面向旅游电子商务的多目标复合评估及优化推荐算法①

戴启艳

(苏州信息职业技术学院 计算机科学与技术系, 苏州 215200)

摘 要: 推荐算法在电子商务系统中具有良好的应用前景, 受到了越来越多关注和重视, 逐渐成为了电子商务领域的研究热点. 在介绍目前主流的推荐算法的基础上, 结合电子商务实际应用需求, 提出了多目标复合评估及优化推荐算法. 并基于途牛旅游网数据, 与一般推荐算法进行比较, 验证了该算法的有效性, 从而为多目标复合评估及优化推荐系统的研究提供了新思路与新方法.

关键词: 旅游电子商务; 多目标复合评估及优化; 推荐系统

Multi-objective Composite Evaluation and Optimization of Recommendation Algorithm for Tourism Electronic Commerce

DAI Qi-Yan

(Department of Computer Science and Technology, Suzhou Information Career Technical College, Suzhou 215200, China)

Abstract: Recommendation algorithm has a good application prospect in e-commerce systems, has been paid more and more attention and recognition, and has become a research hotspot in the field of electronic commerce. Based on the introduction of the current mainstream recommendation algorithm, in combination with the practical application requirements of e-commerce, multi-objective composite evaluation and optimization of recommendation algorithm is proposed. Based on the tuniu.com data, comparing with other recommendation algorithm, this paper verifies the effectiveness of the algorithm, and provides a new way and method of research of composite recommendation system.

Key words: tourism electronic commerce; multi-objective composite evaluation and optimization; recommendation system

随着互联网与信息技术的发展, 电子商务在人们生活、工作、学习及消费等多个领域被广泛应用. 博客、微博、微信等社会网络的新型信息发布方式的不断出现, 以及物联网、移动互联网、云计算等新型新兴信息技术的飞速发展, 使得网络信息量呈指数级增长, 并呈现出多源性、复杂性、数据结构不统一、信息量大等特点. 因而, 如何在海量复杂的信息中快速准确地获取到客户所需的商品, 进行精准化营销成为电子商务企业在市场上的竞争优势, 在这样的背景下, 电子商务推荐系统应运而生.

电子商务推荐系统以客户曾经浏览的历史数据、购买行为和记录、客户所在城市、网站最热卖商品等数据信息建模, 并进行分析, 获取关键数据, 最终将客户可能感兴趣又未曾浏览的商品信息个性化地推荐给客户, 从而使电子商务企业挖掘潜在客户, 扩大企业产品市场, 同时也为客户带来更加人性化的服务. 无论是对电子商务企业还是客户来讲, 推荐系统都具有巨大的发展潜质和应用价值. 目前推荐系统已经被广泛地应用于电子商务、电影、视频网站、社交网络、个性化音乐电台、个性化阅读等. 随着商业发展获利的需求越来越大, 电子商务推荐系统发展迅速, 目前国内主流电子商务平台都已具备推荐功能, 诸如淘宝网、京东商城、苏宁易购、易迅网等.

1 推荐系统算法概述

自 1997 年“推荐系统”概念诞生以来, 推荐系统就成为了电子商务研究应用领域的热点, 在发展过程中, 其定义有很多, 目前被广泛认可和采用的是 Resnick 和 Varian 给出的描述: “它是以电子商务网站为平台, 为消费者提供商品的信息和建议, 协助他们决定应该

① 收稿时间:2015-12-27;收到修改稿时间:2016-03-22 [doi:10.15888/j.cnki.csa.005403]

购买什么产品, 模拟推销人员协助消费者完成购买过程”[1].

进入 21 世纪, 推荐系统迅速渗透到系统电子商务应用中, 亚马逊(Amazon.com)将电子商务推荐系统成功应用于商业活动中, 在亚马逊购买书籍的客户同时能享受到亚马逊免费推荐可能感兴趣书籍的贴心服务. 推荐系统的使用极大地提高了亚马逊的营业额. 据统计, 推荐系统对亚马逊营业额的贡献率在 20%以上.

