应用模糊集理论的多因素兴趣度评价

李 珊,云彩霞,白彦霞,张秋菊,李丽芬
LI Shan,YUN Cai-xia,BAI Yan-xia,ZHANG Qiu-ju,LI Li-fen

北京化工大学 北方学院 信息学院,河北 廊坊 065201
Department of Information,North College of Beijing University of Chemical Technology,Liangfang,Hebei 065201,China
E-mail:lishan_106@yahoo.com.cn

LI Shan,YUN Cai-xia,BAI Yan-xia,et al.Evaluation of many-sided interesting degree based on fuzzy sets theory.Computer Engineering and Applications,2009,45(8):132-134.

Abstract: With the rapid increase of the Internet information,users find what they need more and more difficultly.Aiming at the problem,authors propose a method of evaluating users' interest adopting fuzzy comprehensive evaluation,experiment results indicate that the method has better effect.

Key words: fuzzy sets theory;fuzzy comprehensive evaluation;Web logs;user's actions;user's interest;data mining

摘 要:互联网的信息急剧增加,用户越来越难以找到自己所需要的信息。针对目前个性化服务系统中用户兴趣难以获取的问题,提出了一种模糊综合评判评价用户兴趣的方法,实验表明具有较好的评价效果。

关键词:模糊集理论;模糊综合评判;Web 日志;用户行为;用户兴趣;数据挖掘
DOI:10.3778/j.issn.1002-8331.2009.08.040 **文章编号**:1002-8331(2009)08-0132-03 **文献标识码**:A **中图分类号**:TP311

1 引言

互联网上的信息含量让人无法估量,使得个性化服务技术成为当前研究的一个重要课题。高质量的推荐技术对于个性化服务的质量至关重要。而如何根据用户的兴趣和爱好为其定制个性化推荐内容则是个性化技术研究的关键内容之一。

1965 年,扎德教授在发表的《模糊集合论》论文中提出用"隶属函数"这个概念来描述现象差异的中间过渡,从而突破了经典集合论中属于或不属于的绝对关系[1]。扎德创立的模糊集合是模糊数学的基础,以逻辑真值为[0,1]的模糊逻辑为基础,善于描述属性不分明的事物,是对经典集合的开拓。模糊综合评判是指当评价因素具有模糊性时,综合考虑受多种因素影响的事物或系统并对其进行总的评价。

本文提出一种采用模糊综合评判计算用户兴趣度的方法,用户的兴趣应该从多种因素考虑,数据包括用户的行为信息和 Web 日志中记录的信息。实验表明能在一定程度上提高挖掘用户兴趣的准确性。

2 用户信息的收集

个性化信息服务系统根据用户初次使用系统时注册的个人信息,定制用户的描述文件。描述文件定制好之后,可以由用户自主修改,也可以由系统自适应地修改。当系统要自适应地修改描述文件时,必须收集用户的信息,学习用户的兴趣,从而调整用户兴趣的权重或层次结构。根据信息源,收集用户信息的方法有显式跟踪和隐式跟踪两种。显式跟踪是指系统要求用户对推荐的资源进行反馈和评价,方法简单而直接,但多数用户不愿意进行此项工作。隐式跟踪由系统自动完成,不要求用户提供信息,分为行为跟踪和日志挖掘[2]。

用户的行为可以表现为查询、浏览页面和文章、标记书签、反馈信息、点击鼠标、拖动滚动条、前进、后退等等[2]。文献[3]对浏览页面的时间、鼠标的移动时间、鼠标的点击数和滚动的时间四种行为进行了测试,并与用户的显式评价进行比对,结果发现点击鼠标这类简单的动作不能有效地表示用户的兴趣,而浏览页面和拖动滚动条所花的时间可以有效地表示用户兴趣。

Web 站点的服务器上每天产生大量的 Web 日志数据,其中蕴涵着丰富的用户信息。利用日志可以得到访问页面的频率、页面的访问时间和页面访问顺序等信息[2,4-5]。个性化信息服务系统可以利用由挖掘日志得到的信息创建或更新用户描述文件。尽管日志的信息不够全面,不能完整地表示用户的兴趣,但还是可以从中发现许多有意义的信息[2]。

3 模糊集理论

3.1 模糊集

集合论是古典数学建立的基础,对于一个集合,一个对象属于这个集合,或者一个对象不属于这个集合,两者必居其一,且仅居其一,绝不能模棱两可。这个要求大大限制了数学的应用范围,使其无法处理日常生活中大量的不明确模糊现象与概

作者简介:李珊(1980-),女,助教,主要研究方向为数据挖掘及系统安全等;云彩霞(1981-),女,助教,主要研究方向为无线通信与计算机应用等;白彦霞(1979-),女,助教,主要研究方向为信息检索等;张秋菊(1982-),女,助教,主要研究方向为信号与信息处理等;李丽芬(1982-),女,助教,主要研究方向为信号与信息处理等。

