

开放式用户模型服务平台的设计与实现

王巧容, 陈庆奎, 赵海燕

(上海理工大学 光电信息与计算机工程学院, 上海 200090)

(happywangqiaorong@163.com)

摘要:为了构建一个公共的共享数据的用户模型平台,给各个接入该平台的网站提供更全面、更准确的用户信息,平台提供了数据接口和算法接口用于与第三方网站的交互,重点研究了如何解决来自不同数据源的用户数据的冲突,从而形成统一的用户模型的问题,最终实现了算法和模型以及数据的共享。实验结果表明,该平台更能准确、全面地构建用户模型。

关键词: Web 服务; 用户建模; 数据融合; 算法共享; 数据共享

中图分类号: TP391; TP301.6 **文献标志码:** A

Design and implementation of open user model service platform

WANG Qiao-rong, CHEN Qing-kui, ZHAO Hai-yan

(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200090, China)

Abstract: To build a common user model platform which can share data and provide more comprehensive and accurate user information to all sites accessed to the platform, the platform provided data interfaces and algorithm interfaces to interact with the third-party sites, focused on how to solve the conflict of user data from different data sources in order to form a unified user model, finally achieved sharing of algorithms, models and data. The experimental results show that it is more accurate and comprehensive to build user model on this platform.

Key words: Web service; user modeling; data fusion; algorithm sharing; data sharing

0 引言

在电子商务蓬勃发展的今天,网上购物作为一种新兴事物得到了很多人的认可和喜爱,足不出户,就可以买到喜欢的东西。由于网上的商家越来越多,如何留住用户,并快速准确地向用户推荐合适的商品是每一个网上商家都要考虑的问题。根据已有的用户网上行为对用户进行建模,并根据模型向用户推荐商品是解决这个问题的一个有效途径。研究人员提出了多种用户模型的构建方法^[1-3]和算法^[4-6];很多网站采用了用户建模的方法,以根据用户模型进行个性化产品、服务和内容的推荐^[7-8]等。

在实际应用中,由于某种原因,例如新用户或者网站推出了新的内容,某一网站上相关的数据可能较少,从而造成了用户模型的准确度不高;用户的信息可能分布在不同的网站上,例如一个用户,他可能在当当网上买人文类的书,在卓越网上买科技类和外文书。通过将多个网站上的用户模型数据进行集成和融合,形成统一的用户模型,并为这些网站提供用户模型服务,显然将大大提高用户模型的准确度和全面性。从商业的角度看,由于用户模型准确度的提高将给它带来利润的提高,因此具有共享用户模型信息的动力。另一方面,对于某一运营商,将多个版块(或子网站)中的用户模型进行集成,同样存在必要性。

在构造一个统一的用户模型服务平台时,需要解决一系列关键技术。由于各个网站提供的用户模型数据的格式并不

统一,因此需要解决格式的统一问题;而且对于相同项目,各个网站提供的数据可能也不一致,需要将不同网站的数据进行融合形成一个一致的结果。

本文针对这种需求,提出了一个开放式用户模型服务平台的体系结构,并针对其中的关键技术进行了研究。

1 开放式用户模型服务平台体系结构

图1为开放式用户模型服务平台的体系架构。

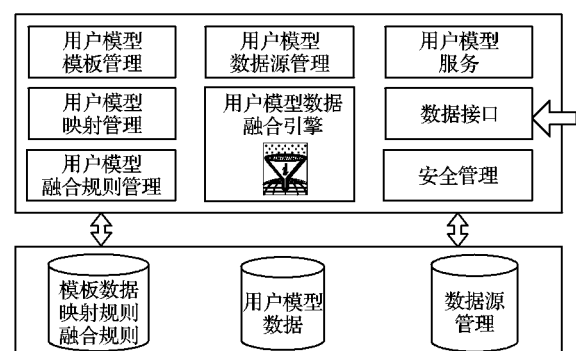


图1 开放式用户模型服务平台

各个用户模型数据源通过注册模块可以动态添加到系统中。不同的领域涉及到的用户模型信息是不同的。因此,在本文的平台中,首先由用户建模专家根据业务或者领域的需要创建用户模型模板,该用户模型模板定义了从哪些方面来描述用户的兴趣,它提供了构建用户模型的规范化公共定义;

收稿日期: 2010-09-06; **修回日期:** 2010-10-26。 **基金项目:** 国家自然科学基金资助项目(60970012; 60873230); 上海信息技术领域重点科技攻关项目(09511501000); 上海重点科技项目(09220502800); 上海市重点学科建设项目(S30501); 上海市科委基础研究重点课题项目(08JC1411700)。

