$$

其中由主观Bayes概率理论知,网页可信度的先验概率 $P(t)$ 表示在不经过网页兴趣度评估机制前的网页可信的程度。经过网页兴趣度评估机制 m 后,网页的可信度用后验概率 $P(t\mid m)$ 表示。

当 $\mathrm{SBF}(t,m)>0$ 时表示经此网页兴趣度评估机制成功,即公式(2):

$$\mathrm{SBF}(t,m)=\frac{(P(\neg t)-P(\neg t\mid m))}{P(\neg t)}\tag{2}$$

所以 $\mathrm{SBF}(t,m)$ 可理解为网页不可信度$(\neg t)$ 的先验概率相对于经过网页兴趣度评估机制后$(\neg t)$ 的条件概率的相对减少值。若 $\mathrm{SBF}(t,m)=0.65$ 表示经过网页兴趣度机制 m 后,$(\neg t)$ 的可信度相对减少了 65%。可见 $\mathrm{SBF}(t,m)>0$ 意味着经过网页兴趣度机制 m 后,网页的可信度增强了。

同理,当 $\mathrm{SBF}(t,m)=0$ 则意味着 $P(t\mid m)=P(t)$,即网页经过某网页兴趣度机制 m 后未改变(即此网页兴趣度评估机制对网页不起作用)。

同理,当 $\mathrm{SBF}(t,m)<0$ 则表示网页兴趣度机制失效,即公式(3):

$$\mathrm{SBF}(t,m)=-\frac{(P(t)-P(t\mid m))}{P(t)}\tag{3}$$

此时 $\mathrm{SBF}(t,m)$ 可理解为网页可信度 t 的先验概率相对于经过网页兴趣度评估机制 m 后 t 的条件概率的相对减少值。若 $\mathrm{SBF}(t,m)=-0.65$ 表示经过网页兴趣度评估机制 m 后,网页的可信度 t 相对减少了 65%。故 $\mathrm{SBF}(t,m)<0$ 意味着经过网页兴趣度评估机制后网页的可信度减弱了。由公式(1) 可得公式(4):

$$P(t\mid m)=\begin{cases}\mathrm{SBF}(t,m)+(1-\mathrm{SBF}(t,m))\times P(t), & \text{若 } \mathrm{SBF}(t,m)\geqslant 0\\ (\mathrm{SBF}(t,m)+1)\times P(t), & \text{若 } \mathrm{SBF}(t,m)<0\end{cases}\tag{4}$$

这样的转化就可以求解网页经过某网页兴趣度评估机制 m 后的可信度值,即 $P(t\mid m)=SB(u)$。

2.2 网页兴趣度规则的加强

网页经过网页兴趣度评估机制可信度,要么增强要么减弱。为方便求解网页兴趣度评估机制增强与减弱情况下的网页的可信度,把这两种结果分别给予了定义:网页兴趣度评估机制增强可信度因子(SABF)和减弱可信度因子(SSBF)。

增强可信度因子(SABF)表示在网页经过网页兴趣度评估机制 m 后的可信度因子,取值在[0,1]之间。同时 SSBF 则表示在网页经过网页兴趣度评估机制 m 后的可信度因子减弱(用 $\neg m$ 表示),取值在[-1,0]之间。分别表示如下:

$$\mathrm{SABF}(t,m)=\frac{P(t\mid m)-P(t)}{1-P(t)}$$

$$SSBF(t,\neg m)=\frac{P(t\mid\neg m)-P(t)}{P(t)}$$

即 $SABF(t,m)$ 与 $SSBF(t,\neg m)$ 对于 $SBF(t,m)$ 具有互斥性。即要么 $SBF(t,m)=SABF(t,m)$，要么 $SBF(t,m)=SSBF(t,\neg m)$，也就是公式(5)：

$$SBF(t,m)=\begin{cases}SABF(t,m),\text{网页兴趣度可信度增强}\\SSBF(t,\neg m),\text{网页兴趣度可信度减弱}\end{cases}\tag{5}$$

