重大管理评论

China Management Review

2015年　第2辑

经济科学出版社

重大管理评论

China Management Review

第2辑

主　　编　刘　星
副 主 编（按姓氏拼音字母排序）
但　斌　杨　俊

经济科学出版社
2015年

图书在版编目（CIP）数据

重大管理评论 . 第 2 辑 / 刘星主编 . -- 北京：经济科学出版社，2015.10
ISBN 978-7-5141-6260-8
Ⅰ . ①重… Ⅱ . ①刘… Ⅲ . ①企业管理 - 文集 Ⅳ . ① F270-53

中国版本图书馆 CIP 数据核字（2015）第 267207 号

责任编辑：黎子民
责任编辑：刘　昕
责任印制：邱　天

重大管理评论（2015 年第 2 辑）
刘星　主编
但斌　杨俊　副主编
经济科学出版社出版、发行　新华书店经销
社址：北京市海淀区阜成路甲 28 号　邮编：100142
总编室电话：88191217　发行部电话：88191537
网址：www.esp.com.cn
电子邮件：esp@esp.com.cn
天猫网店：经济科学出版社旗舰店
网址：http://jjkxcbs.tmall.com
北京万友印刷有限公司 印装
787 × 1092　16 开　12.5 印张　220 000 字
2015 年 10 月第 1 版　2015 年 10 月第 1 次印刷
ISBN 978-7-5141-6260-8　定价 30.00 元

（图书出现印装问题，本社负责调换）
（版权所有　翻印必究）

《重大管理评论》学术委员会

（按姓氏拼音字母排序）

陈　晓	清华大学
贾建民	香港中文大学
李书行	台湾大学
李新春	中山大学
刘　星	重庆大学
陆正飞	北京大学
吕长江	复旦大学
陶志刚	香港大学
Gholamreza Torkzadeh	美国 内华达大学
王重鸣	浙江大学
吴世农	厦门大学
杨　勇	香港中文大学
姚树洁	英国 诺丁汉大学
张宗益	西南财经大学
赵曙明	南京大学
周冠男	台湾政治大学

重大管理评论
2015年第2辑

目 录

此为试读，需要完整PDF请访问：www.ertongbook.com

China Management Review
2015 Volume 2

CONTENTS

面向在线购物决策的消费者信息搜索收益预测方法

陈国青　张明月　王　昊*

摘　要：在线购物环境中，信息搜索是消费者做出购物决策前的重要步骤。尽管电商平台提供了许多决策支持工具，消费者仍然存在搜索“过多”或“过少”的问题。理论上看，当消费者信息搜索的收益增量小于或者等于相应的成本增量时，应该停止搜索。消费者本身不善于估计搜索收益增量是造成搜索量过多或过少的主要原因。针对此问题，本文将消费者的搜索场景形式化表示，并利用消费者对商品的评分作为其收益的度量，提出了一种有效预测搜索收益增量的商务智能方法（CPDM）。该方法在协同过滤预测评分的基础上，进一步考虑了评分的概率分布以及预测值的可信程度信息，以提升预测精度并体现决策中的个性化和不确定性特点。同时，在真实评分数据MovieLens上的实验结果也表明了该方法的有效性。

关键词：消费者信息搜索　收益增量协同过滤　置信度

* 陈国青，教育部人文社会科学重点研究基地清华大学现代管理研究中心，清华大学经济管理学院；张明月，清华大学经济管理学院；王昊，清华大学经济管理学院，中信证券股份有限公司另类投资业务线。基金项目：教育部人文社会科学重点研究基地基金（12JJD630001）；国家自然科学基金（71110107027）。

一、引 言

随着信息技术的发展与广泛应用，人们越来越多地利用信息技术所提供的便利服务来满足其自身在学习、工作、生活和休闲娱乐等方面的需求(Curme et al., 2014)。近年来我国网民总数剧增，互联网普及率迅速提高。根据中国互联网络信息中心（CNNIC）在 2015 年发布的《第 35 次中国互联网发展状况统计报告》，截至 2014 年 12 月，我国网民规模达 6.49 亿，人均周上网时长长达 26.1 小时，网络购物用户规模达到 3.61 亿，较 2013 年年底增加 5953 万人，增长率为 19.7%，我国网民使用网络购物的比例从 48.9%提升至 55.7%。