作者简介: 王巧容(1985-),女,湖北天门人,硕士研究生,主要研究方向:服务计算; 陈庆奎(1966-),男,黑龙江哈尔滨人,教授,博士生导师,主要研究方向:计算机群、并行数据库、并行理论、网络; 赵海燕(1975-),女,河南温县人,高级工程师,主要研究方向:服务计算。

通过用户模型映射管理模块定义的映射规则,不同数据源的信息将映射到统一的用户模型上,平台提供了一个标准的接口,用于定期采集各个第三方网站提供的各自的用户数据并按照映射规则对这些用户数据进行格式转换,将第三方的用户数据进行标准化;在采集第三方数据的过程中,通过安全管理模块对传递的数据进行加密,以免造成商业数据的泄露;而用户模型融合规则管理模块定义了融合规则,以对不同数据源中提供的不一致信息进行统一化(融合)。用户模型数据融合引擎是整个平台的核心,它将定时自动采集各个用户模型数据源的数据进行整合。

整个平台将向外界提供用户模型服务,包括整个用户模型的获取,部分用户模型的获取等;在获取用户模型时,通过安全管理模块验证用户的登录,以确定其有获取该模型的权利。

平台的数据流图如图2所示。

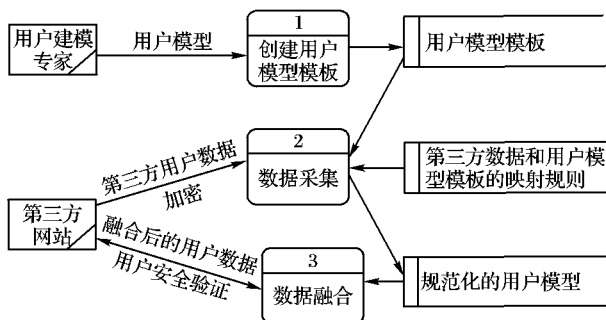


图2 开放式用户模型服务平台的数据流图

为了实现该平台,需要解决一系列的技术问题。其中最为重要的问题有两个:一个是数据格式冲突问题;另外一个数据内容冲突问题。后面两章将对这两个问题的解决方法进行介绍。

2 用户模型数据的格式转化

由于各个网站的独立性和内容差异性,来自不同网站的用户模型数据格式并不统一,需要对它们进行数据转换,使其与用户模型模板文件一致。

通常,各个网站的用户模型数据元素与用户模型模板数据元素存在以下几种对应关系。

- 1) 一一对应的关系。数据源和用户模型模板的元素是一一对应的,但是两者的名称可能一样或者不一样。
- 2) 多对一的关系。数据源元素的信息层次比较低,数据源的多个元素对应于用户模板的一个元素。
- 3) 一对多的关系。数据源元素的信息层次比较高,对应于用户模型模板的多个元素。

例如,假设网站A的用户数据模型为{名称,家用电器,手机通信,手机配件,数码影像,数码配件,日用百货},用户模型模板为{用户名称,家用电器,手机数码,家具用品,服装鞋帽},则两者的元素对应关系如图3所示。

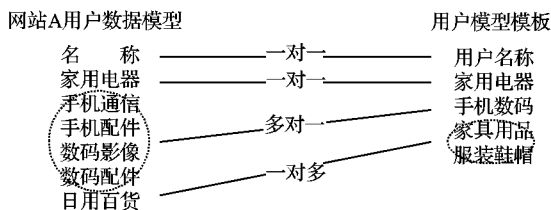


图3 数据对应关系

图3中,网站A的“名称”与用户模型模板的“用户名称”

是一一对应的关系,但两个元素的名称不一致;网站A的“家用电器”与用户模型模板的“家用电器”也是一一对应的关系,两者的名称也是一致的;网站A的“手机通信”、“手机配件”、“数码影像”、“数码配件”是比较低层的产品分类,一起对应于用户模型模板的“手机数码”,是多对一的关系;网站A的“日用百货”是比较高层次的产品分类,对应于用户模型模板的“家具用品、服装鞋帽”,是一对多的关系。

为了完成数据转换,本文采用了XSLT转化技术。各个网站以XML格式定期提交各自的用户数据,用户模型模板也是XML格式的,可以事先建立两者Schema之间的XSLT文件。在运行时,直接应用XML的转换机制达到目的。值得指出的是,对于多对一的关系,本文将数据源的多个元素的值取均值作为目的元素的值;对于一对多的关系,本文将数据源的元素的取值复制到多个目的元素。

