个性化推荐系统中的混合用户偏好获取①

蒋 翀[1] (湖南女子职业大学 现代教育技术中心 湖南 长沙 410004)

费洪晓[2] (中南大学 软件学院 湖南 长沙 410079)

摘 要: 针对个性化推荐系统中用户偏好信息量小的问题，提出了混合用户偏好获取，以相对准确但稀少的显式评分为基础，综合考虑用户页面停留时间、页面滚动时间和鼠标点击次数三项浏览行为，以评分转化规则为依据，得到隐式评分。结合隐式评分和显式评分，构建反映用户偏好的用户-项目矩阵，为个性化推荐算法的实施提供数据基础。实验证明，混合用户偏好获取是可行和有效的。

关键词: 显式评分; 隐式反馈; 混合用户偏好获取

Hyper User Preference Access in Personalized Recommendation System

JIANG Chong[1], FEI Hong-Xiao[2]

(1. Modern Education Technology Center, Hunan Woman's Vocational University, Changsha 410004, China)

(2. School of Software, Central South University of China, Changsha 410079, China)

Abstract: For the sparse data of user preferences in personalized recommendation system, a new hybrid user preference access is presented. The accurate foundation, considers the user residence time, mouse clicks, and page scrolling time, receiving implicict ratings. The user preference matrix is constructed with implicit and explicit rating, providing a data base for a recommended algorithm. Experiments proves that the hyper user preferences access is feasible and effective.

Keywords: explicit ratings; implicit feedback; mixed-mode preference access

1 引言

个性化推荐系统的质量不仅依赖于优秀的推荐技术，而且与大量可用且准确的用户数据密切相关，所以如何高效率地获得优质的体现用户偏好的数据是个性化推荐系统的核心问题之一。目前，获取用户偏好信息主要有两种方法，另一种是显式反馈，主要是通过用户对资源项目评分和填写各种形式的反馈表来实现；一种是隐式获取，主要是通过数据挖掘和应用技术从用户浏览行为中获取偏好和兴趣信息[1]。

本文提出了混合用户偏好获取，以相对准确但十分有限的显式评分为基础，通过浏览器获取用户典型浏览行为，根据评分转化规则得到隐式评分，两者共同组成反映用户兴趣偏好的用户-项目矩阵，通过实验证明混合用户偏好获取是可行和有效的。

2 获取用户偏好信息

目前获取用户偏好信息主要有显式反馈和隐式获取两种方式。显式反馈要求用户中断正常的浏览行为对资源项目进行评分；隐式获取主要是以 Web 使用挖掘、人工智能和知识发现等理论为基础，利用 Ajax、JavaScript 等应用技术分析用户浏览网页时的行为，获取用户偏好信息并转化为结构化数据的过程[2]。

2.1 显式反馈

用户主动参与电子商务网站的反馈活动可以提高个性化推荐系统性能，通过显示反馈获得的信息是用户针对特定项目或网站设置等情况最准确直接的反

① 基金项目:湖南省科技计划基金(2006JT1040)
收稿时间:2010-01-22;收到修改稿时间:2010-03-05

映。不足的是，显式反馈为用户带来了额外的负担，匆忙中填写的反馈信息很大程度上并不能客观的反映用户偏好，显式反馈宣称的高准确性也很难反映在真实数据上[3]。从心理学角度来看，人往往会将自己未来的一些喜好而不是目前的兴趣反映在显式反馈上[4]。

目前通过显式反馈获取用户偏好的情况并不乐观，显式反馈获取的信息也比较有限。为了解决这一矛盾，需要将其个性化推荐系统的概念、作用和工作原理告知所有用户，激励用户参与到这一过程。网站通过用户获取的显式反馈主要有以下三种信息[5]：

(1) 注册信息。用户在注册信息中提供的关于用户偏好有效的信息主要是“兴趣爱好”类的选项。

(2) 评分信息。用户对资源项目的评分主要分为两种情况，一是对已购买的项目进行评价；二是要求用户对系统推荐的项目进行评分。

(3) 文字评价。文字评价主要是指用户对特定项目有过购买经验，以文字形式将自身的经历和感受描述出来，以供所有用户参考借鉴的一种推荐形式。

2.2 隐式获取

这种方式的优势在于整个过程不需要用户的主动参与，不会给用户带来额外的负担，能够得到比显式反馈更丰富的用户偏好信息，进而为推荐算法提供更丰富的训练数据集，缓解数据稀疏问题[6]。但是，由于缺乏用户的主动控制，隐式获取的用户偏好信息具有一定的随机性和不确定性，这是其劣势所在。