网络购物使得消费者通过在线浏览产品信息、发布购物订单即可足不出户完成整个购物环节，为用户提供了便捷的服务，并逐渐成为一种新型的休闲方式和商务体验。消费者与电商平台的交互产生了大量的行为记录，为学者研究网上的消费者行为提供了良好的契机，特别是在大数据背景下（Agarwal and Dhar, 2014; Chen et al., 2012），研究消费者网上行为成为市场营销学、经济学、心理学、信息科学等多个领域关注的重点问题。其中，搜索行为是帮助用户在海量数据中及时获取有用信息的一种重要方式。研究搜索行为不仅能够使消费者更方便和高效地选择并购买商品、为其提供决策支持，也有助于平台改进搜索服务质量，增强用户对网购的信任与依赖程度，有巨大的商业价值。

消费者信息搜索（Consumer Information Search, CIS）是指消费者在购物过程中为了做出满意的决策而收集相关信息，获取有用资料的过程（Solomon, 1999）。随着电子商务市场的日益膨胀，在网络上可以买到的产品种类和数量也在最近几年有着显著的增加（Brynjolfsson et al., 2011），这种状况一方面增大了用户买到满意产品的可能性，另一方面也由于相似产品太多而大大增加了用户浏览产品并做出购买决策的成本（Fleder and Hosanagar, 2009）。随时随地可以获得的海量商品使用户应接不暇，时间和精力的约束条件使其往往无法对全部商品进行比较和评估，因此消费者通常在做出购买决策前进行信息搜索。

一个普遍的搜索场景是顺序搜索（Palley and Kremer, 2014）。对同质化商品而言，消费者按照给定的顺序每次考察一个，经过比较后最终只选择使他最满意的一个商品。例如，用户在 Netflix 网站上搜索电影在周末晚上观看，由

于时间的冲突性，最终他只会选择一部电影。考察商品的过程包括阅读产品描述、从评论中获取有用的信息、查看商家信誉等，通过这一系列活动，消费者对考察的商品做出评估并决定是否要继续搜索。

顺序搜索会给消费者带来收益和成本。一方面，搜索收益包括同等情况下更低的价格、更好的质量、更个性化的设计甚至更高的购买信心等（Ratchford,1982;Hawkins et al., 2007）。除了低价格、高质量等一般性收益外，对于那些时刻关注最新商品动态的消费者，通过搜索获取更多信息本身就是一种享受（Solomon, 1999; Hawkins et al., 2007）。另一方面，在享受收益的同时，消费者也需要承受信息搜索带来的成本（Ratchford, 1982; Lynch and Ariely, 2000; Bakos, 1997; Kuksov, 2004; Sankaranarayanan and Sundararajan, 2010）。随着信息技术的发展，信息搜索效率逐渐提高，但信息搜索的成本仍然不可忽略（Browne et al., 2007; Klein and Ford, 2003; Pun jand Moore, 2009; Brynjolfsson et al., 2011）。并且由于个性化商品的数量逐渐增多，消费者需要耗费比以往更多的精力搜索和比较各式各样的商品。此外，商品搜索会占用消费者的其他时间，例如工作时间等。随着消费者工作效率的提高，即便很少的搜索时间也有可能会引起很大的机会成本。