3 用户模型数据的内容一致性处理

对于同一个用户,不同的网站可能有不同的信息。如张三可能在网站A买了很多手机数码产品,所以网站A认为张三对手机数码产品很感兴趣,而张三很少在网站B购买手机数码产品,则网站B认为张三对手机数码产品不感兴趣。为了提供合理的用户模型信息,需要把各个网站中有关用户模型的内容进行一致性处理。

如果把每一个网站看做一个评价者,用户模型数据内容一致性处理就可以看成为一个群体综合评价问题^[9-10],即将各个网站的用户模型数据(评价)按照某种方式进行合成得到一致的用户模型数据(综合评价结果)。

在应用群体综合评价方法时,依据用户模型融合的特点,需要考虑以下几个方面的问题。

1) 评价值的模糊性。

由于技术原因,或者由于商业隐私方面的考虑,网站提供的用户模型信息一般不会是一个精确值,而更可能是一个模糊值。对此,本文采用三角模糊数来统一表示用户模型中的元素的语言评价,如图4所示为各个语言评价对应的隶属度函数。模糊数的界限一般并不明显,本文用0~0.25表示非常不喜欢,0~0.5表示不喜欢,0.25~0.75表示一般,0.5~1.0表示喜欢,0.75~1.0表示非常喜欢,它们之间互相有交叉。

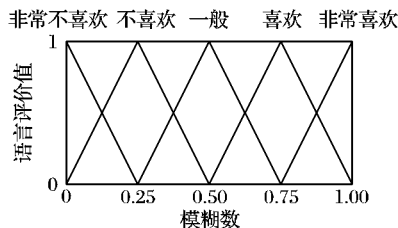


图4 数据模糊表示

2) 网站数据质量差别。

在实际生活中,各个网站的大小、规模、建站时间、知名度、人气等不一样,网站A和网站B对于最终用户模型数据的形成应该具有不同的权重。例如,网站A的用户数远远大于网站B的用户数,则它对用户的评价对合成结果的影响应该比网站B的大。

3) 各个网站评价一致性的影响。

在将各个网站的用户数据进行合成时,除了考虑各个网

站的权威性之外,还需要考虑群体意见的一致性。如:虽然网站 A 的权重比较大,但是其他权重小的网站的意见比较一致,则也要考虑如何尊重多数网站的意见。

目前的群体评价方法一般采用加权合成的方法考虑专家的不同权重对结果的影响,也有一些评价方法考虑了群体意见的一致性来合成评价意见。在将各网站的评价意见进行合成时,既要考虑各个网站的权威性也要兼顾到一致性。本文根据这种思想,使用了将网站权重和一致性进行有机结合的方法,使最终合成的群体评价结果更为合理。

1) 网站权重影响。

依据由于规模、知名度、以往数据准确性等,确定网站的权重,本文将网站 i 的权重记为 w_i 。

2) 数据一致性影响。

首先定义两个三角模糊数的相似度: 设 $R_i = (l_i, m_i, r_i)$, $R_j = (l_j, m_j, r_j)$ 为两个三角模糊数, $S(R_i, R_j)$ 为其相应的隶属度函数, 则称:

$$S(R_i, R_j) = \frac{\int_x \min(R_i(x), R_j(x)) dx}{\int_x \max(R_i(x), R_j(x)) dx} \quad (1)$$

为两个三角模糊数的相似度。

由于用户模型元素的取值用三角模糊数来表达, 因此有下面的定义: 两个网站 N_i 与 N_j 中对用户模型某个元素的取值分别为两个模糊数 $R_i = (l_i, m_i, r_i)$, $R_j = (l_j, m_j, r_j)$, 将两个元素取值之间的一致度 S_{ij} 定义为 R_i 和 R_j 的相似度, 即 $S_{ij} = S(R_i, R_j)$ 。

为了衡量某一网站用户模型元素取值与其余所有网站用户模型元素取值的一致性, 定义该网站的用户模型元素的平均一致度 S_i 为:

$$S_i = \frac{1}{n-1} \sum_{j=1, j \neq i}^n S_{ij} \quad (2)$$

为了便于比较, 将平均一致度进行归一化, 从而得到各网站针对用户模型某一元素取值的相对一致度, 网站 N_i 元素取值的相对一致度 S_i' 如式(3)所示:

$$S_i' = S_i / \sum_{i=1}^n S_i \quad (3)$$

网站 N_i 的用户模型元素取值的相对一致度反映了它与其他网站上对该元素取值的一致性程度, 其值越大越说明它的取值代表了大多数网站的取值。为了集结多数网站的取值, 获得一致性程度高的结果, 应该考虑各网站的相对一致度的影响。