推荐算法是推荐系统中最核心、关键的部分, 算法的优劣决定了推荐系统的应用效果和性能. 现阶段对于推荐算法的研究主要集中于如何提高推荐效率和推荐效果两个方面. 目前主流的推荐系统算法主要有三种类型: 基于协同过滤的推荐算法、基于内容的推荐算法、基于社交网络的推荐算法.

1.1 基于协同过滤的推荐算法

基于协同过滤的推荐算法是最早被研究也是应用最为广泛的算法, 主要基于用户行为展开研究. 其基本思想是对目标用户进行相似性计算, 依据相似用户群的兴趣爱好对目标用户生成预测评价, 从而将计算得出的推荐结果推送给目标用户. 根据相似度比较对象的不同, 基于协同过滤的推荐算法又可分为基于用户的推荐和基于项目的推荐.

基于用户的推荐算法, 其基本思想是: 首先利用最相邻计算寻找目标用户的最近邻居, 然后根据最近邻居产生推荐结果. 该算法最核心的问题是如何寻找目标用户的最近邻居, 而这一问题是通过计算目标用户与其他用户之间的相似度来解决, 相似度越高, 则两个用户越相近.

基于项目的推荐算法, 其核心是: 根据大部分用户对项目的评分来进行预测和推荐. 此算法通过计算项目之间的相似性, 以寻找目标项目的最近邻居, 根据当前用户对最近邻居项目的评分来预测目标项目的评分, 并选择评分最高的若干项目推荐给目标用户[2].

协同过滤推荐算法的优点有: ① 适用于复杂的非结构化对象的推荐, 依据用户行为数据挖掘用户新的兴趣爱好. ② 以用户为中心进行推荐, 可以挖掘到内容上不完全相似的对象, 用户看到推荐结果具有惊喜感. ③ 随着用户规模的不断增大, 算法的推荐性能逐步上升. 但基于协同过滤的推荐算法也存在不足之处: ①若是新用户, 推荐系统中没有此用户的历史信息, 则推荐效果较差. ②若是老用户加入新项目后, 协同过滤算法也无法对此项目进行计算评分而将其推荐给其他用户. ③由于用户和推荐项目数量庞大, 部分项目无法得到用户的评分而不被推荐的现象无法避免.

1.2 基于内容的推荐算法

基于内容的推荐算法根据用户历史兴趣爱好数据建立用户模型, 提取推荐对象的内容特征, 与用户模型中的用户兴趣爱好进行比较, 将具有最大相似性的推荐对象推送给用户[3]. 这种算法无需借助物品评分系统, 具有很强的独立性, 可以接收用户的反馈信息进行学习和改进来提高推荐质量.

基于内容的推荐算法简单有效, 推荐结果直观且容易理解, 无需借助复杂抽象的专业领域知识就可完成推荐, 能够推荐新项目或是当前不十分流行的项目. 但是, 这种推荐算法容易受推荐对象特征的提取能力影响, 推荐质量会因部分难以提取特征的项目而降低, 从而降低用户的满意度[4]. 由于需要在推荐对象特征与兴趣偏好相匹配时才能获得推荐, 所以基于内容的推荐算法推荐很难发现推荐对象新的兴趣爱好.

1.3 基于社交网络的推荐算法

随着社交网络的流行, 基于社交网络的营销开始凸显出巨大的商业价值和商业潜力, 诸如基于微信平台的微商. 基于社交网络的推荐算法是以社交网络中人与人之间的关系为基础对象, 使用协同过滤推荐算法进行推荐的一种方法. 基于社交网络的推荐算法可以分为基于邻域的社交网络推荐和基于网络结构的社交网络推荐[5,6].

基于邻域的推荐算法: 假设给定一个社交网络及其用户行为数据集, 社交网络列出了用户之间的好友关系, 用户行为数据集给出了不同用户的历史行为和兴趣数据, 在这种情况, 给用户推荐好友喜欢的物品集合.

基于网络结构的推荐算法: 社交网站中存在两种关系, 一种是用户对物品的兴趣关系, 一种是用户之间的社交网络关系. 通过图模型来表示这两种关系, 便于对用户进行个性化推荐.