从成本收益的角度来看，消费者希望增加信息搜索的净收益，即消费者信息搜索收益减去相应的信息搜索成本（Ratchford, 1982），但是同时面对海量的商品，消费者自己并不善于估计考察商品的效用或者收益增量（Zwick et al., 2003; Diehl and Zauberman, 2005; Diehl, 2005）。为了给消费者提供更好的服务，增强消费者的满意度和忠诚度，许多网商（如 Amazon, Netflix, Taobao 等）开发了各种各样的消费者决策支持系统，包括价格比较引擎、搜索引擎、推荐系统等，用来帮助消费者提高信息搜索效率，增加信息搜索净收益（Wang and Benbasat, 2009; Xiao and Benbasat, 2007; Häubl and Trifts, 2000; Alba et al., 1997）。然而，有研究表明，即便在决策支持系统的帮助下，实际中消费者也往往存在浏览“过多”或“过少”的现象（Diehl, 2005; Zwick et al., 2003）。

一般来说，考察商品的数量并不是越多越好，消费者需要及时地停止搜索，避免浪费时间。因为虽然浏览越多的产品可能使消费者获得越高的收益，但也会增加其成本。最理想的情况是消费者在净收益最大时停止浏览产品，即：当考察商品的收益增量小于或者等于相应的成本时，消费者应该停止考察商品（Stigler, 1961; Nelson, 1970; Weitzman, 1979; Ratchford, 1982），从而使网上产品选择过程的效用最大化。因此，需要提供相应的网上互动决策辅助功能的

浏览停止策略，帮助消费者判断是否已经获得最大效益，从而判断是否需要继续浏览和考察其他商品。尽管搜索收益和搜索成本对消费者决策都起着重要的作用，但每多考察一个商品占用的消费者的时间和精力基本是不变的，也就是搜索成本是一个相对稳定的变量。而搜索收益却会因为考察商品的价格、质量等属性的不同而不断变化。因此在假定搜索成本增量不变的情况下，预测消费者考察商品的收益增量是设计停止策略的重要环节，也是消费者网上产品选择过程的重要问题。

本文正是从预测消费者考察商品的收益增量出发，考虑单点预测值的分布信息和置信度因素，设计了一种有效的预测搜索收益增量的商务智能方法，以帮助消费者更好的解决搜索“过多”或“过少”的问题。文章第二部分的文献综述主要介绍了对消费者搜索收益的相关研究，包括定义和其他预测方法；第三部分介绍了置信度的度量方式，以及我们新提出的预测方法，并举例说明新方法如何预测搜索收益增量；第四部分用数据实验验证了方法的有效性；最后总结全文并指出未来的研究方向。

二、相关研究工作

消费者搜索信息的过程往往被定义成为顺序搜索问题(McCall, 1970; Nelson, 1970；Lippman and McCall, 1976；Ratchford and Srinivasan, 1993; Palley and Kremer, 2014)，即按照一定顺序每次浏览和考察一个商品，最终从中选择使他最满意的一个。线下顺序搜索的场景包括应届生对工作机会的选择、金融企业对投资方案的评估和选择等；线上对应的场景在体验型商品和搜索型商品中都有所体现，例如，在视频网站上根据用户口碑选择感兴趣的视频观看、摄影爱好者根据参数配置信息选择适合自己的相机进行购买等。由于商品的同质性，用户最终只会选择其中一个来购买（Zwick et al., 2003; Wright, 1975; Mogilner et al., 2012）。具体而言，消费者在每次考察完商品后都需要决定是否接受当前的产品、或召回以前考察过的其他产品、或继续搜索(Weitzman, 1979; Adam, 2001)。为了做出这样的决策，消费者需要在每多考察一个商品的搜索成本和未来可能获得的不确定收益之间进行平衡（Branco et al., 2012; Ghose et al., 2012）。从理论上来讲，当消费者信息搜索的收益增量小于或者等于相应的成本增量时，消费者应该停止搜索。

（一）消费者搜索收益增量的定义

一般来说，消费者信息搜索收益和收益增量难以定义和测量（Putrevu and Ratchford, 1997；Diehl, 2005; Reutskaja et al., 2011）。在实际搜索场景中，面对大量待考察的商品，通常情况下消费者会采取“考虑-选择”两阶段决策法（Moe, 2006；Zwick et al., 2003）。这一过程包括两个部分：消费者考察商品后挑选出比较满意的一部分，形成候选集合；接着经过仔细比较，从候选集中选出最终购买的商品。从这一过程来看，考察更多的商品可能使消费者获得效用更高的商品。因此，消费者信息搜索收益可以用最终购买商品的效用来衡量（Wang et al., 2011）。

