网站用户偏好度的数据挖掘模型

赵雪梅,朱恩亮
(盐城工学院 实验教学部,江苏 盐城 224051)

摘要:数据挖掘和数据库知识发现是当前国际科技界的一个研究热点。这是一个介于统计学、模式识别、人工智能、机器学习、数据库技术以及高性能并行计算等领域的交叉新兴学科,具有极为广泛的应用前景。在基于统计学观点的基础上讨论了网站用户偏好度的数据挖掘模型,设计了一个网站用户信息浏览偏好度的数据挖掘模型。

关键词:数据挖掘;统计学;模型;信息;偏好度
中图分类号:TP311.1 **文献标识码**:A **文章编号**:1671-5322(2009)01-0075-04

随着计算机技术和电子数据获取方面的不断进展以及因特网和各种局域网的广泛普及,人们获得的数据正以前所未有的速度急剧增加,最近几十年产生了很多超大型数据库,遍及超级市场销售、银行存款、天文学、粒子物理、化学、医学以及政府统计等领域。例如,美国著名零售商MalMart每天要做2 000万次交易;美国电报电话公司AT &T每天有1亿多用户在远程网络上呼叫2亿多次。在这个充满数据的数字化、信息化时代,如此规模甚至更大的数据库将是人们不得不面对的一个越来越突出的问题。如何从这些大型数据库中发现有用的信息、模式和知识?如何开发有效的挖掘方法?已成为众多科技工作者共同关注的焦点。在过去一个称为"数据挖掘"和"数据库知识发现"(Data Mining & Knowledge Discovery in Database,简称DM & KDD)的新领域得到了快速发展[1,2],这是一个介于统计学、模式识别、人工智能、机器学习、数据库技术以及高性能并行计算等领域的交叉新学科,已在经济、商业、金融、天文等行业得到了成功的应用,在国际上掀起了一股空前的研究热潮。我国学者在这个领域也已开展了很多研究[3,4],但涉及的人员主要来自计算机科学及相关领域,其它专业的研究者相对较少,究其原因可能是由于学科相隔、交流不够所致。

随着Internet的迅速发展,如前所述,各种信息以指数级的速度增长,类型也越来越多。人们面对太多的信息无法选择和消化。Interne上信息资源分布的广泛性又给用户寻找感兴趣的信息增加了困难,用户不知道如何更有效地发现自己所需的信息资源。当前我们主要采用搜索引擎来检索Web上的信息,大多数搜索引擎缺少主动性,没有考虑用户的兴趣偏好。

为了适应用户不断增长的信息需求,研究人员纷纷从人工智能中寻找突破口。在许多探索性研究中,个性化主动信息服务(Personalized Active Information Service, PAIS)作为一种崭新的智能信息服务方式,应用前景广泛,十分引人注目[5-9]。PAIS的特征是信息服务系统根据每个用户的信息需求和用户的个性化模式,主动搜寻相关信息,并且利用在线智能推荐服务或者推送技术,准确地将用户所需的信息传送到相应的用户。

在智能个性化主动信息服务中最重要的服务是个性化信息推荐。作为人工智能的一个重要研究领域,数据挖掘近年来有了广泛的应用。因此,两者的结合—基于数据挖掘的Web个性化信息推荐服务日益成为一个重要的研究课题。

目前已经存在很多个性化信息推荐系统,但是仍然存在一些问题,主要包括:

① 多数个性化信息推荐系统针对的是注册用户,较少考虑非注册用户的信息推荐;

② 多数个性化信息推荐系统对新用户和访

收稿日期:2008-12-29

作者简介:赵雪梅(1975-),女,江苏盐城人,讲师,主要研究方向为数据挖掘、计算机教学。

万方数据

问站点较少的用户的信息推荐考虑不够,因为新用户和浏览站点较少的用户被系统收集的用户信息较少,采用某些推荐算法并不合适;

③ 大多数个性化信息推荐系统没有考虑用户是否有新颖信息需求的偏好。

1 网站数据的特点

Web上有海量的数据信息,怎样对这些数据进行复杂的应用成了现今数据库技术的研究热点。数据挖掘就是从大量的数据中发现隐含的规律性的内容,解决数据的应用质量问题。充分利用有用的数据,废弃虚伪无用的数据,是数据挖掘技术的最重要的应用。相对于Web的数据而言,传统的数据库中的数据结构性很强,即其中的数据为完全结构化的数据,而Web上的数据最大特点就是半结构化[10]。所谓半结构化是相对于完全结构化的传统数据库的数据而言。显然,面向Web的数据挖掘比面向单个数据仓库的数据挖掘要复杂得多。据统计,有99%的网站信息相对99%的用户来说都是无用的。虽然这看起来并不是很明显,但一个用户只关心网站上的很小的一部分信息却是事实,网站上所包含的其余信息对于这个用户来说是不感兴趣的,而且往往会"淹没"所希望得到的信息,鉴于此,网站上的数据有如下特点。