具体来说，能够反映用户偏好的行为主要分为以下几个方面[7]：

(1) 反复进行或持续的行为。一般来说，用户在某个页面上停留的时间，鼠标的移动和点击，鼠标的滑动和拖动滚动条都是反映用户偏好的重要指标。

(2) 标记行为。这类行为主要是指用户收藏、保存和打印页面等动作，由于是用户出于某种目的而进行的主动行为，不会让用户感觉到是额外的负担。

(3) 操作行为。用户通过各种方式拷贝当前页面内容，或者在新页面内搜寻相关内容也是在一定程度上反映用户兴趣。

(4) 浏览路径。如果用户在每次打开页面时都按照固定的路径进行浏览，这也就反映出用户对某类或某种资源的特别偏好。

(5) 消极行为。如果把能够反映用户对某类或某个资源项目兴趣的行为称为积极行为，那么反映用户对资源无兴趣或厌恶的行为就被称为消极行为。

3 混合用户偏好获取

为了保留隐式和显式获取用户偏好方式的优点，尽量避免其缺点，本文提出采用混合用户偏好获取方式获得用户兴趣信息。显式反馈获得用户对资源项目直接的评分数据，隐式方式推测用户对浏览过的资源项目的评分，主要依据包括①页面停留时间，②鼠标点击次数，③鼠标移动时间，架构表示如图 1 所示。

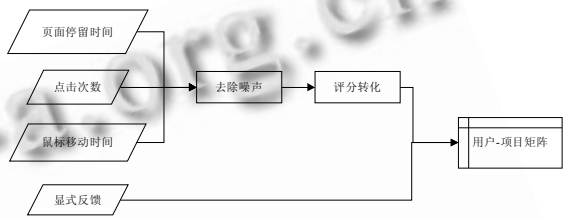

图 1　混合用户偏好获取

其中，去除噪声主要是指忽略过小或过大的页面停留时间和鼠标移动时间，对去除噪声后数据进行评分转化，推测得出用户对资源项目的隐式评分。隐式评分与显式反馈得到的显式评分在表现形式上都是若干用户对项目的评分，通过用户-项目矩阵 R_0 表示，也就是说在矩阵 R_0 中并不区分显式或隐式评分，如果用户对特定项目既有显式评分又有隐式评分，则以显示评分为准。假定用户总数是 M，项目总数是 N，创建所有用户对所有项目的评分表是没有必要的，因为肯定有用户没有对任何项目进行过评分，也有项目没有被任何用户评过分。系统将创建 m*n 项目-用户矩阵 R_0，其中 $m \leqslant M$，表示对项目评过分的用户数，$n \leqslant N$，表示被评价过的项目数，具体表示如图 2 所示。

$$R_0=\begin{pmatrix} Rt_{1,1} & \cdots\cdots & Rt_{1,k} & \cdots\cdots & Rt_{1,n} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ Rt_{s,1} & \cdots\cdots & Rt_{s,k} & \cdots\cdots & Rt_{s,n} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ Rt_{m,1} & \cdots\cdots & Rt_{m,k} & \cdots\cdots & Rt_{m,n} \end{pmatrix}$$

图 2　用户-评分矩阵

其中，$Rt_{s,k}$ 表示用户 s 对项目 k 的评分，通常是在规定范围内的一个数值。

本文设定用户的评分是一个 0~5 之间的数字，表示用户对项目的喜好程度。将浏览行为转化为隐式评

分的规则描述如下：

(1) 若用户 s 购买项目 k，但并未对项目 k 进行评分，$Rt_{s,k}=3.0$；

(2) 若用户 s 对项目 k 评分为 SCORE，$Rt_{s,k}=SCORE$；

(3) 若用户在某项目 k 页面鼠标移动时间 t_m 为 0，$Rt_{s,k}=0$；

(4) 若用户在某项目 k 页面鼠标移动时间 $tm>=0$，则根据页面停留时间 t_p 与点击次数 f 与给定的除“0”外的 10 个分数档对应，具体转化规则如关系表 1 所示。

