基于相关物品的电子商务智能推荐系统研究

童　启，刘　强，许赛华，胡益广
（湖南工业大学 计算机学院，湖南 株洲 412007）

【摘　要】以大数据时代为背景，改变传统推荐系统的设计思路，给出一个“相关物品”智能推荐系统体系结构。将用户即时购买需求和历史偏好相结合，提出一种改进的基于粗糙集的属性约简算法，用于数据预处理阶段提取用户实时需求商品类的特征，在线分析阶段采用引入兴趣域的聚类算法挖掘用户实时关注商品的相似商品集，离线用协同过滤推荐挖掘相关商品集，将在线部分与离线部分的相关商品集融合，按照点击率预估对集合进行排序，形成推荐，以解决推荐系统在实时性、扩展性、智能性和实时性与精准平衡性方面存在的问题。

【关键词】推荐系统；粗糙集；数据挖掘
【中图分类号】TP18；TP311.52【文献标识码】A　【文章编号】1674-0688(2019)12-0079-02

信息过载时代，智能推荐系统在电子商务领域的地位不断提高，它能最大化地体现电子商务网站的优势。电子商务智能推荐系统是通过一定方法帮助人们找到自己需要的物品，它涉及的技术很多，其核心是推荐算法。随着电子商务系统规模的扩大，目前电子商务推荐系统在准确率、多样性、实时性等方面存在很多问题。文章以物品—物品的角度，将多种推荐算法融合，构建面向电子商务的推荐系统。

1　推荐算法在应用中存在的问题

1.1　实时性和扩展性问题

大部分推荐算法是在小数量级上评估的。而随着电子商务的发展，大型购物网站是千万级顾客、百万级商品的大数据集。面对这样的大数据集，算法会遭受严峻的性能和计算量问题，达不到实时性、扩展性的要求。例如，采用传统协同过滤来产生推荐，最坏的情况是 O（$M\times N$）（其中 M 是用户数量，N 是物品数量），如果用户或物品数量级不大，算法的执行也接近于 O（$M+N$）。解决此问题的思路是可以根据用户实时关注的物品类关键属性挖掘相似物品集合，使算法的在线计算规模与用户数量无关，仅与用户实时购买需求的物品大类数量有关。

1.2　智能化问题

大部分的推荐系统仅抓住了用户在过去一段时间内的行为偏好，并没有对用户当前行为的短期意图进行考虑，而这些意图正是最能反映用户当前需求的一些信息。传统的推荐系统并未能很好地解决智能化推荐的问题。解决此问题的思路是将用户的历史偏好信息与在线的实时购买需求信息相结合并推荐给用户，力求给用户最需要的推荐，解决智能化问题。

1.3　实时性与推荐满意度平衡问题

根据用户在线行为的短期意图做推荐时，大部分算法需要处理大量数据不断迭代，执行时间较长，而执行快速的算法在用户满意度指标上又有欠缺，因此算法在实时性与满意度两个方面无法找到一个平衡点。解决此问题的思路是采用在线部分和离线部分，分别挖掘相似物品集合，然后按权重融合，再进行推荐。

2　相关物品思路的推荐引擎架构

架构包括离线处理和在线分析两个部分。离线处理部分根据用户的信息如年龄、性别、职业、用户的历史购物信息及历史物品的评价信息，采用协同过滤算法挖掘用户关注物品的相似物品集合。在线部分考虑用户实时性需求，在服务器日志中收集用户在线行为，如浏览的网页信息、收藏的物品信息，研究用户对物品的兴趣度，接着对感兴趣物品类信息进行数据预处理时，采用改进的基于粗糙集属性约简算法，提取用户实时关注的物品特征，如物品的款式、价格、颜色、购买人数、评价分等，抽取出特征后采用混合数据聚类标签算法挖掘用户实时购买需求的相似物品集合。两类集合按照权重进行融合，如果用户的浏览带无目的性或者已经完成购物，则以离线部分的相似物品集合为主，如果用户的浏览行为有明确的购物目标，则以在线部分的相似物品集合为主。最后用决策树算法对相似物品集合按照点击率预估进行排序，产生推荐。

3　推荐系统设计及算法研究

3.1　隐式获取数据

推荐算法分析设计的基础是用户行为数据，在线部分通过

【基金项目】湖南省教育厅科学研究项目“面向电子商务的相关物品智能推荐系统的研究”（项目编号：15C0399）。
【作者简介】童启，男，湖南工业大学计算机学院讲师，研究方向：计算机应用技术、数据分析。

