

文章编号: 1000 - 5463 (2014) 05 - 0008 - 08

个性化推荐算法研究

陈洁敏, 汤庸*, 李建国, 蔡奕彬
(华南师范大学计算机学院, 广州 510631)

摘要:随着全球信息总量的爆炸式增长,信息超载问题无法避免且日趋严重化. 个性化推荐系统是当前解决信息过载问题的有效技术. 该文首先阐述了推荐系统概念定义及其三大组成模块,其次深入分析了个性化推荐算法,详细讨论了当前主流的四大类推荐算法:基于内容的推荐算法、协同过滤推荐算法、基于知识的推荐算法和混合的推荐算法,从多角度对各算法的优缺点进行对比,然后阐述了常用评价方法、评测指标及对测试标准进行分类,并且介绍了常用数据集,最后展望个性化推荐未来研究热点.

关键词:推荐系统; 推荐算法; 协同过滤; 个性化; 信息过载

中图分类号: TP391 **文献标志码:** A doi:10. 6054/j. jscnun. 2014. 06. 021

Survey of Personalized Recommendation Algorithms

Chen Jiemin, Tang Yong*, Li Jianguo, Cai Yibin
(School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: As the explosive growth of global information, the information overload is one of the most critical problems and is becoming more and more serious. Recommendation systems are one of the powerful ways to solve the problem. The definition of recommendation system is first introduced. A comparison study is conducted based on the four main recommendation algorithms: content-based recommendation, collaborative filtering recommendation, knowledge-based recommendation and hybrid recommendation. The evaluation methods, evaluation metrics and recommendation benchmarked datasets are also presented. At last the difficulties and future directions of recommendation systems are given.

Key words: recommendation systems; recommendation algorithms; collaborative filtering; personalization; information overload

在网络数据爆炸的年代,人们面对的是海量信息,例如亚马逊上面有数百万种独特的商品,Google Music 曲库有过千万首歌,Del. icio. us 上面有超过 10 亿的网页收藏,淘宝在线商品数已经超过了 8 亿件,新浪微博用户数及腾讯微信用户数均超过 5 亿. 用户在海量的信息中难以找到自己感兴趣的信息,这就是所谓的“信息过载 (information overload)”问题,搜索引擎和推荐系统是目前解决该问题的主要技术. 与搜索引擎相比,用户更青睐于使用个性化推荐系统,因为它能主动地从用户注册信息、用户浏览日志、历史评分记录和项目信息等方面进行分析,从

而挖掘用户的兴趣偏好和项目的特征,然后为用户实现个人感兴趣信息的私人定制,并根据用户需求和项目信息的变化及时调整推荐的内容和服务方式,实现“以用户为中心”的个性化服务. 个性化推荐系统具有良好的发展和应用前景,目前电子商务网站、社交网站、电影和视频网站、个性化音乐网络电台、个性化阅读和个性化广告等领域都使用了不同形式的推荐系统并获得巨大的效益. 本文阐述了推荐系统概念定义,对比各类推荐算法优点和缺点,总结了常用数据集和评测指标,最后提出未来面临的主要问题及可能的研究方向.

收稿日期: 2014 - 02 - 01

基金项目: 国家 863 项目 (2013AA01A212); 国家自然科学基金项目 (61370229, 61272067); 国家科技支撑计划项目 (2013BAH72B01); 广东省自然科学基金团队研究项目 (S2012030006242); 广东省重大科技专项计划项目 (2012A080104019); 广东省科技计划项目 (2011B080100031)

* 通讯作者: 汤庸, 教授, Email: ytang@m. scnu. edu. cn.

1 推荐系统概念定义

推荐系统利用输入的数据来预测系统对象未来潜在的喜好和兴趣,将个性化推荐对象的清单推送给用户.因此,系统用户、推荐对象及个性化推荐算法是推荐系统的重要组成部分.推荐系统根据用户注册信息、历史行为数据和需求为用户建模,例如用户年龄、职业、爱好、社交网络关系等,同时也根据推荐对象的相关信息来构造对象模型,最后推荐算法根据推荐策略、用户信息和对象信息为用户进行个性化推荐.一般的推荐系统模型流程如图1所示^[1-2].

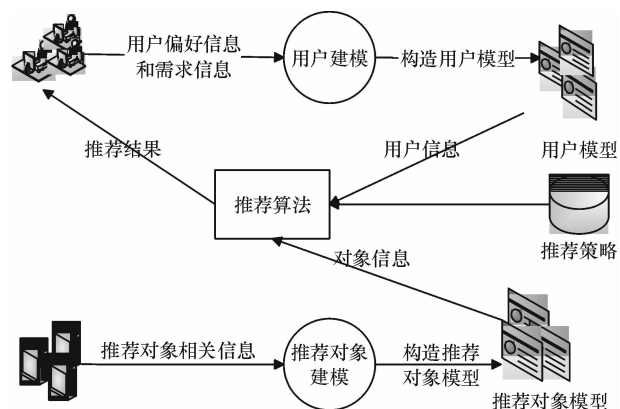


图1 推荐系统模型

Figure 1 The model of recommendation systems

2005年,Adomavicius和Tuzhilin^[3]总结了推荐系统的形式化定义:设 U 是推荐系统中所有用户(users)的集合, C 是有可能被推荐的物品(items)的集合,在用户集合和项目集合的笛卡尔乘积之上定义一个效用函数 f ,可以衡量物品 c 对用户 u 的推荐度,比如: $f:U \times C \rightarrow R$,其中 R 是一个全序关系集合(如一定范围内非负整数或非负实数的全序集和),用来表示用户对项目的评价级别,效用函数性能的优良与推荐效果密切相关.对于任意用户 $u \in U$,可以按照下式找到推荐度最大的项目 c'_u :