表 1 页面停留时间与评分转化关系

页面停留时间（second）	$Rt_{s,k}$
$5\leqslant t_p\leqslant 10$	1+0.1*f
$10\leqslant t_p\leqslant 15$	2+0.1*f
$15\leqslant t_p\leqslant 20$	3+0.1*f
$20\leqslant t_p\leqslant 25$	4+0.1*f
$25\leqslant t_p$	5

(5) 其他情况下，$Rt_{s,k}=0$。

混合用户偏好获取方式保留了相对准确的显式评分，也就是说如果用户对某项目有显式反馈，则不再通过分析浏览行为的方式去获得隐式评分。如果用户并未对一些浏览过或购买的反映其兴趣的项目进行显式评分，那么在分析用户兴趣偏好时忽略这些信息显然是不合理的。因此，通过获取的页面停留时间、鼠标移动时间和点击次数，根据上述评分转化规则，得到用户对项目的隐式反馈，获取更为丰富的用户偏好数据，进而提高个性化推荐质量。

最终得到的用户-项目矩阵 R0 是一个既有显式评分又有隐式评分的矩阵，其中大部分数据是通过评分转化规则而得到的隐式评分。作为个性化推荐系统中用户兴趣建模的源数据，R0 将被进一步分割成链式结构，构建基于线性衰减的用户兴趣模型，应用到基于协作过滤的个性化推荐系统中。

4 实验结果与分析

为了证明混合用户偏好获取方式的可行性，更直观的了解用户行为与偏好的关系，本文通过混合用户偏好获取实验来验证两者之间的关系，记录鼠标点击次数，垂直和水平滚动轴滚动时间等信息，同时对当前页面进行显式评分，将获取的所有用户的行为数据与显式评分作比较，以确定某一浏览行为与用户显式反馈之间的关系。

考虑到用户浏览行为的普遍性和实验数据分析的便利性，利用过滤后数据分析用户 3 种行为①页面停留时间，②鼠标点击次数，③页面滚动时间与显式评分的关联。实验结果通过箱线图来表示，用于反映一组或多组连续型定量数据分布的中心位置和散布范围。它是利用数据中的五个统计量：最小值、第一四分位数、中位数、第三四分位数与最大值来描述数据的一种方法，图中每组数据白色矩形区域的上边缘和下边缘分别表示第一四分位数和第三四分位数，两条线段的最低点和最高点分别表示最小值和最大值，中位数在结果分析中列出。通过箱线图可以看出数据集的分布情况，特别是用于几个样本的比较。

实验结果和对比分析详述如下：

(1) 用户页面停留时间与显式评分对比

浏览器记录的页面停留时间是用户进入某页面与离开该页面之间的时间间隔。实验结果如图 3 所示。

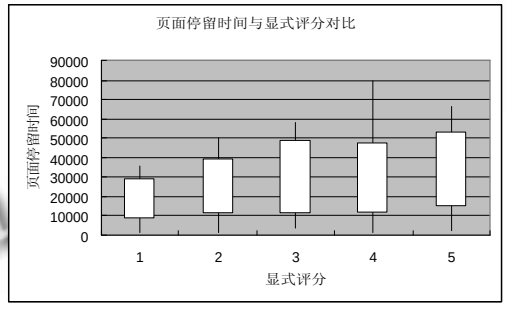

图 3 页面停留时间与显式评分对比

图 3 中未表示出来的中分位数分别为(12311，17192，20199，22395，23049)。图中时间单位为毫秒。从这组数据和图 3 可知，虽然从第三组数据到第四组数据，最大值减少，但从白色区域分布和中分位数变化来看，用户在页面的停留时间是随着显式评分的提高而呈现稳步增长的趋势，可见页面停留时间的变化可以在一定程度上反应用户兴趣的高低。