收稿日期:2008-09-15 **修回日期**:2008-12-08

念,例如,“年轻”、“年老”之类的概念均没有明确的定义。Zadeh于1965年提出的模糊集[6]概念是对普通集合的一种推广,并奠定了模糊数学的理论基础。

定义1 假设 U 是一个论域,U 上的一个模糊集合 $\widetilde{A}$ 是由 U 上的一个实值函数

$$\mu_A: U\to[0,1]$$

表示。对于 $u\in U$,$\mu_A(u)$ 称为 u 对于 $\widetilde{A}$ 的隶属度,而 μ_A 称为 $\widetilde{A}$ 的隶属函数。通常用 $\widetilde{A}(u)$ 表示 $\mu_A(u)$。

$\mu_A(u)$ 的值表示 u 属于 $\widetilde{A}$ 的程度。$\mu_A(u)$ 的值越接近1,u 属于 $\widetilde{A}$ 的程度就越高;相反,$\mu_A(u)$ 的值越接近0,u 属于 $\widetilde{A}$ 的程度就越低。

3.2 模糊数

定义2 假设 $\widetilde{A}\in\widetilde{R}$,称 $\widetilde{A}$ 为一个模糊数,如果它是正则凸模糊集;称 $\widetilde{A}$ 为闭模糊数,如果它是正则闭凸模糊集;称 $\widetilde{A}$ 为有界闭模糊数,如果它是正则有界闭凸模糊集;称 $\widetilde{A}$ 为有限闭模糊数,如果它是有限闭凸模糊集。

定义3 三角模糊数 (a_1,a_2,a_3) 隶属函数为:

$$\widetilde{A}(x)=\begin{cases}0 & x<a_1\\ (x-a_1)/(a_2-a_1) & a_1\leqslant x\leqslant a_2\\ (a_3-x)/(a_3-a_2) & a_2\leqslant x\leqslant a_3\\ 0 & x>a_3\end{cases}$$

3.3 模糊综合评判方法[1]

综合评判是指综合考虑受多种因素影响的事物或系统并对其进行总的评价,当评价因素具有模糊性时,这样的评价被称为模糊综合评价,又称模糊综合评判。

模糊综合评判的数学模型:

首先建立影响评价对象的 n 个因素组成的集合,称为因素集:

$$U=\{u_1,u_2,\cdots,u_n\}$$

然后建立由 m 个评价结果组成的评价集:

$$V=\{v_1,v_2,\cdots,v_m\}$$

再对各因素分配的权值,建立权重集,即表示为权向量:

$$\widetilde{\boldsymbol{A}}=(a_1,a_2,\cdots,a_n)$$

式中 a_i 为对第 i 个因素的加权值,一般规定 $\sum_{i=1}^{n}a_i=1$。

对第 i 个因素的单因素模糊评价为 V 上的模糊子集

$$\widetilde{R}_i=(r_{i1},r_{i2},\cdots,r_{im})$$

于是单因素评价矩阵 $\widetilde{\boldsymbol{R}}$ 为:

$$\widetilde{\boldsymbol{R}}=\begin{bmatrix}r_{11} & \cdots & r_{1m}\\ \vdots & & \vdots\\ r_{n1} & \cdots & r_{nm}\end{bmatrix}$$

则对该评判对象的模糊综合评价 $\widetilde{\boldsymbol{B}}$ 是 V 上的模糊子集 $\widetilde{\boldsymbol{B}}=\widetilde{\boldsymbol{A}}\circ\widetilde{\boldsymbol{R}}$。

根据权重集 $\widetilde{\boldsymbol{A}}$ 与单因素模糊评价矩阵 $\widetilde{\boldsymbol{R}}$ 合成,进行模糊综合评价求取评价模糊子集 $\widetilde{\boldsymbol{B}}$,一般有五种模型。其中一个模型为 $M(\wedge,\vee)$,根据 $\widetilde{\boldsymbol{B}}=\widetilde{\boldsymbol{A}}\circ\widetilde{\boldsymbol{R}}$,可以写为:

$$\widetilde{\boldsymbol{B}}=(a_1,a_2,\cdots,a_n)\circ\begin{bmatrix}r_{11} & \cdots & r_{1m}\\ \vdots & & \vdots\\ r_{n1} & \cdots & r_{nm}\end{bmatrix}=(b_1,b_2,\cdots,b_n)$$

其中,$\widetilde{\boldsymbol{B}}$ 中第 j 个元素 b_j 可由下式计算

$$b_j=\bigvee_{i=1}^{n}(a_i\wedge r_{ij}),j=1,2,\cdots,m$$

这种方法主要通过取小及取大两种运算,因此,称该种模型为 $M(\wedge,\vee)$ 模型。这种方法当因素比较多时,对每一因素的加权值必然很小,会导致评价结果不理想。因此,对权系数 a_i 加以修正,并归一化。