1.1 异构数据库环境

从数据库研究的角度出发,Web网站上的信息也可以看作一个数据库,一个更大、更复杂的数据库。Web上的每一个站点就是一个数据源,每个数据源都是异构的,因而每一站点之间的信息和组织都不一样,这就构成了一个巨大的异构数据库环境。如果想要利用这些数据进行数据挖掘,首先,必须要研究站点之间异构数据的集成问题,只有将这些站点的数据都集成起来,提供给用户一个统一的视图,才有可能从巨大的数据资源中获取所需的东西。其次,还要解决Web上的数据查询问题,因为如果所需的数据不能很有效地得到,对这些数据进行分析、集成、处理就无从谈起。

1.2 半结构化的数据结构

Web上的数据与传统的数据库中的数据不同,传统的数据库都有一定的数据模型,可以根据模型来具体描述特定的数据。而Web上的数据非常复杂,没有特定的模型描述,每一站点的数据都各自独立设计,并且数据本身具有自述性和动态可变性。因而,Web上的数据具有一定的结构性,但因自述层次的存在,从而是一种非完全结构化的数据,这也被称之为半结构化数据。半结构化是Web上数据的最大特点。

1.3 解决半结构化的数据源问题

Web数据挖掘技术首要解决半结构化数据源模型和半结构化数据模型的查询与集成问题。解决Web上的异构数据的集成与查询问题,就必须要有一个模型来清晰地描述Web上的数据。针对Web上的数据半结构化的特点,寻找一个半结构化的数据模型是解决问题的关键所在。除了要定义一个半结构化数据模型外,还需要一种半结构化模型抽取技术,即自动地从现有数据中抽取半结构化模型的技术。面向Web的数据挖掘必须以半结构化模型和半结构化数据模型抽取技术为前提。

2 网站用户偏好度模型

当一个的用户浏览网站上的信息时,他会受到很多因素的影响,如信息标题,网页之间的链接,网页的外观等等。显然,这些因素之间很难用一种精确的方法表示出来。但是我们可以在合理的假设之下,建立他们之间的模型,为此我们可以假设:

用户偏好度是其影响因素的线性函数。

于是,令用户偏好度为y,影响它的因素假设有$x_1,x_2,\cdots,x_n$共n个参数,因此,我们可以得到如下的线性关系:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k+\varepsilon$$

其中,ε是随机变量,且一般假设$\varepsilon\sim D(0,\sigma^2)$。为了估计未知参数$\beta_0,\beta_1,\beta_2,\cdots,\beta_k$我们可以对$y$和$x_1,x_2,\cdots,x_k$同时做$n$次独立观察试验,得到$n$组观察样本值$(y_t,x_{t1},x_{t2},\cdots,x_{tk},t=1,2,\cdots n(n>k+1)$,他们满足关系

$$y_t=\beta_0+\beta_1x_{t1}+\beta_2x_{t2}+\cdots+\beta_kx_{tk}+\varepsilon_t,t=1,2,\cdots n(n>k+1) \quad (1)$$

其中,$\varepsilon_1,\varepsilon_2,\cdots\cdots,\varepsilon_n$互不相关且均是与$\varepsilon$同分布的随机变量。我们把上面的式子用矩阵来表示,于是令

$$Y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix},\beta=\begin{pmatrix}\beta_0\\\beta_1\\\vdots\\\beta_k\end{pmatrix},\varepsilon=\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{pmatrix}$$

万方数据

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

这样①式变为

$$Y = X\beta + \varepsilon$$

其中 X 为已知的 $n \times (k+1)$ 阶矩阵,我们称为用户偏好度的资料矩阵,β 为 $k+1$ 为的未知列向量,ε 是满足

$$\begin{cases} E(\varepsilon) = 0 \\ Cov(\varepsilon, \varepsilon) = \sigma^2 I \end{cases}$$

的 n 维随机变量。这样,我们就可以得到如下的用户偏好度模型

$$\begin{cases} Y = X\beta + \varepsilon \\ E(\varepsilon) = 0 \\ Cov(\varepsilon, \varepsilon) = \sigma^2 I \end{cases}$$

这样,有统计学知识,我们得到未知系数 β 的最小二乘估计为

$$\hat{\beta} = (X'X)^{-1}X'Y$$

3　模型的运用

我们现在以通用商品销售系统来做试验,后台的界面如图 1 所示。当人们想购买商品时,会浏览商品的信息,这样我们可以模拟如下一组数据(见表 1),其中用户偏好度用商品信息的点击量来表示,试验中,我们只考虑一个影响因素—商品信息的标题的知识量(0 到 100 之间的实数)。

