**基于触摸屏幕设备的阅读兴趣度计算方法**

**摘要：**针对个性化新闻资讯推荐服务中用户主动提供的阅读兴趣度评分比较稀疏的问题，提出一种基于触摸屏幕设备上用户操作行为数据的获取、处理和分析，在不涉及阅读内容的情况下，计算用户阅读兴趣度的方法。采用四元的线性回归模型，对用户阅读时操作触屏设备而产生的多种多样化行为数据进行综合分析和计算，实验表明该方法能够得到合理有效的兴趣度计算模型。

**关键词：**个性化推荐；用户行为；阅读兴趣度；回归分析

**1 引言**

在任何类型的推荐系统中，尽可能地获取用户准确的喜好（兴趣），是产生准确推荐结果的重要工作，对于现在更加强调个性化的推荐服务来说，更是不可轻视的前提[1]。早在十多年前，互联网刚刚进入Web时代而处于蓬勃发展的时期，就有学者通过研究用户的网页浏览行为而获取用户的兴趣，进而提供互联网新闻推荐服务[2, 10]。如今智能手机等移动设备的快速发展和普及，使人们日常生活中的阅读行为变得更加碎片化，尤其是新闻资讯类的信息获取，在智能移动终端的帮助下，几乎覆盖了人们的全部生活。因此对于新闻资讯类服务来说，要提供高质量的个性化推荐服务，将服务基础扎根于终端用户行为的获取与分析是十分必要的。

用户兴趣度的计算可以间接地为用户填充评分矩阵，对于解决基于协同过滤方法的推荐系统的数据稀疏性问题有很大帮助[7, 8, 16]，在此基础上对用户做聚类分析也有助于提高推荐质量[6]。本文介绍的是新闻资讯推荐系统中，在不进行阅读内容分析的情况下，基于阅读行为的用户兴趣度计算方法。

**2 相关工作**

在最新的研究成果中，学者Haggai Roitman等人提出了一个用户唯一性（个性化）建模方法并用实验验证了用户个性化对于推荐系统的重要性[3]。近年来，将信息检索领域的TFIDF算法引入到兴趣度计算方法方面的研究也不在少数，也有学者结合神经网络做研究[4]，但具体实现和应用的代价相对来说很大。学者孟祥武关于移动推荐系统的研究中指出移动终端在获取多元化的用户行为和获取方法方面具有很大优势[9]。

传统的基于Web网页浏览的用户兴趣度获取方法研究已经比较成熟，国内学者在用户兴趣建模方面的研究成果也已经很多[5]。在用户兴趣度量化方面，用户浏览时间依然被普遍认为是兴趣度的核心因素，因此在浏览时间获取方面有许多深入研究[11, 12]，页面的滚动次数也被认为是重要因素[11, 13]。此外，页面的浏览次数、超链接的点击次数、页面的点击次数等因素都被大量采用于相关研究。在兴趣度量化方法方面，文献[11]提出采用多元线性回归法来量化用户兴趣度，文献[12]的方法是根据经验赋予不同参数权值来计算，文献[17]中采用的是Logistic曲线模型和双曲线模型，文献[14, 15]中的方法都基于领域本体，实现代价比较大。

目前，基于触摸屏幕控制的移动智能终端设备是用户阅读行为的重要载体，但这方面的用户阅读兴趣度研究还并不多。

**3 用户阅读行为数据获取**

本文研究内容所涉及的所有数据均来源于实验室现有信息推送系统（针对校园用户）。系统向用户提供种类全面的新闻资讯订阅源，由用户自由订阅，并将最新采集到的内容推送到用户手机终端，并在终端上以网页的形式自适应重新排列后单页显示。人机交互方式与其他基于触屏控制的阅读方式完全相同。本文方法中使用到的行为数据有：阅读时间长度、阅读页面长度、阅读完整度、页面拖动（触控停止后页面静止）次数、页面自由滑动（触控停止后页面做减速滑动）次数。

**3.1 阅读时间的获取**

阅读时间长短计算的准确与否直接关系到兴趣度计算的准确性，因此必须尽可能全面地考虑各种可能影响计算的操作。

**3.1.1 页面滚动对实际阅读时间的影响**

在传统的Web页面阅读时长计算方法中，页面滚动的影响通常忽略不计，这是因为1）传统PC的计算机的屏幕相对于手机来说大很多，相对于手机来说不需要那么多的页面滚动操作；2）如今智能手机的页面大多是“卷轴”形式的，即人机交互通常采用拟物化操作以降低使用难度，所以通常情况下手指滑动页面的次数很多，在认真阅读的情况下，手指经常“拖”着页面在滑动。通过统计分析，发现大多数时候，在一次页面阅读过程中，页面处于滚动状态的时间占到总时间的10-25%。