4 模糊综合评价用户兴趣度

本文实验所用的数据是通过 Web 浏览器收集来的,用 Microsoft Visual Basic.Net 语言实现了 Web 浏览器的设计,此浏览器具有其他浏览器的基本功能,还具有用户浏览网页时记录用户行为的功能。数据包括跟踪用户行为所得到的浏览页面的时间和页面滚动的时间,在日志中记录的访问页面的时间和统计得到的页面访问频率,以及用户浏览的 URL 地址。收集得到的数据记录在数据库内。每当用户输入新的 URL 地址或关闭正在浏览的页面时,要求用户对刚刚浏览过的页面根据兴趣进行评分,作为用户的显式信息,存入数据库。

对已得数据进行数据预处理,包括数据清洗:把日志文件中的后缀为 gif、jpg、jpeg、GIF、JPG、JPEG 等的记录删除;用户识别:采用的技术就是基于日志/站点的方法,应用启发式规则来进行;会话识别:设置的 *timeout* 值为 30 min。在数据上应用模糊综合评判方法,判断出用户感兴趣的页面。

(1)确定被评判因素集 $U=\{u_1,u_2,\cdots,u_n\}$,选择“浏览页面的时间”、“页面滚动的时间”、“访问页面的时间”和“页面访问频率”这几个属性为评判因素。

(2)确定备择集 $V=\{v_1,v_2,\cdots,v_m\}$,又称评语集,“不感兴趣”、“一般”、“感兴趣”和“非常感兴趣”为评语集。

(3)单因素评判,即对单个被评判因素 u_i 评判,得到 V 上的模糊集 $\{r_{i1},r_{i2},\cdots,r_{im}\}$,即从 U 到 V 的一个模糊映射。从而构成评判矩阵。

$$\boldsymbol{R}: U\times V\to[0,1],\boldsymbol{R}=\begin{bmatrix}r_{11} & \cdots & r_{1m}\\ \vdots & & \vdots\\ r_{n1} & \cdots & r_{nm}\end{bmatrix}$$

即将单个被评判因素 u_i 分别评判,先将单个被评判因素 u_i 应用公式(1)所示的 min-max 方法标准化为 0-1 之间的数。

$$v'=\frac{v-min}{max-min}(new\_max-new\_min)+new\_min \qquad (1)$$

其中,v' 为标准化后的数值,v 为要进行标准化的数值,min、max 为被标准化的数值中的最小值、最大值,new_min、new_max 为标准化后的最小值、最大值,这时 $new\_min=0$,$new\_max=1$。

标准化处理后的单个评判因素 u_i 应用图1所示的隶属函数分别得到 V 上的模糊集 $\{r_{i1},r_{i2},\cdots,r_{im}\}$。

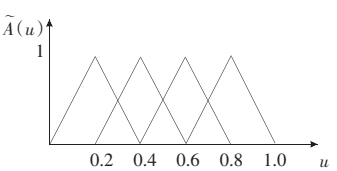

图1　模糊化隶属函数

所有单个评判因素构成了模糊矩阵 $\boldsymbol{R}=\begin{bmatrix}r_{11} & \cdots & r_{1m}\\ \vdots & & \vdots\\ r_{n1} & \cdots & r_{nm}\end{bmatrix}$。

(4)确定权重集,给每个被评判因素赋一个权重值 w_i,建立权重集。$W=\{w_1,w_2,\cdots,w_n\}$,本实验分别设为 0.25、0.25、0.25、0.25,即表示"浏览页面的时间"、"页面滚动的时间"、"访问页面的时间"和"页面访问频率"这几个因素与用户兴趣的关联程度基本相同。

(5)计算模糊综合评判的结果,结果集表示为:$\boldsymbol{B}=\boldsymbol{W}\cdot\boldsymbol{R}$,即

$$\boldsymbol{B}=(w_1\ w_2\ \cdots\ w_n)\begin{pmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & & \vdots \\ r_{n1} & \cdots & r_{nm} \end{pmatrix}=(b_1\ b_2\ \cdots\ b_m)$$

本文实验应用第 3.3 节介绍的模型,主要通过取小及取大两种运算。

(6)模糊综合评判结果集的单值化。利用加权平均的方法单值化结果集。

$$P=\frac{\sum_{i=1}^{m} a_i b_i^k}{\sum_{i=1}^{m} b_i^k}$$

常数 k 的值人为确定。本文实验 $k=2$,$a_i=0.1,0.2,0.3,0.4$,$i=1,2,3,4$。这样若用户对该网页"感兴趣"或"非常感兴趣",计算所得的值则较大。