3) 网站权重和数据一致性的综合影响。

通过上面的分析可知, 网站的评价对于合成结果的影响大小与网站的权重、网站取值的相对一致度有关。因此。将这两项进行合成得到用户模型元素取值的合成权重:

$$C_i = \alpha_1 W_i + \alpha_2 S_i' \quad (4)$$

其中 $\alpha_1 + \alpha_2 = 1$ 。对式(4)中系数的取值, 若只考虑权威的作用而不考虑一致性的影响, 则 $\alpha_1 = 1, \alpha_2 = 0$ 。若只考虑网站取值一致度的影响, 则 $\alpha_2 = 1, \alpha_1 = 0$ 。

4 系统实现

为了验证上述方法的可用性, 本文实现了一个平台原型。假定数据来自三个不同的网站: A 网站, B 网站, C 网站, 考虑到三个网站的用户规模大小不同, 假设这三个网站的权重分别为 0.5、0.3 和 0.2。

在该例子中, 采用用户所购买的产品树及权重来描述用户

的兴趣。其中, 权重表示了用户对该产品的兴趣程度, 通常各网站可通过用户对该产品的浏览次数或者购买次数来获得相关的权重。由于各个网站的商业性质, 它们提供的数据一般都不是一个确定的数据, 通常采用一种模糊的表示, 假设该例子中分为 6 类: “非常喜欢”、“喜欢”、“一般”、“不喜欢”、“非常不喜欢”, 各个网站提交的用户张三的兴趣如表 1~3 所示。

表 1 网站 A 提供的张三的数据

类别	兴趣	类别	兴趣
家用电器	非常喜欢	数码影像	一般
手机通信	一般	数码配件	不喜欢
手机配件	不喜欢	日用百货	一般

表 2 网站 B 提供的张三的数据

类别	兴趣	类别	兴趣
家用电器	喜欢	家具用品	一般
手机数码	喜欢	服装鞋帽	喜欢

表 3 网站 C 提供的张三的数据

类别	兴趣	类别	兴趣
家用电器	不喜欢	家具用品	非常喜欢
手机数码	一般	服装鞋帽	非常喜欢

最终, 系统将根据 A、B、C 网站提交的用户兴趣数据的格式, 对比分析字段, 创建合适的 XSLT 转换文件, 导入数据并进行数据融合, 得到综合的用户兴趣数据。

整个系统的使用过程如下。

1) 用户建模专家添加用户模型模板, 添加完成后的用户模型模板界面如图 5 所示。

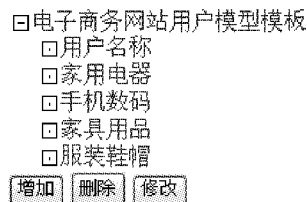


图 5 电子商务网站用户模型模板

2) 创建数据转换文件。首先将第三方网站提交的数据转换为三角模糊数, 即将模糊表示的 XML 文件转换为用三角模糊数表示的 XML 文件, 然后根据各个网站数据与用户数据模板的数据映射关系, 编写转换用的 XSLT 文件, 对于每个网站, 该 XSLT 文件只需编写一次。

在该例子中, 只有 A 网站的数据需要转换, 图 6 所示为其数据转换 XSLT 文件。

3) 各个网站以 XML 格式定期提交各自的用户数据。系统根据第一次编写的相应的 XSLT 文件对数据进行转化, 得到用户模型模板格式的用户数据。

网站 A 的源数据和转换数据如图 7 所示。

4) 最后, 平台采用式(4)所示的方法将三份数据融合, 在这里, 取 $\alpha_1 = 0.6, \alpha_2 = 0.4$, 得到数据融合的结果如图 8。