2.3 扩展网页兴趣度可信度因子

前文的定义是建立在网页兴趣度可信度因子完全可信的情况，然而由于网页浏览人数、网页浏览频率、网页浏览时长等评估机制，在总的网页兴趣度评估中各占有一定的比例，故在网页兴趣度中要综合考虑兴趣度机制与网页兴趣度规则两方面的可信度。下面对网页兴趣度可信度因子进行扩展，用来表示某网页兴趣度评估机制可信度为 $B(m)$ 情况下的兴趣度可信度因子。

这里 $B(m)$ 表示对此网页兴趣度评估机制 m 的相信程度。可表示成 $B(m)=P(m\mid s)$，s 表示与此网页兴趣度评估机制 m 有关的所有观察。故得：

定理 1：在网页兴趣度机制为 $B(m)$ 情况下网页可信度 t 的扩展条件概率表示为 $P(t\mid m_{B(m)})$，用公式(6)表示如下：

$$P(t\mid m_{B(m)})=\begin{cases}P(t\mid\neg m)+\dfrac{P(t)-P(t\mid\neg m)}{P(m)}\times B(m),\\\qquad\text{若 } 0\leqslant B(m)<P(m)\\P(t)+\dfrac{P(t\mid m)-P(t)}{1-P(t)}\times(B(m)-P(m)),\\\qquad\text{若 } P(m)\leqslant B(m)\leqslant 1\end{cases}\tag{6}$$

又已知概率公式：

$$P(t\mid s)=P(t\mid m)P(t\mid s)+P(t\mid\neg m)P(\neg m\mid s)\tag{7}$$

此公式已由 Duda 等人于 1976 年作了证明。

在所有观察 s 下，$P(m\mid s)$ 与 $P(t\mid s)$ 的关系存在三种特殊情况：

① 当 $P(m\mid s)=1$ 时，$P(\neg m\mid s)=0$ 则 $P(t\mid s)=P(t\mid m)$；表示此网页兴趣度机制肯定可信。

② 当 $P(m\mid s)=0$ 时，$P(\neg m\mid s)=1$，则 $P(t\mid s)=P(t\mid\neg m)$；表此网页兴趣度机制肯定不可信。

③ 当 $P(m\mid s)=P(m)$ 时，$P(t\mid s)=P(t)$，表示观察 s 与 m 无关。

利用这三种特殊情况，分别取了三个特殊点，再利用分段线性插值公式可得 $P(m\mid s)$ 的函数 $P(t\mid s)$ 的解析表达式(参见图 1)，即公式(8)：

$$P(t\mid s)=\begin{cases}P(t\mid\neg m)+\dfrac{P(t)-P(t\mid\neg m)}{P(m)}\times P(m\mid s),\\\qquad\text{若 } 0\leqslant P(m\mid s)<P(m)\\P(t)+\dfrac{P(t\mid m)-P(t)}{1-P(m)}\times(P(m\mid s)-P(m)),\\\qquad\text{若 } P(m)\leqslant P(m\mid s)\leqslant 1\end{cases}\tag{8}$$

可见公式(6) 与公式(8) 可转化。要想利用公式(8) 须首先知道 $P(m\mid s)$ 与 $P(m)$ 之间关系。故要先求 $P(m)$ 的值，下面来求 $P(m)$。

定义 4　网页兴趣度机制对应的评估可信度因子的特征值可用三元组 $(B(m),SABF(t,m),SSBF(t,\neg m))$ 来表示。假如某网页兴趣度评估机制可信度为 $B(m)$，若网页兴趣度评估机制增强则为 $SABF(t,m)$，反之则为 $SSBF(t,\neg m)$。

这样可求出某网页兴趣度机制的先验概率。即公式(9)：

$$P(m)=\frac{P(t)\times SSBF(t,\neg m)}{P(t)\times SSBF(t,\neg m)-(1-P(t))\times SABF(t,m)}\tag{9}$$

故由公式(6) 可知若 $B(m)>P(m)$ 时说明某网页兴趣度评估机制可信，且网页兴趣度评估机制的可信度随 $B(m)$ 的增大而增大。反之若 $B(m)<P(m)$ 则某网页兴趣度机制不可信。当 $B(m)=0$ 时，某网页兴趣度评估机制完全不可信。但当 $B(m)=P(m\mid s)=P(m)$ 时不能表明网页兴趣度机制可信或不可信。