随着互联网数据量的爆发性增长以及数据复杂性的提高, 推荐算法也将趋向复合推荐方向发展. 本文主要基于旅游电子商务对复合推荐算法进行系统研究.

2 评估算法概述

评估算法就是根据一系列的准则判断推荐算法的

准确性, 随着越来越多的学者提出各种推荐算法, 那么急切需要能评估推荐结果的算法. 通常采用两种常用的评价指标评估推荐算法的结果, 即预测偏差(prediction shift)和命中率(hit radio).

1) 预测偏差

当直观评估推荐算法的推荐结果对用户选择的影响程度时, 通常会选择预测偏差. 预测偏差是指推荐算法与用户选择之间的变化. 预测偏差的平均值是所有项目和所有用户的加权平均. 当然预测偏差的平均值越大, 说明推荐效果越好. 反之, 表明该推荐算法的效果不好.

用 I 表示推荐项目的集合, 用 U 表示用户的集合, 用 $\Delta u,i$ 表示用户 u 对项目 i 的预测值, 即 $\Delta u,i$ 的公式为:

$$\Delta_{u,i} = q_{u,i} - p_{u,i} \tag{1}$$

其中, q 表示推荐的预测值, p 则是用户最后的选择结果.

2) 命中率

当项目 i 的预测偏差较高时, 项目 i 不被用户推荐的概率较高, 因为它会被其他项目预测偏差影响. 因此, 本文引入了一个新的评价标准: 命中率.

命中率是指项目被用户选择前 N 个被推荐的项目的概率值. R_u 表示在推荐之后, 用户主动推荐的项目的集合, 用 $H_{u,i}$ 表示项目 i 是否进入 *top-N* 的推荐列表序列. 如果 $i \in R_u$, 则 $H_{u,i}=1$, 否则 $H_{u,i}=0$. 对于项目 i 的命中率计算如下:

$$HitRatio_i = \frac{\sum_{u \in U} H_{u,i}}{|U|} \tag{2}$$

同样, 平均命中率是指所有项目命中率的均值, 如下:

$$\overline{HitRatio_i} = \frac{\sum_{i \in I} HitRatio_i}{I} \tag{3}$$

3 多目标复合评估及优化推荐算法提出

传统的推荐系统提供给用户一系 $t \in N$ 列的推荐(recommendation), 通常每个推荐只包含一项数据(item). 例如旅游推荐时, 只推荐一个景点或一条路线. 本文提出的复合推荐(composite recommendations)算法, 每个推荐包含多项数据. 例如对于同一个旅游推荐, 可以包含景点, 美食, 路线等多项数据, 这样可以在旅游者的预算(时间和金钱)范围内, 结合兴趣点推荐出对旅游者而言最有价值的旅游计划. 复合推荐结构如图 1 所示.

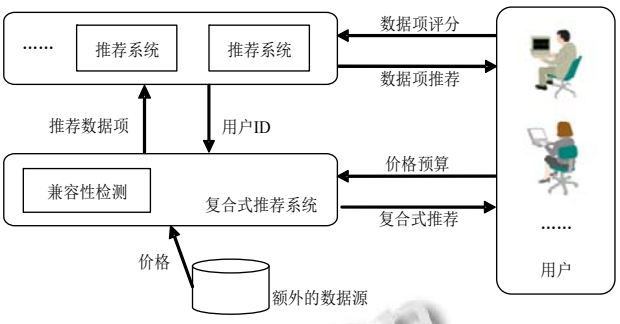

图 1　复合推荐系统结构图

3.1 问题描述

本文首先定义 N 个数据项 *items*, 每个数据项包含评分和价格, 用户集 $u \in U$. 假定 $v_{u(t)}$ 是对用户 u 数据项 t 的评分, $c(t)$ 是数据项 t 的价格. 那么对于一部分数据项 $R \subset N$, 本文定义: $C(R)=\sum_{t \in R} c(t)$ 和 $V(R)=\sum_{t \in R} v(t)$. 定义价格预算 B, 如果 $c(P)<B$, $P \subset N$, 则称 P 是可执行的.