假设消费者集合为 C，商品集为 S，消费者 $c \in C$ 按照给定的顺序考察商品并获得收益和产生成本，其在购买并使用商品 $s \in S$ 后获得的效用记作 u_{cs}。由于消费者最终只从可选集合中选择使他最满意的那一个商品购买，因此当他考察完某个集合的商品后获得的搜索收益应为商品效用的最大值（Diehl, 2005; Moe, 2006）。将消费者 c 已经考察的商品集合记做 S_c，剩余的商品集合为 $\tilde{S}_c$，则有 $S = S_c \cup \tilde{S}_c$。考察完商品集合 S_c 后获得的收益为：$\max_{s \in S_c}(u_{cs})$，若该消费者继续考察剩余的商品集合 $\tilde{S}_c$，其获得的信息搜索收益增量可以表示为：

$$B\left(c, \tilde{S}_c \mid S_c\right) = B\left(c, S_c \cup \tilde{S}_c\right) - B\left(c, S_c\right) = \max_{s \in S_c \cup \tilde{S}_c}\left(u_{cs}\right) - \max_{s \in S_c}\left(u_{cs}\right) \tag{1}$$

通常，消费者的效用难以度量（Samuelson and Nordhaus, 2001），但随着信息技术的发展，买家对商品的评分可以作为效用的一种评测标准。具体来说，在网上购物结束后，卖家或者商城会要求消费者给所购买的商品评分。例如，在观看电影之后，消费者会对所看的电影给出 1～5 分的评分（Adomavicius and Tuzhilin, 2005）。这样的评分反映了商品对消费者的效用，进而可以用来测量收益增量（Bhattacharjee et al., 2006）。因此，根据消费者对商品的评分，收益增量可以定义为：

$$B\left(c, \tilde{S}_c \mid S_c\right) = \max_{s \in S_c \cup \tilde{S}_c}\left(r_{cs}\right) - \max_{s \in S_c}\left(r_{cs}\right) \tag{2}$$

其中 r_{cs} 为消费者 c 如果购买并使用商品 s 后对商品 s 的评分，$r_{cs} \in R$，R 为评分集合。

（二）相关预测方法

消费者只有在购买并使用产品后才会给出代表自己偏好的评分，因此在搜索和考察商品的过程中，消费者对于商品的评分都是未知的，这正是预测的难点。

此为试读，需要完整PDF请访问：www.ertongbook.com

在推荐系统相关研究中，协同过滤（Collaborative filtering, CF）是最主要的一类评分预测方法，Amazon, Netflix 以及其他一些大型电子商务公司的推荐系统都采用此种类型方法来预测消费者对商品的评分（Koren, 2010; Pathak et al., 2010）。CF 方法根据“相似的消费者具有相似偏好”的假设，可以利用历史数据预测消费者未来的评分，记作 r_{cs}^{CF}。因此，CF 可以作为预测消费者搜索收益增量的一种方法，其预测结果为：

$$E^{CF}\left(B\left(c,\tilde{S}_c|S_c\right)\right)=\max_{s\in S_c\cup\tilde{S}_c}\left(r_{cs}^{CF}\right)-\max_{s\in S_c}\left(r_{cs}^{CF}\right) \tag{3}$$

从式（3）中可以看出，CF 预测搜索收益的关键点在于准确地预测每个单点的评分值。具体的预测方法可进一步细分为启发式方法（memory-based）和基于模型（model-based）的方法（Adomavicius and Tuzhilin, 2005）。具体算法包括基于用户的最近邻方法、基于产品的最近邻方法、矩阵分解、神经网络、贝叶斯模型（Hofmann, 2003; Koren, 2010; Chien and George 1999; Sarwar et al., 2001）等。