为此提出可信度因子扩展公式。用 $SBF(t,m_{B(m)})$ 表示在可信度为 $B(m)$ 情况下的网页兴趣度可信度因子。即公式(10)：

$$SBF(t,m_{B(m)})=\begin{cases}\dfrac{P(t\mid m_{B(m)})-P(t)}{1-P(t)},\text{若 } P(m_{B(m)})\geqslant P(t)\\\dfrac{P(t\mid m_{B(m)})-P(t)}{P(t)},\text{若 } P(m_{B(m)})<P(t)\end{cases}\tag{10}$$

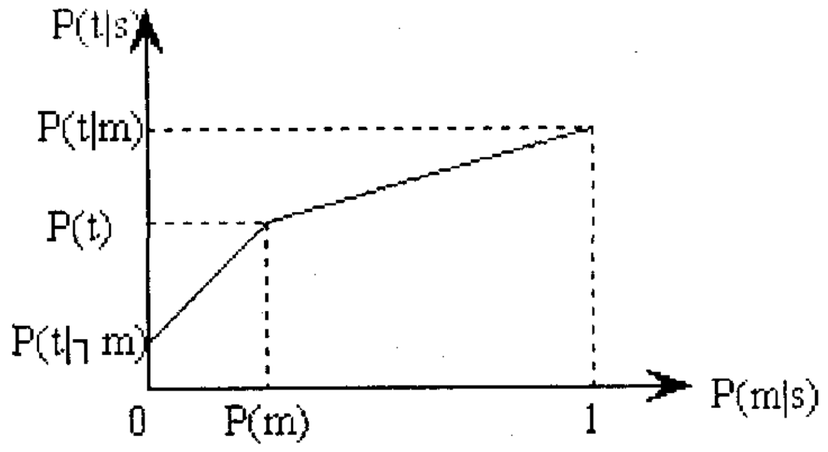

图 1　线性插值函数

3　组合网页兴趣度可信度因子

网页经过某网页兴趣度评估机制(如浏览人数、浏览频率、浏览时长这些风险机制) m_i，则相对应的可信度因子为 $SBF(t,m_i)$，(其中 $i=1,2,3$)。那么网页经

过所有网页兴趣度评估机制相应网页兴趣度可信度因子为 SBF($t, m_1m_2m_3$),称为网页兴趣度评估机制的并行组合公式。同时 t 与 $\neg t$、各网页兴趣度评估机制之间相互独立。那么可得公式(11):

$$O(t \mid m_1m_2m_3) = \prod_{i=1}^{3} \lambda(t, m_i)O(t) \qquad (11)$$

其中:

$$\lambda(t, m_i) = \frac{P(m \mid t)}{P(m \mid \neg t)}$$

即表示网页经过网页兴趣度可信度 t 关于网页兴趣度机制 m 的似然率。$O(x) = \frac{P(x)}{1 - P(x)}$ 表示几率函数。则可得 $P(x) = \frac{O(x)}{1 + O(x)}$,在此 λ 可看作一个几率修改因子,显然当 $\lambda > 1$ 时,后验几率大于先验几率,相应地网页兴趣度评估机制支持网页的可信度;当 $\lambda < 1$ 时,后验几率小于先验几率,相应地网页兴趣度评估机制减弱网页的可信度。再由公式(11) 得公式(12):

$$P(t \mid m_1m_2m_3) = \frac{O(t \mid m_1m_2m_3)}{1 + O(t \mid m_1m_2m_3)} = \frac{\prod_{i=1}^{3} P(t \mid m_{iB(m_i)})P(\neg t)}{\prod_{i=1}^{3} P(t \mid m_{iB(m_i)})P(\neg t) + \prod_{i=1}^{3} P(\neg t \mid m_{iB(m_i)})P(t)} \qquad (12)$$