假定一个复合式推荐的实例 I, 价格预算 B, 整数 K, *top-k* 背包问题 $P_1,\ldots,P_k$, 如果 P_i 是可执行的, 那么 $P_1,\ldots,P_k$ 就有最高的评分, 即 $v(P) \le v(P_i)$, $P \notin \{P_1,\ldots,P_K\}$.

3.2 推荐算法过程

3.2.1 实例优化算法

假设给复合式推荐一个实例 I, 定义了 BG 为背景成本信息, $S = \{t_1,\ldots,t_n\}$ 为目前可以预测到的数据项. $\bar{v}$ 为首个数据项的评分, 评分值在这个推荐包里是递减的, 那么对于每个 $i \in \{1,\ldots,n\}$ 都有 $v \in \{1,\ldots,n \cdot \bar{v}\}$. 定义 $SS_{i,v}$ 为 $\{t_1,\ldots,t_i\}$ 的子集, 并且它的总评分是 v 而总价格是最小的, 又定义了 $C(i, v)$ 为 $SS_{i,v}$ 的价格(如果 $SS_{i,v}$ 不存在, 那么 $C(i, v)=\infty$). 那么拟多项式时间算法[7,8]能通过递归方法计算所有的 $C(i, v)$ 来获得最优背包问题的解(4). 然后选择 $SS_{n,v}$ 中总的价格要小于或者等于预算 B, 即 $\max\{v|C(n, v) \le B\}$.

$$C(i+1,v) = \begin{cases} \min\{C(i,v), c(t_{i+1}) + C\{i, v - v(t_{i+1})\}\} & if\ v(t_{i+1}) \le v, e \\ C(i,v) & otherwise \end{cases} \tag{4}$$

定义背景成本信息 $BG = c_{min}$, 即为所以数据项的价格最小值, 上界 V^* 意味着在预算范围内的最大评分. $V_{\min} = \min_{t \in S} v(t)$, 为了获得最理想的价值, 即在价格小于预算价格, 评价最高, 本文通过 MaxValBound 算法得到上界 V^*. 初始化 V^*, 循环迭代 $C(i, v)$ 直到 V^* 最大, 伪代码如下:

算法 1 $MaxValBound(S,C,B,BG)$

$V^* \left|\frac{B}{C_{\min}}\right| \times v_{\min}$

for $v \in \{1,\ldots,n \bullet \bar{v}\}$ *do*

if $C(n,v) < B$

$V^* = \max\left\{V^*, v + \left|\frac{B - C(n,v)}{c_{\min}}\right| * v_{\min}\right\}$

return V^*

假设上界 V^*是最优解, 接下来需要求 *top-1* 的复合式推荐, 即实例优化算法. 每个数据项都由数据源依次取出, 当接受到一个新的数据项, 我们用拟多项式时间算法 InsOpt-CR 取得一个优化的解 R_0, 然后通过 MaxValBound 算法得到上界值 V^*, 如果 $v(R^1) \geq \frac{1}{a} \times V^*$, 算法停止计算. 算法伪代码如下:

算法 2 $InsOpt-CR(N,B,BG)$

$S \leftarrow$ *An empty buffer*

While TRUE do

$t \leftarrow N.getNext()$

$S.Insert(t)$

$(R^0, C) \leftarrow OptimalKP(SB)$

$V^* = MaxValBound(S,C,B,BG)$

$fv(R^0) \geq \frac{1}{2} \times V^*$

return R^1

上述为 *top-1* 的复合式推荐, 关于 *top-k* 的复合式推荐, 首先假定 *RI* 是一系列的可行的推荐包, $R^I = \{R \mid R \subseteq N \wedge c(R) \leq B\}$. 根据[9,10], 定义$\alpha$近似求解, $R_k = \min(k, |R^I|)$, 那么 $R \in R_k, R' \in R^I \setminus R_k$, 则 $v(R) \geq \frac{1}{\alpha} \times v(R')$.