对于页面滚动对阅读时长的影响，采取滚动中暂停计时的办法即可消除。但是有一种特殊情况需要考虑。我们在数据采集过程中，发现不少用户有在缓慢滑动页面的过程中阅读内容的习惯，而这一行为恰恰反映出了用户对当时正在阅读的某一些内容很感兴趣，因此当检测到用户长时间“拖”着页面滑动的时候，应该开启时长计时器继续计时。

**3.1.2 用户暂停阅读的影响**

判断用户是否暂停阅读也是不可缺少的，因此要设置一个合理的阈值来暂停阅读计时。一个几乎所有智能手机设备都能实现的判断方式是触控操作超时法，即页面在一定时间内没有接收到任何操作则暂停计时。如何设置阈值是重点，我们采取的方法是判断当前屏幕显示的内容在一般情况下需要多少阅读时间，例如针对iPhone5/5s，根据经验我们设置为12秒。也可以设计一个计算方法，根据当前显示内容的多少来推算超时阈值大小。

**3.1.3 阅读计时的起点**

通过实验发现，阅读计时不能以页面加载完成为起点，因为一些包含图片等超链接资源的页面有时候加载时间比较长（可能长达数秒甚至加载失败），而一般情况下用户在看到页面内容后就开始了阅读。因此时间起点除了页面加载完成，还应该有其他触发点，比如页面已显示或者点击操作（与具体终端系统平台有关）。

**3.2 阅读内容长度**

在阅读内容长度不同的情况下，不同的阅读时间长短并不能准确地表示兴趣度，为解决阅读时长的归一化问题，文献[18]提出的方法是用时长与页面内容字节数的比值来替代时长。在页面内容为文本为主的情况下，该方法适用，但对于包含超链接资源等内容的复杂页面而言，这个方法并不妥。本文采取的方法是获取阅读过的页面长度，来做归一化处理。不使用页面总长度，是因为通过实验发现这样做而得出的模型显著性相对略低。不同的终端系统获取方法可能有所差异，在iOS系统中，获取页面最大纵向偏移量即可计算出结果。在页面内容短于屏幕显示长度的情况下，我们采取根据实际内容长度对比屏幕大小进行换算的方式重新计算。

**3.3 阅读内容完整度**

通过实验发现，用户是否完整阅读了某一内容可以显著反映出用户的兴趣度高低。这一参数的获取很简单，判断页面是否滚动到最底部即可。完整阅读则值为“1”，否则为“0”。在页面内容的长度短于手机屏幕内容显示长度的情况下，完整阅读度为“1”。但通过实验我们发现，有些新闻资讯的尾部常常包含一些拓展阅读、相关阅读等等内容，而用户常常对这些内容不感兴趣，因此有的时候页面不一定能被滚动到底部。因此在我们的系统中，当用户将页面滚动超过某一临界值后，即判定用户完整阅读过该内容。例如在针对iPhone 5/5s的应用中，我们设置这一临界值为页面总长度的百分之90%。

**3.4 页面拖动次数**

观察得知，基于触摸屏幕控制的阅读操作中，拖动和滑动操作有着比较显著的差异。拖动操作往往体现出用户正专注于某部分内容，而滑动操作则比较明显地体现出用户想要快速略过某些内容的意图，因此本文的方法区分这两种操作，并希望通过实验说明这一区分的必要性。

拖动和滑动页面的方向本文的方法不加区分，因为实验表明区分方向并不能带来更好的结果。

关于拖动次数的获取，过滤掉那些无效的操作是必须的。因为观察发现，连续快速的拖动和滑动是常常出现的，其中的连续动作是无意义的，例如连续三次快速的滑动，其实和一次比较大幅度的滑动意义相当。因此本文的方法是只计算“有效动作”，方法是每一次页面静止之后，在一定时长之内页面不再滚动，则判定为一次“有效动作”。例如，通过实验，在iOS系统中，设置这一时长为0.5秒。