将数据融合得到的结果代入三角模糊数, 重新转换为模糊表示, 可以得到张三的兴趣爱好如表 4 所示。

表 4 融合多个网站数据所得到的张三的兴趣爱好

类别	兴趣	类别	兴趣
家用电器	非常喜欢	家具用品	一般
手机数码	喜欢	服装鞋帽	非常喜欢

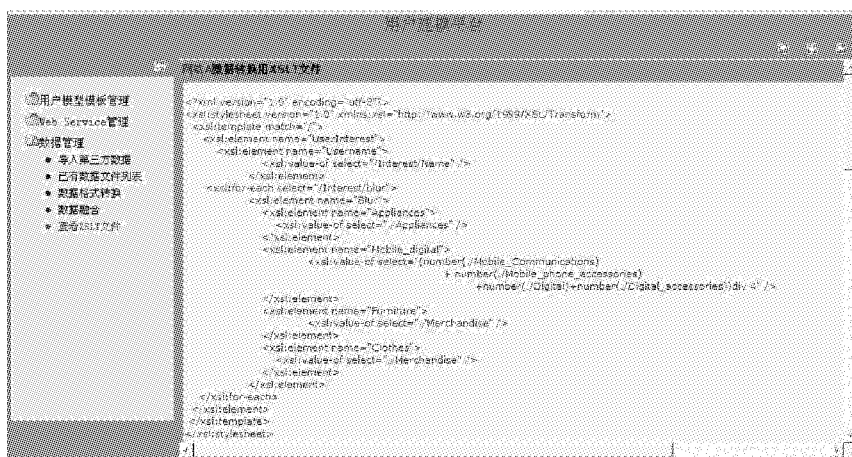


图6 XSLT文件截图



图7 数据转换

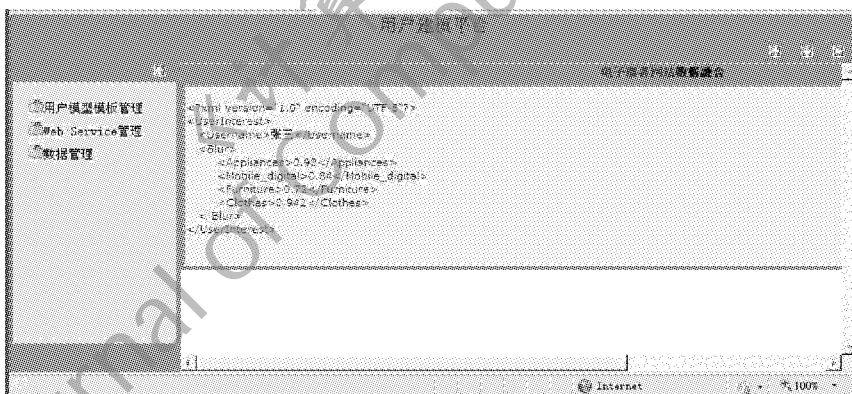


图8 数据融合

5 结语

本文介绍了一个开放式用户模型服务平台,该系统首先由用户建模专家根据业务或者领域的需要创建用户模型模板文件,然后对传入的用户数据进行格式化,使网站数据格式与模型模板文件的格式一致,然后融合各个不同网站的数据,对用户建模。本文详细介绍了实现该系统的主要思想和关键技术,并实现了一个面向网上购物的例子,证实了系统的可行性。

参考文献:

- [1] 戴洪钧. 面向个性化服务的用户建模相关问题研究[J]. 情报方法, 2006(3): 77-79.
- [2] 方惠敏, 杨国胜, 丁文珂. 基于人性化网站界面设计的用户建模[J]. 计算机技术与发展, 2008, 18(2): 187-190.
- [3] 应晓敏. 面向 Internet 个性化服务的用户建模技术研究[D]. 长沙: 国防科学技术大学, 2003.
- [4] SYMEONIDIS P, NANOPOULOS A, MANOLOPOULOS Y. Feature-weighted user model for recommender systems [C]// UM07: Proceedings of the 11th International Conference on User Modeling, LNCS 4511. Berlin: Springer, 2007: 97-106.
- [5] ZHANG H, SONG H T. Construction of ontology-based user model for Web personalization [C]// UM07: Proceedings of the 11th International Conference on User Modeling, LNCS 4511. Berlin: Springer, 2007: 67-76.
- [6] 周彩兰, 王鹏. 基于空间向量模型的用户建模算法改进[J]. 计算机与数字工程, 2010, 38(2): 15-17.
- [7] 当当网. 个性化推荐[EB/OL]. [2010-10-24]. <http://reco.dangdang.com/>.
- [8] Amazon.com. Online shopping for electronics, apparel, computers, books, DVDs & more [EB/OL]. [2010-10-24]. <http://www.amazon.com>.
- [9] DYER J S, SURLIN R K. Group preference aggregation rules based on strength of preference [J]. Management Science, 1979, 25(9): 22-34.
- [10] 司艳杰, 魏法杰. 基于二元语义的项目成功度群体综合评价方法[J]. 系统工程, 2009, 27(3): 73-78.