则可得在 m_1, m_2, m_3 的可信度 t 的扩展条件概率表示为 $P(t \mid m_{1B(m_1)}m_{2B(m_2)}m_{3B(m_3)})$,即某网页在三网页兴趣度评估机制的可信度值为:

$$BF(u) = P(t \mid m_{1B(m_1)}m_{2B(m_2)}m_{3B(m_3)}) = \frac{\prod_{i=1}^{3} P(t \mid m_{iB(m_i)})P(\neg t)}{\prod_{i=1}^{3} P(t \mid m_{iB(m_i)})P(\neg t) + \prod_{i=1}^{3} P(\neg t \mid m_{iB(m_i)})P(t)} \qquad (13)$$

其中 $P(t \mid m_{iB(m_i)})$ 可由公式(6)分别求出,然后代入即可。

4　举　例

假设在一个网页兴趣度评估机制中,$P(t) = 0.6$,则该网页要经过三种风险评估机制。其中浏览人数评估机制三元组为(0.7,0.7,−0.1);浏览频率评估机制三元组为(0.7,0.6,−0.2);浏览时长评估机制三元组为(0.6,0.6,−0.3)。则通过网页兴趣度评估机制的可信度因子计算,由公式(9) 分别可求得在这三种评估机制下的先验概率为:

$$P(m_1) = P(m_{人数}) = 0.1765$$
$$P(m_2) = P(m_{频率}) = 0.333$$
$$P(m_3) = P(m_{时长}) = 0.429$$

又由公式(4) 求得在这三种评估机制下的可信度值分别为:

$$P(t \mid m_1) = 0.88 \quad P(t \mid \neg m_1) = 0.54$$
$$P(t \mid m_2) = 0.84 \quad P(t \mid \neg m_2) = 0.48$$
$$P(t \mid m_3) = 0.84 \quad P(t \mid \neg m_3) = 0.42$$

又根据定理 1 可分别求出它们的扩展条件概率:

$P(t \mid m_{1B(m_1)}) = 0.96645$　$P(t \mid m_{2B(m_2)}) = 0.8202$

$P(t \mid m_{3B(m_3)}) = 0.7026$

最后求出此网页在总的网页兴趣度机制的可信度为:

$BF(u) = P(t \mid m_{1B(m_1)}m_{2B(m_2)}m_{3B(m_3)}) = 0.9952$,再代入网页兴趣度机制可信度因子扩展公式得:SBF($t, m_{1B(m_1)}m_{2B(m_2)}m_{3B(m_3)}$) = 0.988。所以此网页经过网页兴趣度总评估机制可信度因子值为 0.988,可信度值为 0.9952,故说明此网页是兴趣度比较高的。

5　结　论

信息技术的飞速发展,使网络的结构及信息的有效性等诸要素也存在着不确定性。文中借鉴专家系统的主观 Bayes 方法,与传统数据挖掘技术中的使用挖掘相结合[8],根据服务器端记录的日志,提出了网页兴趣度的可信度思想,并给出了可信度模型,处理了网页兴趣度中的不确定性因素,从而合理地根据人们的需求更好地、定期地改进站点结构,使网页兴趣度在此模型下达到最高。

参考文献:

[1] 钟　珞,张开松,李三得,等. Web 使用挖掘研究及实现[J]. 微机发展,2005,15(1):33－35.

[2] 蒋外文,喻兴标,熊东平. Web 使用挖掘研究[J]. 微机发展,2005,15(8):37－40.

[3] 魏恒义,管旭东,杨怡玲,等. Web 日志挖掘中的数据预处理的研究[J]. 计算机工程,2000,26(4):66－67.

[4] 靳风荣,郑雪峰. Web 日志挖掘的预处理过程及算法[J]. 微型电脑应用,2004,20(6):44－45.

[5] 张仰森. 人工智能原理与应用技术[M]. 北京:高等教育出版社,2003.

[6] 谷秀岩. Web 使用挖掘的研究[J]. 计算机工程与应用,2005,41(16):175－177.

[7] 田盛丰. 人工智能原理及应用[M]. 北京:北京理工大学出版社,1993.

[8] 耿　桦,李　媛,朱　炜,等. Web 搜索中的数据挖掘技术研究[J]. 计算机科学,2005,32(4):37－40.