此外，许多研究还提出“评分的预测值”存在置信度或可靠性的问题（Hernando et al., 2013; Mazurowski, 2013），即每个预测结果的可信程度是不一样的。从消费者角度来看，“置信度”（confidence）可以理解成多大程度上该预测值是准确的。预测值的置信度取决于该预测结果的计算过程，这意味着即使两个预测值是相同的，它们的置信程度也可能是不同的。例如，最近邻方法中通过计算用户之间的相似度为目标消费者找到“邻居用户”，并基于邻居用户的历史评分数据预测目标消费者的评分。因此，直观来看，邻居用户的数量越多，对某个商品的评价越一致，则该预测值越准确，对应的置信度越高。

三、新的预测方法

利用 CF 方法预测消费者的个性化评分可以作为估计搜索收益增量的一种方法，然而其缺陷在于只使用了单点的预测信息，而没有考虑评分的概率分布信息。其次，CF 没有具体分析每个预测值的计算过程，而认为每个预测值的可信程度是一样的，导致对搜索收益增量的估计不准。本节在传统协同过滤方法的基础上，考虑了以上两个方面的不确定性信息，即（1）评分的概率分布信息，（2）预测评分单点值的可信程度信息，最终形成新的预测消费者搜索

收益增量的方法，即考虑置信度的个性化概率分布方法（Confidence-based Personalized Distribution Method, CPDM）。

（一）概率分布信息

根据定义，消费者 c 考察完集合 S_c 的所有商品后获得的收益是 $\max_{s\in S_c}(r_{cs})$。在经济学和市场营销学中，不同商品对相同或者不同消费者的效用被看作是独立的随机变量（Zwick et al., 2003），而此处消费者的评分 r_{cs} 可以看作是商品 s 给用户 c 带来的效用，因此有：

$$\max_{s\in S_c}(r_{cs})=\sum_{r\in R}\Pr\left(\max_{s\in S_c}(r_{cs})\geq r\right)=\sum_{r\in R}\left(1-\prod_{s\in S_c}\Pr(r_{cs}<r)\right) \tag{4}$$

由公式（4）可知，求解 $\max_{s\in S_c}(r_{cs})$ 的关键在于得到概率分布 $\Pr(r_{cs}<r)$。若对于消费者的真实打分 r_{cs} 没有任何先验知识，则有：

$$\Pr(r_{cs}<r)=\frac{r-1}{|R|},\forall s\in S_c \tag{5}$$

公式（5）对概率分布的求解方法无法把不同的商品区分开来，即认为任何一个商品给消费者 c 带来的效用都是一样的，这显然与真实场景是不一致的。

由于协同过滤方法可以得到消费者 c 对产品 s 的评分预测值，即 r_{cs}^{CF}，我们便可以根据历史数据计算 r_{cs} 的后验概率分布 $\Pr(r_{cs}<r|r_{cs}^{CF})$。因此，公式（4）可进一步细化为：

$$\max_{s\in S_c}(r_{cs})=\sum_{r\in R}\left(1-\prod_{s\in S_c}\Pr\left(r_{cs}<r\middle|r_{cs}^{CF}\right)\right) \tag{6}$$

正如文章第二部分所述，协同过滤方法包括多种具体的算法，如启发式方法和基于模型的方法等。因此，在公式（6）中，给定一个具体的评分预测方法 CF 及其预测评分 r_{cs}^{CF}，我们便可以得到一个相应的搜索收益预测结果 $\max_{s\in S_c}(r_{cs})$。

（二）预测值的可信程度信息

如前所述，在计算评分的后验概率分布时，评分预测值 r_{cs}^{CF} 是重要的输入数据，而根据 CF 得到的评分预测值 r_{cs}^{CF} 是具有不确定性的，这在公式（6）中没有体现。由于 CF 在预测过程中依据的数据的规模、质量不同而使得每个预测值的可信度不同，这一指标可以用置信度来衡量。一般来说，置信度被定义成一个标量，和每个预测值一一对应，即可以记作 $\left(r_{cs}^{CF},\ conf\left(r_{cs}^{CF}\right)\right)$。