3.2.2 贪婪算法

虽然实例优化算法能够得到, 但是需要精确的算法解决背包问题, 这样计算量太大. 相对于用精确的算法计算背包问题, 本文采用一种简单的贪婪启发式算法计算一个高质量的包[11]. 相对于例优化算法, 本文采用贪婪启发式算法 GreedyKP 代替了拟多项式时间算法 OptimalKP 找到一个高质量的 *RG*, 并且采用启发式算法 MaxHeuristicValBound 算法计算上界值, 伪代码如下.

算法 3 $MaxHeuristicValBound(S,B,BG)$

$r \leftarrow \frac{v_{\min}}{c_{\min}}$

$Sort\ \ S = \{t_1,\ldots,t_n\}\ \ by\ \ value/\cos t\ \ ratio$

$m = \max\{\{m \mid \frac{v(t_m)}{c(t_n)}\} \geq r \wedge c(R_m) \leq B\}$

$R_m = \{t_1,\ldots,t_m\}$

if $m == n$

$V^* = v(R_m) + r^*(B - c(R_m))$

Else

$V^* = v(R_m) + \max\{r, \frac{v_{m+1}}{c_{m+1}}\}(B - c(R_m))$

return V^*

3.2.3 实验分析

基于途牛旅游网数据, 本文选取五组数据集进行实验, 表 1 展示了 *top*-5 复合式推荐的推荐质量与推荐效果. 本文采用每个推荐包总的评分(*SUM*)和每个推荐包的平均数据项的评分(*AVG*)作为测量方法.

表 1 不同复合式推荐算法的质量比较

	数据组 1		数据组 2		数据组 3		数据组 4		数据组 5	
	SU	AV	SU	AV	SU	AV	SU	AV	SU	AV
Optimal	427	46.7	426	46.6	425	46.7	424	46.7	423	46.6
InsOpt-CR-Top-k	386	47.5	385	47.4	385	47.3	384	47.2	383	47.2
Greedy-CR-Top-k	384	47	381	47	380	46.8	379	46.7	379	46.7

表 1 已经证实了本文提出的算法能够得到 top-K 复合式推荐的最优评分. 并且, 从每个数据项的平均评分来看, 本文提出的算法通常能得到高质量的推荐包, 但是 Optional 算法只能提供低价格, 低评分的推荐包. 因此, 为更高效地提供高质量的推荐包, 本文通过放弃一些低质量的数据项.

为了更好地评价返回来的推荐包的质量, 本文采用 Normalized Discounted Cumulative Gain(NDCG) [12] 去测试 *top-K* 复合式推荐的返回结果. 假设 Optional 算法的返回结果是 $R_0=\{P_{10},\ldots,P_{K0}\}$, 我们的算法则是 $R_\alpha=\{P_{1\alpha},\ldots,P_{K\alpha}\}$. 改进的 NDCG 分数是指总的推荐包在不同位置的权重值, 以下是 NDCG 的定义[13]:

$$NDCG(R^0, R^\alpha) = \sum_{i=1}^{k} \frac{\log(1 + \frac{v(p_i^0) - v(p_i^\alpha)}{v(p_i^0)})}{\log(1+\alpha)} \quad (5)$$

改进的 NDCG 的理想分数是 0, *top-K* 复合式推荐返回的包几乎和 Optional 算法返回的一样. 改进的

NDCG 的最糟糕的分数是 $\sum_{i=1}^{k}\frac{\log 2}{\log(1+i)}$，返回的包的总评分为0. 图2展示了实例优化算法和贪婪算法返回的 *top-k* 包的途牛数据集的 NDCG 分数. 很明显, 虽然贪婪算法需要大量的计算时间, 但是比实例优化算法返回整体 *top-k* 包的质量的相似度更高.