为了研究信息标题的知识量与用户偏好度(信息的点击量)之间的规律,我们以信息的标题的知识量 x 为横坐标,以用户偏好度(信息的点击量)y 纵坐标,将这些数据点在笛卡尔坐标系里标出,如下图 2。

表 1　商品信息点击量与信息标题知识量的关系

Table 1　Merchandise traffic information and heading information the title of the relationshio between knowledge

点击量	143	145	146	147	149	150	153	154	156	157	158	159	160	162	164	155
知识量	88	85	88	91	92	93	93	95	98	97	96	98	99	100	100	96

图 1　后台界面

Fig. 1　Background interface

显然,信息标题的知识量 x,用户偏好度(信息的点击量)y 之间的关系大致满足线性关系,由模型我们可得 x 和 y 之间的关系为

$$\hat{y} = -15.080\,0 + 0.610\,0x$$

4　结束语

数据挖掘对统计学理论和实践提出了许多困难而根本的问题,统计学对此并非都有答案,但统

万方数据

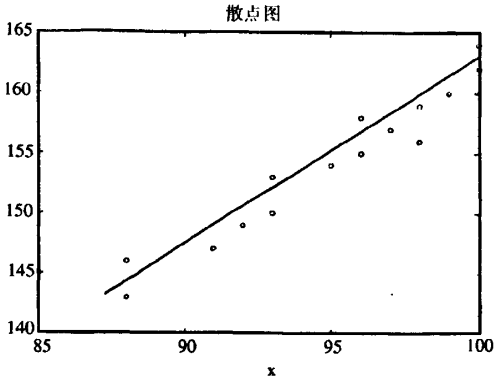

图2 信息标题的知识量对用户偏好度的关系

Fig. 2 the amount of information the title of the knowledge of the relationship between the degree of user preferences

计学为探索解决方案提供了有用而实际的框架,为数据挖掘提供了丰富的方法和结果。已经有越来越多的研究者认识到,统计学对数据挖掘是必要的,数据挖掘需理解统计学原理,统计学需理解数据挖掘问题的本质,二者的结合必将对数据挖掘技术产生深远的影响。

我们需要信息处理的技术和方法,特别是从大数据集产生有效科学结果的技术和方法。统计学已和其它学科一起为数据挖掘提供了很多极具潜力的思想与方法。数据挖掘的未来也许主要在于将这些来自不同学科的方法有效地,创造性地实施和集成。不过,任何时候,用分析的艺术和人脑的能力从数据中综合新知识仍然是任何机器无法超越的。

本文基于统计学的观点,得到了网站用户信息浏览偏好度的数据挖掘模型,为同类问题提出了一种新的思路。

参考文献:

[1] 王能斌. 数据库系统教程[M]. 北京:电子工业出版社,2002.

[2] 孙荣恒. 概率论与数理统计[M]. 重庆:重庆大学出版社,2000.

[3] 王琴,杨宗凯,吴砥. 基于工作流和 JSP/Servle 技术的网上项目申报与管理系统设计[J]. 计算机应用研究,2006(12):181 - 184.

[4] 王实,高文. 数据挖掘中的聚类方法[J]. 计算机科学,2000,27 (4) :42 - 45.

[5] Heckerman D. Bayesian networks for data mining[J]. Data Mining and Knowledge Discovery,1997,1:79 - 119.

[6] 林士敏,田凤占,陆玉昌. 贝叶斯学习、贝叶斯网络与数据采掘[J]. 计算机科学,2000,27 (10):69 - 72.

[7] Hand D J . Data mining:statistics and more[J]. The American Statistician,1998,52(2):112 - 118.

[8] Hand D J. Statistics and data mining:intersecting disciplines[J]. ACM SIGKDD Explorations Newsletter,1999,1(1):16 - 19.

[9] Hosking J R M, Pednault E P D , Sudan M. A statistical perspective on data mining[J]. Future Generation Computer Systems,1997,13:117 - 134.

[10] 格罗思,侯迪. 数据挖掘—构筑企业竞争优势[M]. 宋擒豹,译. 西安:西安交通大学出版社,2001.

The Data Mining Model for the Preference Degree of Website Users

ZHAO Xue-mei, ZHU En-liang

(Department of Experiment Teaching, Yancheng Institute of Technology, Jiangsu Yancheng 224051, China)

Abstract: Data mining and knowledge discovery in database have now been an active research area in scientists and engineers. This is a new discipline lying at the interface of statistics, pattern recognition, artificial intelligence, machine learning, database technology, high performance and parallel computing, and so forth. The purpose of this paper is to discuss the data mining model of preference degree for website users. We present a data mining model for the preference degree of information browsed of website users.

Keywords: data mining; statistics; model; information; preference degree

(责任编辑:沈建新; 校对:张英健)

万方数据