在经典的协同过滤方法中（如最近邻方法或者矩阵分解方法），预测值的计算依赖于三种类型的数据：与消费者相关的数据、与商品相关的数据、与历史评分相关的数据。因此，每个预测值的置信度也可以通过分析这三种类型的数据而得到。马祖洛斯基（Mazurowski，2013）分别从这三种类型数据的角度

出发，共提出了六种计算置信度的方法，可以直接应用到我们的方法中。这里我们采用其中一种方法来得到置信度信息，即 *Support for Items*。具体地，预测值 r_{cs}^{CF} 的“置信度”定义为商品 s 在历史数据中被评分的数量，即：

$$conf\left(r_{cs}\right)=\left|I_s\right| \tag{7}$$

其中 I_s 表示所有包含商品 s 的评分构成的集合。

直观来看，商品 s 被评价的数量越多，则基于此得到的预测结果越准确，所以用该种方法计算得到的 $conf(r_{cs}^{CF})$ 与 r_{cs}^{CF} 的准确率呈正向关系。表 1 展示了用该种方法计算得到的置信度水平与预测准确率之间的关系：低置信度对应低预测准确率，高置信度对应高预测准确率。从而验证了 *Support for Items* 方法在计算置信度方面的有效性。

表 1　置信度水平与预测准确率的关系

数据子集名称	置信度水平	*RMSE*	*MAE*
lowConfSets	低	0.994	0.785
ranConfSets	中	0.906	0.711
highConfSets	高	0.886	0.683

注：RMSE指均方根误差（root mean square error），MAE指平均绝对误差（mean absolute error），是衡量预测准确率的经典指标。

此外，由于用该种方法计算得到的“置信度”是连续值，我们需要进一步离散化。这里采用十分位点离散化方法，即将原始数据集合基于十分位数划分成 10 个相等大小的子集。例如，假设原始集合中共包含 1000 个置信度的值，需要离散化成 10 个不同置信度水平。则基于十分位点划分后，第一个子集（对应置信度水平为 1）包括 100 个置信度取值最低的点，第十个子集（对应置信度水平为 10）包括 100 个置信度取值最高的点，以此类推。

（三）考虑置信度的个性化概率分布方法

通过使用上面介绍的置信度计算和离散化方法，便可以得到每个预测值对应的置信度水平，即 $\left(r_{cs}^{CF}, conf\left(r_{cs}^{CF}\right)\right)$，其中 $conf\left(r_{cs}^{CF}\right)\in\{1,2,\cdots,10\}$。因此，在求解消费者搜索收益时，公式（4）中的概率分布 $\Pr\left(r_{cs}<r\right)$ 在综合考虑预测值及其对应的置信度水平后，利用历史数据求得 $\Pr\left(r_{cs}<r\,|\,r_{cs}^{CF},\ conf(r_{cs}^{CF})\right)$。从而得到相应的预测消费者搜索收益增量的方法，即考虑置信度的个性化概率分布方法（Confidence-based Personalized Distribution Method, CPDM）：

$$E^{CPDM}\left(B\left(c,\tilde{S}_c|S_c\right)\right)$$
$$=\sum_{r\in R}\left(1-\prod_{s\in S_c\cup\bar{S}_c}\Pr\left(r_{cs}<r\left|r_{cs}^{CF},conf\left(r_{cs}^{CF}\right)\right.\right)\right)-\sum_{r\in R}\left(1-\prod_{s\in S_c}\Pr\left(r_{cs}<r\left|r_{cs}^{CF},conf\left(r_{cs}^{CF}\right)\right.\right)\right)$$
$$=\sum_{r\in R}\left(\prod_{s\in S_c}\Pr\left(r_{cs}<r\left|r_{cs}^{CF},conf\left(r_{cs}^{CF}\right)\right.\right)-\prod_{s\in S_c\cup\bar{S}_c}\Pr\left(r_{cs}<r\left|r_{cs}^{CF},conf\left(r_{cs}^{CF}\right)\right.\right)\right) \quad (8)$$