**3.5 页面滑动次数**

页面滑动次数的获取方式和拖动次数的获取方式相似，不再赘述。不同的系统平台具体获取方法有所差异，但有效性判断方法一致。

**4 兴趣度计算的多元线性回归方法**

文献[11]中引入了内驱力理论相关知识，解释了用户的信息反应和行为在刺激强度和诱因一定的情况下，完全取决于习惯强度和内驱力，在本文的研究内容里，可以解释为在一般情况下，用户对于阅读内容的评分取决于用户的行为习惯。因此，本文的兴趣度计算方法是收集用户的阅读行为数据与主动给出的兴趣度评分，通过多元线性回归方法，拟合出回归系数，进而为用户的每一次阅读计算出用户的兴趣值（用户不主动评分的情况下）。

**4.1 多元线性回归模型**

假设是与有关的随机变量，对于和的每一组值，若有：

（1）

其中都是与无关的未知参数，是随机误差，则称（1）为多元线性回归模型，而称：

（2）

为线性回归方程。

参数可以采用最小二乘法拟合得出。若回归方程和回归系数通过显著性检验，此模型才具有实用意义。

**4.2 二元的方法（R-2）**

阅读时间的长短和页面的滚动次数是公认的最重要的行为数据，文献[11]采用这两项数据作为回归模型的自变量，本文将该方法适配到我们的系统中作为参照实验。其中页面滚动次数为拖动次数和滑动次数的和。

设用户对页面内容的兴趣度为，阅读这一页面内容的时间为，有效滚动页面的次数为，则该方法的回归模型为：

（3）

**4.3 四元的方法（R-4）**

通过实验比对，最终确定采用一个四元的线性回归模型作为本文的兴趣度计算方法。四个自变量分别为：阅读时间与页面内容阅读长度的比值、阅读完整度、滑动次数、拖动次数，回归模型为：

（4）

**5 实验**

**5.1 实验方法**

本实验从相关性、显著性、奇异点率、平均拟合误差等指标上，并参照对比文献[11]的方法适配而来二元模型，评估所提出的四元线性回归模型。

实验使用MATLAB的统计工具包中的regress方法来求解回归模型，得出相关检验数据并进一步计算出奇异点概率和平均拟合误差。实验中使用的regress方法的调用格式为：

其中为因变量向量，为自变量矩阵，为置信水平（缺省为0.05）。是回归方程的系数矩阵，是各系数对应的置信区间矩阵，是残差向量，即真实值与计算值的差值，是残差置信区间矩阵。向量中包含四个用于检验回归模型的统计值，是相关系数，越接近1回归方程越显著，是F-统计量，该值越大说明方程越显著。是对应的概率，当时回归模型成立。是估计误差的方差。

本实验中，因变量就是用户阅读页面后给出的兴趣度，自变量为，置信水平设置为缺省值0.05。

在调用regress方法得到一些数据后，进一步根据残差和残差置信区间计算出现奇异点的比率和回归方程的平均拟合误差。

**定义 1**（奇异点）残差过大的拟合点，也就是拟合值和真实值差距过大的点，称为奇异点，从数据上分析，是指残差的置信区间不包括0的数据点，即残差置信区间的上界小于0或者下界大于0的点。

因此奇异点比率就是奇异点的个数与总数据量的比值。

**定义 2**（平均拟合误差）真实值与拟合值差值的绝对值的平均值称为平均拟合误差。

其中为行为数据组数。

平均拟合误差是判断线性回归模型可用性的一个直观指标。

**5.2 实验数据与环境**

实验数据是我们系统中的带有用户主动兴趣评分的真实行为数据，数据规模为[?]组，数据采自iOS客户端。系统中提供的阅读源包含时政新闻、科技新闻、社会新闻、娱乐体育新闻、校园新闻等众多内容，大量阅读内容包含图片、表格、超链接等资源，内容具有普遍性。实验数据格式如表1所示。

表1 实验数据格式说明

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 用户id | 内容id | 阅读时长 | 阅读内容长度 | 阅读完整度 | 滑动次数 | 拖动次数 | 兴趣度评分 |
| \* | \* | 9.4 | 653 | 0 | 2 | 2 | 3.32 |
| \* | \* | 23.4 | 1434 | 1 | 2 | 12 | 5.14 |
| \* | \* | 18.1 | 692 | 1 | 2 | 5 | 5 |

表1中的阅读时长单位为秒，内容长度的单位是屏幕显示点单位。兴趣度评分的可选范围为1-10分，在我们的iOS客户端中“打分”操作由拖动“滑杆”控件实现，因此打出的分数不是离散的，所以评分中出现小数，但不影响实验。