$$\forall u \in U, c'_u = \arg \max_{c \in C} f(u, c). \quad (1)$$

2 推荐算法的主要分类

推荐系统利用不同的信息源为用户提供预测和项目的推荐,推荐算法在这个过程中起着重要的作用.根据可用信息不同,例如标签、信任度、社交网络关系、人口统计信息等,及对各种评测指标的考虑,目前流行的推荐算法主要分为四大类:基于内容的推荐算法、协同过滤推荐算法、基于知识的推荐算法

和混合的推荐算法(图2).下面对该4类算法进行介绍并对比优缺点.

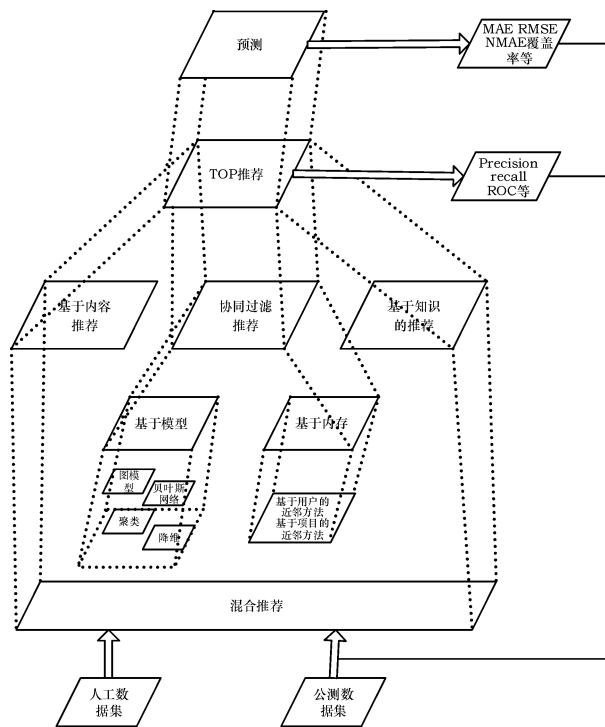


图2 常用推荐算法分类

Figure 2 Classification of recommendation algorithms

2.1 基于内容的推荐算法

基于内容的推荐(content-based recommendation)也称为基于内容的信息过滤推荐,它不需要用户对推荐对象进行评价,而是把推荐对象的内容特征抽取出来,然后从用户以往选择对象的内容特征去学习用户的偏好兴趣,最后与用户偏好兴趣匹配度较高的对象将被推荐给用户.

在基于内容的推荐算法中, $f(u, c)$ 为被推荐对象 c 对用户 u 的效用函数,主要利用对象内容特征和用户资料模型.对象内容特征($Content(c)$)一般采用基于TF-IDF^[4]权重的向量空间模型(Vector Space Model)模型进行表示^[5].用机器学习的方法来获取用户的资料模型($ContentBasedProfile(u)$),例如神经网络、决策树、贝叶斯分类算法等^[2].基于内容的推荐算法中的效用函数 $f(u, c)$ 可表示为^[3]:

$$f(u, c) = score(ContentBasedProfile(u), Content(c)). \quad (2)$$

Score的计算可以采用不同方法,例如可以用二者的余弦相似度计算,如:

$$f(u, c) = \cos(\mathbf{w}_u, \mathbf{w}_c) = \frac{\sum_{i=1}^k w_{i,u} w_{i,c}}{\left(\sqrt{\sum_{i=1}^k w_{i,u}^2} \sqrt{\sum_{i=1}^k w_{i,c}^2} \right)}. \quad (3)$$

最后按所得到的 f 数值对对象排序,将最前面

的项目作为推荐对象. 虽然该算法的推荐结果比较符合用户的喜好, 但缺乏新颖度和惊喜度.

2.2 协同过滤推荐

协同过滤推荐(collaborative filtering recommendation)是推荐系统中最为流行且应用广泛的算法, 它基于系统里其他用户的评分记录或其他历史数据(例如, 在亚马逊上用户购买商品的历史记录). 目前 2 种主要的协同推荐技术是基于内存的协同推荐和基于模型的协同推荐, 前者利用历史数据来寻找相似的推荐项目, 后者通过历史数据构造预测模型, 再通过模型进行评分预测, 它们的不同在于对用户-项目评分矩阵的分析和使用方式上的差异.