从公式（8）可以看出，CPDM 方法相较于 CF 的优势在于考虑了两方面的信息：概率分布与预测值的置信度。此外，为了验证新方法中考虑置信度信息的必要性，并具体分析置信度信息带来的效果提升程度，这里我们将不考虑置信度时的方法记做 PDM（Personalized Distribution Method，个性化概率分布方法），此时搜索收益的预测结果为：

$$E^{PDM}\left(B\left(c,\tilde{S}_c|S_c\right)\right)=\sum_{r\in R}\left(\prod_{s\in S_c}\Pr\left(r_{cs}<r\left|r_{cs}^{CF}\right.\right)-\prod_{s\in S_c\cup\bar{S}_c}\Pr\left(r_{cs}<r\left|r_{cs}^{CF}\right.\right)\right) \quad (9)$$

值得指出的是，PDM 可以看作是 CPDM 方法的一种特例，即在 CF 的基础上只考虑了概率分布信息，而忽略 CF 预测过程的不确定性，认为所有的预测值 r_{cs}^{CF} 的可信程度是一样的，例如 $conf(r_{cs}^{CF})=10,\forall s\in S,c\in C$。同样地，在后续的应用举例和真实数据实验部分，为了验证置信度信息所起的作用，我们分别对 CF、PDM 和 CPDM 的表现进行分析比较。

（四）应用举例

下面我们通过一个例子对 CF，PDM，CPDM 方法做出简单的说明。假设系统中有 3 种不同的商品记作 (s_1,s_2,s_3) 和 5 个不同的消费者 (c_1,c_2,c_3,c_4,c_5)，表 2 列出了这 5 个消费者对 3 种商品的真实评分和预测得分。其中，预测得分是根据某种 CF 算法得出，而真实评分只有在消费者购买并使用该商品后才会得出，在浏览过程中是不可见的。

表 2　系统中已有评分历史数据

商品		c_1	c_2	c_3	c_4	c_5
s_1	预测得分 r_{cs}^{CF}	5	5	5	5	4
	真实评分 r_{cs}	4	5	4	5	5
s_2	预测得分 r_{cs}^{CF}	4	1	2	3	1
	真实评分 r_{cs}	5	1	2	3	2
s_3	预测得分 r_{cs}^{CF}	4	4	3	5	3
	真实评分 r_{cs}	4	4	4	5	2

根据表 2 中的历史评分数据，可以计算得到 PDM 和 CPDM 方法中的概率

分布，从而进一步得到预测的消费者搜索收益增量。考虑如下场景：消费者c_1在考察完商品s_1后，希望知道继续考察s_2和s_3的收益增量是多少。下面我们分别用CF，PDM和CPDM来预测其收益增量。

（1）消费者c_1的真实搜索收益增量为：5-4=1。

（2）CF方法的预测结果：

$$E^{CF}\left(B\left(c,s_2,s_3\middle|s_1\right)\right)=\max_{s\in s_1\cup s_2\cup s_3}\left(r_{cs}^{CF}\right)-\max_{s\in s_1}\left(r_{cs}^{CF}\right)=5-5=0$$

（3）PDM方法的预测结果：根据表2中的历史评分数据计算得到的概率分布$\mathrm{P}(r_{cs}<r|r_{cs}^{CF})$如表3所示，因此有：

$$E^{PDM}\left(B\left(c,s_2,s_3\middle|s_1\right)\right)=\left(5-0.4\times0.5\times0.5\right)-\left(5-0.4\right)=0.3$$