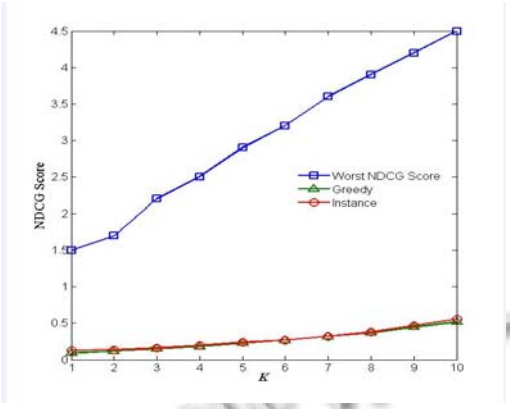

图2 贪婪算法和实例优化算法对途牛数据的 NDCG 分数

3.2.4 结论

图 3 展示了本文提出的算法在途牛数据集上的运行时间. 对于途牛的数据集, 贪婪算法在运行时间和价格效果上更好. 实例优化算法 Instance 有一个低的价格, 但是它的运算速度随着 K 增长地很快, 因为它需要解决很多背包问题实例, 以约束它的数据项.

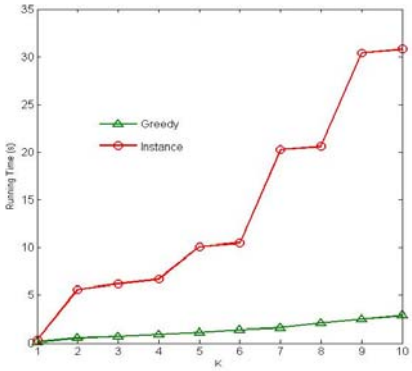

图3 贪婪算法和实例优化算法对途牛数据的运行时间

4 总结

本文以电子商务推荐系统应用为背景, 对大数据发展以及电子商务主流推荐系统算法做了介绍, 在此基础上, 以途牛网旅游数据为例, 提出了复合推荐算法. 与其他两种推荐系统算法进行比较, 证明了复合推荐算法的合理性与有效性, 为复合推荐系统的研究提供了新的研究思路.

参考文献

1 许海玲,吴潇,李晓东,阎保平.互联网推荐系统比较研究.软件学报,2009,20(2):350–362.

2 Bell RM, Koren Y. Improved neighborhood-based collaborative filtering. In: Berkhin P, Caruana R, Wu X D, eds. Proc. of the 13th International Conference on Knowledge Discovery and Data Mining. New York. ACM. 2007. 7–14.

3 黎明,徐德智.一种结合基于项目和用户的个性化推荐算法.小型微型计算机系统,2011,(4):611–613.

4 陈志敏,李志强.基于用户特征和项目属性的协同过滤推荐算法.计算机应用,2011,31(7):1748–1751.

5 Herlocker J, Konstan J, Tervin LG, Riedl J. Evaluating collaborative filtering recommender systems. ACM Trans. on Information Systems, 2004, 22(1): 50–53.

6 张光卫,李德毅,李鹏,康建初,陈桂生.基于云模型的协同过滤推荐算法.软件学报,2007,18(10):2403–2411.

7 Mobasher B, Burke R, Bhaumik R, Williams C. Towards trustworthy recommender systems: An analysis of attack models and algorithm robustness. ACM Trans. on Internet Technology, 2007, 7(4): 2301–2338.

8 Williams C, Bhaumik R, Burke R, Mobasher B. The impact of attack profile classification on the robustness of collaborative recommendation. Proc. of the 2006 Web KDD Workshop, held at ACM SIGKDD Conference on Data Mining and Knowledge Discovery (KDD'06). Philadelphia. ACM. 2006. 1–10.

9 Fagin R, Lotem A, Naor M. Optimal aggregation algorithms for middleware. Journal of Computer and System Sciences, 2003, 66(4): 614–656.

10 Nakagawa M, Mobasher B. A hybrid web personalization model based on site connectivity. Web KDD Workshop at the ACM SIGKKDD International Conference on Knowledge Discovery and Data Mining. Washington DC. ACM. 2003. 1–11.

11 张付志,张启凤.融合多系统用户信息的协同过滤算法.计算机工程,2009,21:258–260.

12 王卫平,刘颖.基于客户行为序列的推荐算法.计算机系统应用,2006,(9):35–38.

13 张恺,秦亮曦,宁朝波,李文阁.改进评价估计的混合推荐算法研究.微计算机信息,2010,26(12-3):193–194.