(2) 页面滚动时间与显式评分对比

页面滚动时间是指用户拖动滚动轴和滑动鼠标滚轴的时间总和。实验结果如图 4 所示。

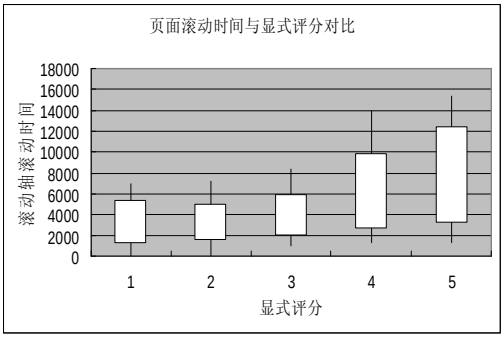

图 4　页面滚动时间与显式评分对比

评分 1-5 五组数据的中分位数为(2833，3486，4523，5639，6826)，从该组数据与图 4 来看，从第一组数据到第二组数据，白色分布区域减少，第三四分位数降低，但是从整体来看，白色分布区域、最大值和中位数都是随着显式评分的提高而逐步增长，由此可得，页面滚动时间可以成为用户偏好的隐式指标之一。

(3) 鼠标点击次数与显式评分对比

鼠标点击次数是指用户在当前页面内点击链接的次数，实验结果如图 5 所示。

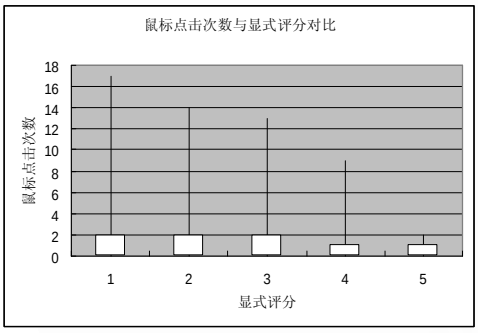

图 5　鼠标点击次数与显式评分对比

显式评分 1-5 五组数据的中分位数分别为(1，1，1，0，0)，从现有数据来看，不管是白色区域分布，还是最大值和中位数，都无法得出用户在当前页面点击链接的次数与用户对当前页面的兴趣度存在直接关系，但可作为一个辅助指标判断用户的活跃程度。

从本文进行的三项五组数据对比情况来看，页面停留时间和页面滚动时间表现较好，能够在一定程度上反映用户偏好的变化，鼠标点击次数并未表现出与用户兴趣的明显关联。同时，就单个用户行为而言，并不能完全反映其兴趣偏好，需要综合考虑多项能够反映用户兴趣的行为数据。

5　结束语

用户偏好信息的获取一直是个性化推荐系统的一个瓶颈，也是实现个性化推荐十分重要的一环，直接关系到最终的推荐质量。本文提出了混合用户偏好获取方式，以用户主动提供的显式评分数据为基础，对用户的典型浏览行为进行分析处理，得到丰富的隐式反馈数据，将两者相结合，得到大量能够反映用户兴趣偏好的基础数据，为个性化推荐算法的实施打下良好基础。通过实验证明，选取的用户浏览行为能够反映用户的兴趣偏好，证明了混合用户偏好获取的有效性和可行性。

参考文献

1 Gadanho SC, Lhuillier N. Addressing Uncertainty in Implicit Preferences. Proc. of the ACM Conference on Recommender Systems, 2007:97－104.

2 Gadanho SC, Lhuillier N. Addressing Uncertainty in Implicit Preferences. Proc. of the ACM Conference on Recommender Systems, 2007:97－104.

3 赵银春,付关友,朱征宇.基于 Web 浏览内容和行为相结合的用户兴趣挖掘.计算机工程, 2005,31(12):93－94.

4 Geyer W, Dugan C, Millen DR, Muller M, Freyne J. Recommending topics for self-descriptions in online user profiles. Proceedings of the ACM Conference on Recommender Systems, 2008:59－66.

5 邢春晓,高凤荣,战思南等.适应用户兴趣变化的协同过滤推荐算法.计算机研究与发展, 2007,44(2):296－301.

6 费洪晓,蒋翀,徐丽娟.基于树状向量空间模型的用户兴趣建模.计算机技术与发展, 2009,19(5):79－81.

7 Zhang Y, Koren J. Efficient Bayesian Hierarchical User Modeling for recommender systems. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007:47－54.