表3 PDM方法中的概率分布$\mathrm{P}(r_{cs}<r|r_{cs}^{CF})$

r	1	2	3	4	5
$r_{cs}^{CF}=1$	0	0.5	1	1	1
$r_{cs}^{CF}=2$	0	0	1	1	1
$r_{cs}^{CF}=3$	0	0	1/3	2/3	1
$r_{cs}^{CF}=4$	0	0	0	0	0.5
$r_{cs}^{CF}=5$	0	0	0	0	0.4

（4）CPDM方法的预测结果：由于本例中历史数据过少，并不知道商品s_1,s_2,s_3的总体被评分数量，因此无法计算每个预测值的具体置信度水平。这里我们可以认为每个预测值的置信度水平均不同，即$|I_{S_1}|\neq|I_{S_2}|\neq|I_{S_3}|$，并分别记作$conf_1$，$conf_2$，$conf_3$。概率分布$\left(\mathrm{P}(r_{cs}<r\,|\,r_{cs}^{CF},conf(r_{cs}^{CF}))\right)$可以由该商品下的评分数据得到，如表4所示。因此有预测结果：

$$E^{CPDM}\left(B\left(c,s_2,s_3\middle|s_1\right)\right)=5-\left(5-0.5\right)=0.5$$

表4 CPDM方法中的概率分布$\left(\mathrm{P}(r_{cs}<r\,|\,r_{cs}^{CF},conf(r_{cs}^{CF}))\right)$

r	1	2	3	4	5
$r_{cs}^{CF}=5,s=s_1,conf_1$	0	0	0	0	0.5
$r_{cs}^{CF}=4,s=s_2,conf_2$	0	0	0	0	0
$r_{cs}^{CF}=4,s=s_3,conf_3$	0	0	0	0	1

对三种方法预测结果的比较见表5，可看出PDM和CPDM均优于CF方法，同时，考虑置信度后（CPDM）比没有考虑置信度时（PDM）对搜索收益$B(c,s_2,s_3\,|\,s_1)$的预测结果更接近真实值，也说明了置信度在预测搜索收益增量

过程中的重要性。

表 5　三种方法对搜索收益增量的预测结果

	$\max\left(r_{cs}\middle\|s\in\{s_1,s_2,s_3\}\right)$	$\max\left(r_{cs}\middle\|s\in\{s_1\}\right)$	$B\left(c,s_2,s_3\middle\|s_1\right)$
真实值	5	4	1
CF	5	5	0
PDM	4.9	4.6	0.3
CPDM	5	4.5	0.5

四、实　验

正如前文所述，经典的推荐算法“协同过滤”可以作为一种预测消费者信息搜索收益增量的基本方法，和我们新提出的方法进行比较。因此本节利用真实评分数据和相关理论模拟出消费者信息搜索场景，然后将 CF 和 CPDM 方法应用在搜索场景中，估计其搜索收益，最后对所预测的结果进行分析和比较，以验证我们方法的有效性。

由于 CPDM 将协同过滤预测得分作为算法输入的一部分，因此每种具体的协同过滤算法都对应着一种 CPDM 算法。本部分实验选取了两种应用最广泛的协同过滤算法来分别作为启发式方法和基于模型方法的两个典型代表，即基于用户的最近邻算法（Nearest Neighborhood, NN）和矩阵分解算法（Matrix Factorization, MF）。具体地，最近邻算法中邻居的个数设置为 40，相似度计算方法采用的是皮尔森相关系数，矩阵分解算法中潜在特征个数设置为 10。

此外，为了验证置信度在预测搜索收益增量过程中所起的作用，这里还将 PDM 作为一种预测方法，和 CF、CPDM 的表现一起进行评估。因此，实验结果比较可以相应的分成两组：

表 6　实验分组情况

	协同过滤方法	个性化概率分布方法	考虑置信度的个性化概率分布方法
比较分析（1）	最近邻算法 CF_{NN}	PDM_{NN}	$CPDM_{NN}$
比较分析（2）	矩阵分解算法 CF_{MF}	PDM_{MF}	$CPDM_{MF}$

此为试读，需要完整PDF请访问：www.ertongbook.com