完成实验的计算机软硬件环境为：AMD Athlon(tm)II X4 645 Processor 3.10Hz 处理器；4GB 内存；Windows 8.1 专业版 64位操作系统；代码编写、实验计算和画图工作在MATLAB R2014a中完成。

**5.3 实验结果**

实验以5组数据为步进，分析以上列出的指标，以及在数据组数不断增大情况下的变化情况。置信水平始终设置为0.05。所有实验得出的F-统计量对应的概率均远远小于置信水平，因此说明模型始终成立。估计误差的方差也始终在1.0左右，同样说明模型是合理的，因此不做图分析。

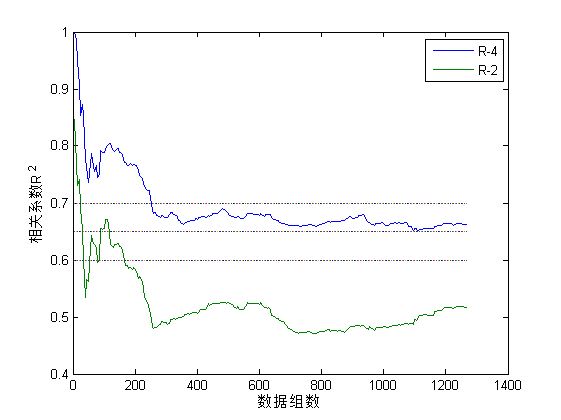


图1 不同数据组数下两种回归模型的相关系数对比

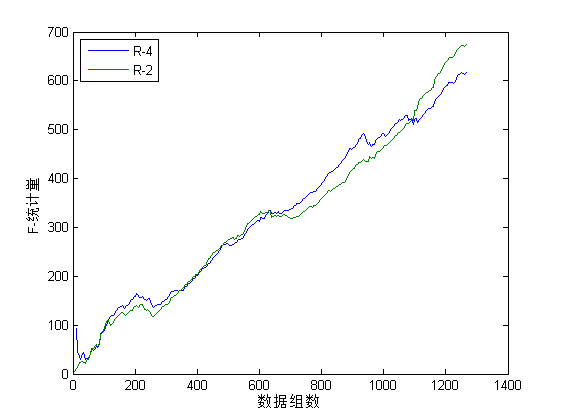


图2 不同数据组数下两种回归模型的F-统计量对比

分析图1可以发现，在现有数据下，数据组数小于200时，模型的相关性很不稳定，因此显著性也不能保证。在数据组数大于200的情况下，本文的四元的方法在相关性方面要优于二元的方法，介于0.65与0.7之间，并且相对来说更加稳定。

由图2发现，两种方法的F-统计量基本处于相同水平，并且随着数据组数的增大，总体呈现上升趋势，说明两种模型的显著性都能得到保证。由于四元的方法自变量较多，因此F-统计量更容易受到奇异点的影响，所以在该指标上四元的方法没有表现出优势。