5 实验结果评估方法与分析

本文得到 6 组实验结果,将应用模糊综合评判方法计算得到的兴趣度的评价与用户的真实评价进行比较。实验结果用曲线图表示,与用户真实评价对应的曲线进行比较,与用户真实评价对应曲线的接近度值越小,表示曲线的走势越接近,与用户真实评价越一致。应用公式(2)计算曲线接近度(CAD)。

$$CAD=\sum|M_i-U_j| \quad (i=1,\cdots,n;j=1,\cdots,n) \tag{2}$$

其中,$M_i(i=1,\cdots,n)$表示应用模糊综合评判方法对应的曲线函数计算得到的点 i 的兴趣度,U_j 表示应用用户真实评价对应的曲线函数计算得到的点 j 的兴趣度,将所有点对应的差值的绝对值相加,所得和值表示模糊综合评判方法对应曲线与用户真实评价对应曲线的接近度。

实验结果如表 1 所示。用户 2 所得数据最不接近用户真实评价,用户 4 次之,用户 6 居第三位,用户 1 和用户 5 并列第四位,用户 3 最接近用户真实评价。分析用户 2、用户 4 和用户 6 的原始数据,发现用户的真实评价与其浏览行为不符。用户评价很高,但分析浏览器记录下数据表示用户对此网页不感兴趣;用户评价很低,但分析浏览器记录下的数据表示用户对此网页很感兴趣。分析用户 1、用户 5 和用户 3 的原始数据,用户的真实评价与其浏览行为基本相符。从整体上看,模糊综合评判方法对应的曲线与用户真实评价所对应的曲线的变化趋势基本相同。

表 1 模糊综合评判方法与用户真实评价的接近度

用户	CAD	用户	CAD
用户 1	18.0	用户 4	22.6
用户 2	30.0	用户 5	18.0
用户 3	15.0	用户 6	21.0

由以上结果可以看出,用户的兴趣与用户的行为密切相关,单纯考虑日志记录不能准确表示用户的真实兴趣。

6 结束语

由于对个性化服务系统实际需要,研究者通过各种方式挖掘用户的兴趣。但是用户的兴趣是多方面的、变化的,跟踪和学习用户的兴趣仍是一个最基本和难以解决的问题,有待进一步的研究。本文将用户行为的信息和 Web 日志的信息结合起来,应用模糊综合评判方法挖掘用户的兴趣,实验证明此方法能较准确表示用户的真实兴趣。

参考文献:

[1] 李士勇.工程模糊数学及应用[M].哈尔滨:哈尔滨工业大学出版社,2004.

[2] 曾春,邢春晓,周立柱.个性化服务技术综述[J].软件学报,2002,13(10):1952-1961.

[3] Claypool M,Le P,Waseda M,et al.Implicit interest indicators[C]//Campbell M.Proceedings of the ACM Intelligent User Interfaces Conference(IUI).New York:ACM Press,2001:14-17.

[4] 冯鉴,姚敏.基于模糊理论的因特网个性化服务应用[J].计算机应用与软件,2004,21(7):74-76.

[5] 王勋,凌云,费玉莲.基于 Web 日志和缓存数据挖掘的个性化推荐系统[J].情报学报,2005,24(3):324-328.

[6] 胡宝清.模糊理论基础[M].武汉:武汉大学出版社,2004:202-208.

[7] Wong C,Shiu S,Pal S K.Mining fuzzy association rules for web access case adaptation[C]//Fourth Internat Conf on Case-Based Reasoning,July,2001:213-220.

[8] 孟雷.多因素顾客满意度总体评价计算方法研究[J].世界标准化与质量管理,2003,8:15-17.

(上接 92 页)

[3] 耿方萍,朱祥华.基于本体的网络资源表示研究[J].计算机应用,2003,23(4):4-6.

[4] 李伟,徐志伟.一种网格资源空间模型及其应用[J].计算机研究与发展,2003,40(12):1756-1762.

[5] 张传富.仿真网格资源管理系统关键技术研究[D].长沙:国防科技大学研究生院,2006.

[6] 徐志伟,李伟.织女星网格的体系结构研究[J].计算机研究与发展,2002,39(8):923-929.

[7] Geng L,Price J D,Srinivas T K.Network information models and one vision architecture[J].Bell Labs Technical Journal,1998,3(4):208-221.

[8] IETF RFC 1213 Management information base for network management of TCP/IP-based Internet:MIB-II[EB/OL].http://www.ietf.org/rfc/rfc1213.txt.

[9] OpenLDAP foundation:Community developed LDAP software[EB/OL].http://www.openldap.org.

[10] Nabrzyski J,Schopf J M,Weglarz J.Grid resource management-state of the art and future trends[M].Boston:Kluwer Academic Publishers,2003.