隐式获取法，如获取用户浏览网页的行为和内容，包括浏览、查询的物品及某物品浏览的次数、频率、停留时间等，以及行为操作（收藏、加入购物车、下单等）。以上这些行为反映了用户对物品有不同兴趣，可以将用户兴趣与行为的联系用公式(1)表达。其中，c为消费者用户，i为物品，$A_b(c)$为用户c操作物品的行为集合，f_b为用户行为b所代表的兴趣权重，$p_{c,i}$为用户c对物品i的综合兴趣度。计算$p_{c,i}$得出用户在线状态感兴趣的物品集合。

$$p_{c,i_{c,i}\in A_b(c)}=\sum f_b \tag{1}$$

3.2 提取物品关键属性

对感兴趣物品的属性进行数据预处理。一是连续属性离散化，如商品的价格属性，可以按规则进行离散化，区分为高、中、低。二是对不完备信息系统进行补齐处理，采用粗糙集中数据不可分辨关系理论来处理或直接删除数据。数据挖掘的结果很大程度上依赖于预处理的效果。

属性约简是指在保证数据划分一致的前提下，通过去除不需要的属性，以简化大数据环境下的复杂程度，降低计算维度。因为粗糙集理论不需要先验信息，可以采用基于差别矩阵的属性约简算法提取用户感兴趣物品的关键属性。用等价类代替单个元素构造差别矩阵，将差别矩阵与属性区分力相融合，提高算法效率。

3.3 根据物品关键属性运用混合数据聚类标签算法挖掘相似物品集

基于“相关物品”推荐的思想，即在用户对物品i感兴趣时，给用户推荐和i相似的物品，在线挖掘考虑用户的实时购物需求。根据属性约简提取的物品特征属性对物品进行聚类。在聚类算法中引入物品特征属性兴趣域，采用融合兴趣域的混合数据聚类标签算法。根据物品特征属性的范围可以建立物品特征属性兴趣域D_j。在D_j中每个属性取值都有一个范围，相似物品数据集中每个属性与D_j中对应属性范围的关联用d_{ij}表示，当混合数据集的属性值在范围内，d_{ij}=1，否则d_{ij}=0。公式(2)中，数据集S_i与物品特征属性兴趣域D_j关联度的权值用C_{ij}表示。公式(3)中，f_{ij}为数据集关于兴趣域的隶属度，其中N为物品特征属性个数。f_{ij}取值范围是0到1，f_{ij}值越大，说明S_i归属D_j的集成度越高，反之，f_{ij}值越小，S_i归属D_j的集成度越低。

$$C_{ij}=\sum d_{ij} \tag{2}$$

$$f_{ij}=\frac{C_{ij}}{N} \tag{3}$$

采用K-prototypes算法对物品特征属性信息进行聚类处理，构建物品特征属性兴趣域。通过公式计算d_{ij}和f_{ij}，对混合数据集数值聚类标签输出。引入物品特征属性兴趣域提升了聚类结果对用户的针对性。此外，可以对关键属性设置感兴趣的权重，进一步提高推荐的满意度。在线分析部分主要是用于满足用户实时需求的推荐，由于受到物品关键属性特征提取能力的限制，所以很难满足推荐的新颖性和惊喜度。

3.4 在线物品集与离线物品集融合，排序产生推荐

离线部分采用基于用户的协同过滤算法产生相似物品推荐集。离线部分是在线部分的补充，为用户提供新的兴趣推荐。融合时在用户属性中加入一个表示行为状态的字段，如“1”表示无需求浏览，“2”表示购物中，“3”表示购物后3种状态值。根据状态值按不同的权重对在线物品集和离线物品集进行融合。

接下来是对融合的候选物品集进行排序。因为推荐系统的不同评测指标和用户的不同需求状态，物品与用户的相关度会不同，因此本文重点考虑用户的反馈，使用用户对推荐物品的点击率作为指标来对物品进行排序。点击率预估问题可以转化为传统的两类分类问题。考虑电子商务网站物品特征属性少的情况，使用决策树的算法，可以发挥决策树算法解决非线性问题的优势。

4 结论

文章提出“相关物品”思路的电子商务推荐系统架构。在线部分的计算仅需考虑用户实时关注的商品类别中商品数量，以解决大数据环境下推荐的实时性、扩展性问题；按照用户状态行为的属性值对两个商品集进行融合，使得系统能够同时满足用户的实时需求和对新物品的兴趣，以解决推荐的智能性问题；在线分析与离线推荐相结合，以解决实时性与精准性的平衡问题。需要进一步研究降低精简属性集对决策效能的影响和数据聚类标签算法中引入的兴趣域容忍度参数对推荐正确率的影响。

参 考 文 献

[1] 邓爱林．机器学习和数据挖掘在个性化推荐系统中的应用[J]．中国计算机学会通讯，2013(9)．

[2] 苏庆等．改进模糊划分聚类的协同过滤推荐算法[J]．计算机工程与应用，2019(5)．

[3] 雷曼等．基于标签权重的协同过滤推荐算法[J]．计算机应用，2019(3)．