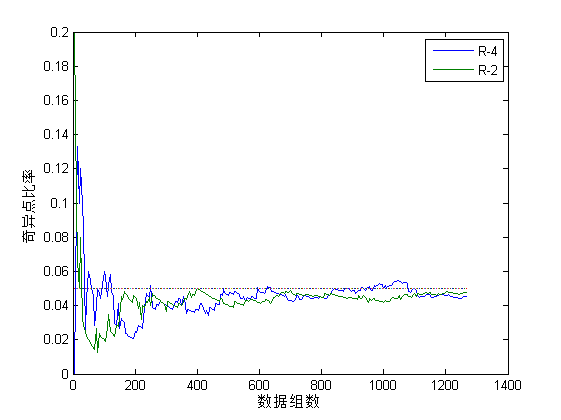


图3不同数据组数下的奇异点比率

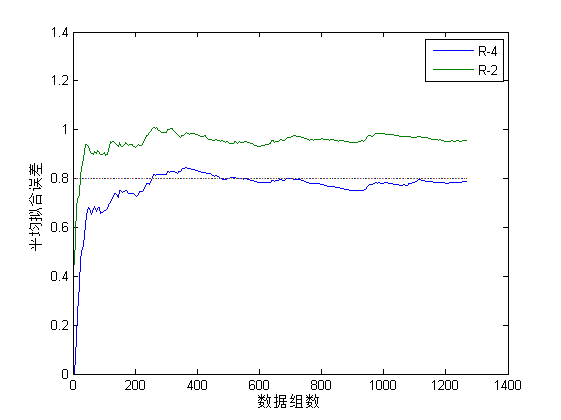
****

图4不同数据组数下的平均拟合误差

图3说明了在0.05置信水平下，两种方法的出现奇异点的比率基本都在4-5%之间，也没有呈现明显差异，都在合理水平之内。

由图4可以发现本文的四元的方法在平均拟合误差方面始终处于比较良好的水平，基本不高于0.8，也就是拟合值与真实值的偏差基本不超过0.8，对于本系统采用的10分制评分而言，这样的结果完全可以接受。二元的方法在该指标上也始终略差与四元的方法。

表2展示的是不同数据组数下的所有自变量的拟合（回归）系数。分别是自变量时间长度比、阅读完整度、滑动次数、拖动次数的系数。

表2 不同数据组数下的拟合系数

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 200 | 0.6109 | 0.5625 | 0.1514 | 0.3550 | 1.5662 |
| 400 | 0.6244 | 0.5831 | 0.1942 | 0.2308 | 1.2895 |
| 600 | 0.6491 | 0.5328 | 0.2260 | 0.2060 | 1.0748 |
| 800 | 0.6426 | 0.5846 | 0.1813 | 0.1960 | 1.0534 |
| 1000 | 0.5972 | 0.6990 | 0.1767 | 0.1987 | 1.0715 |
| 1200 | 0.5745 | 0.7569 | 0.1977 | 0.1512 | 1.0923 |
| 1300 | 0.5770 | 0.7535 | 0.1988 | 0.1499 | 1.1118 |

分析表2可知，在数据组数增大的情况下，各系数都有所变化。阅读完整度对于兴趣度的贡献在增大，并且滑动次数和拖动次数的影响也有明显区别，说明本文的四元方法考虑阅读完整度，并将两种滑动操作区分为两个自变量是有必要且有效的。

**6 结论**

本文以新闻资讯推送系统为应用背景，提出了一种移动触屏终端上基于用户操作行为数据分析的阅读兴趣度的计算方法。实验表明，该方法在尽可能多地使用移动智能终端多样化行为数据的情况下，能够用很小的计算代价得到比较显著的回归方程，完全具有可用性。在与本文类似的应用中，当用户向系统的主动的阅读兴趣度报告十分稀疏的情况下，本文的方法是一个提升推荐系统工作质量的有效解决方案。

**参考文献：**

[1] Estimating the magic barrier of recommender systems: a user study. SIGIR 2012

[2] Discovering user interests from Web browsing behavior: an application to Internet news services. HICSS 2002

[3] Modeling the uniqueness of the user preferences for recommendation systems. [SIGIR 2013:777-78](http://www.dblp.org/db/conf/sigir/sigir2013.html#RoitmanCME13).

[4] 基于商品属性的电子商务推荐系统研究[D].华中科技大学 博士学位论文

[5] 国内用户兴趣建模研究进展\_孙雨生[J]. 情报杂志 2013

[6] 基于用户聚类的异构社交网络推荐算法\_陈克寒[J]. 计算机学报 2013

[7] 基于双重邻居选取策略的协同过滤推荐算法\_贾冬艳[J]. 计算机研究与发展 2013

[8] 基于核方法的User\_Based协同过滤推荐算法\_王鹏[J]. 计算机研究与发展 2013

[9] 移动推荐系统及其应用\_孟祥武[J]. 软件学报 2013

[10] 个性化服务技术综述\_曾春[J]. 软件学报 2002

[11] 个性化服务中基于行为分析的用户兴趣建模[J]. 计算机工程与科学 2005

[12] 基于行为分析的用户兴趣建模[J]. 情报杂志 2009

[13] 移动环境下的个性化推荐用户兴趣建模研究[D]. 北京邮电大学 硕士

[14] 基于领域本体的细粒度用户兴趣建模研究[J]. 情报学报 2010

[15] 个性化服务中用户兴趣建模与更新研究[J]. 情报学报 2010

[16] Improving the Performance of Recommender Systems by Alleviating the Data Sparsity and Cold Start Problems. IJCAI 2013

[17] 面向个性化服务的用户兴趣建模及应用研究[D]. 重庆大学 硕士

[18] 基于层次向量空间模型的用户兴趣建模研究[J]. 现代计算机 2010