2.2.1 基于内存的协同推荐 基于内存的协同推荐(memory-based collaborative filtering)也称为启发式的协同推荐, 主要直接利用用户的历史数据来提供预测结果, 例如用户-电影的评分矩阵. 根据考虑角度的不同, 可以分为基于项目(Item-based)和基于用户(User-based)的协同推荐. 前者是根据项目之间的相似度来寻找与目标项目近似的项目集, 后者则利用用户之间相似的兴趣偏好来获得近邻用户集, 这 2 种基于近邻的推荐算法主要依赖用户数与项目数的比例, 当用户数大大超过项目数时, 基于项目的系统推荐能提供更准确的推荐, 反之亦然.

基于用户的协同推荐的基本步骤: 首先寻找与用户历史评分行为相似的其他用户形成近邻用户集, 然后根据近邻用户的历史偏好或评分来为目标用户进行推荐或评分预测. 在 User-based 模型中, 用户之间的相似度可以选择不同的相似度函数来计算, 例如皮尔逊相关系数:

$$\text{sim}(u, v) = \frac{\sum_{i \in I_u \cap I_v} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_u \cap I_v} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_u \cap I_v} (r_{vi} - \bar{r}_v)^2}}, \quad (4)$$

其中 r_{ui} 为用户 u 对项目 i 的实际评分, \bar{r}_u 和 \bar{r}_v 分别为用户 u 和用户 v 的评分平均值, I_u 和 I_v 分别为用户 u 和用户 v 评过的项目集合, 利用用户之间的共同评分来衡量他们的相似度, 每个用户的评分用向量 $\mathbf{R}_{n \times m}$ 表示, 其中 n, m 分别为用户数及项目数. 计算向量间的距离便可得到用户之间的相似度的计算. 另外, 可以选择的相似度计算方法还有余弦相似度计算、Jaccard 系数等.

确定邻居集合后, 由于不同用户对项目的评分标准存在差异, User-based 模型先对评分进行归一化, 同时由于用户评分存在波动性, 可以引入对评分波动情况的考虑, 例如 Z-scores 归一化方法, 通过

引入用户评分偏差值 σ 来平衡用户间的差异:

$$\hat{r}_{ui} = \bar{r}_u + \sigma_u \frac{\sum_{v \in N_i(u)} \text{sim}(u, v) * (r_{vi} - \bar{r}_v) / \sigma_v}{\sum_{v \in N_i(u)} |\text{sim}(u, v)|}, \quad (5)$$

其中 \hat{r}_{ui} 为用户 u 对项目 i 的预测评分, $N_i(u)$ 为与用户 u 相似的近邻用户集, σ_u 和 σ_v 分别表示目标用户 u 和近邻用户 v 评分的偏差值, 式(5)还对近邻用户的评分进行了加权平均值的处理, 该方法是目前预测计算中简单又有效的方法.

2.2.2 基于模型的推荐算法 由于用户规模和项目数量的快速增长, 数据集的稀疏问题会更严重, 例如 netflix 的电影评分数据集将近 99% 的数据缺失, 那么基于内存的协同算法的计算量将大规模增长, 同时由于数据的稀疏问题, 其推荐结果的质量无疑会下降. 因此, 文献[6]、[7]提出了基于模型的推荐算法(model-based collaborative filtering), 主要思想是利用用户对项目的评分来构造一个评分预测模型, 其中使用了多种机器学习和数据挖掘的计算模型^[7-8], 然后通过该模型实现对未知评分的预测. 常用模型有贝叶斯网络^[9-10]、聚类算法^[11-12]、降维的技术、图模型和回归模型等.

(1) 基于贝叶斯网络的协同过滤. 贝叶斯网络是一种基于概率推理的图形化网络, 也是目前不确定知识表达和推理领域最有效的理论模型之一. 贝叶斯网络包括两部分: 有向无环图及条件概率表. 其中图里的节点代表随机变量, 节点间的有向边代表了节点间的概率依赖关系, 条件概率表中每一行代表一对节点之间的条件概率值, 该值反映了关系的强度^[6].

基于朴素贝叶斯的协同过滤算法认为用户或者项目都具有某些吸引人的特征, 用户对该类特征的偏好是短期稳定的, 因此可以利用贝叶斯理论分析用户或项目特征值. 假设用户和项目属性组成特征集 $X = \{x_1, x_2, \dots, x_n\}$, 将推荐项目分为 2 类: 推荐给用户的项目集合 c_1 和不推荐给用户的项目集合 c_2 , $p(x_i | c_1)$ 、 $p(x_i | c_2)$ 分别表示项目在推荐、不推荐给用户的项目集合中出现特征值 x_i 的概率, 然后选择概率最大的类别作为项目所属类别. 令 X 表示特征集合, C 表示类别集合, 则朴素贝叶斯分类模型:

$$c = \arg \max_{c \in C} p(c_i) \prod_{x \in X} p(x | c_i). \quad (6)$$

(2) 基于聚类的协同过滤. 根据聚类对象的不同可分为基于用户聚类和基于项目聚类的协同过滤. 该方法首先依据用户-项目评分矩阵对用户或项目进行聚类, 然后通过计算目标用户或项目与各类别的相似度, 为其找到所属的类别, 最后通过所属

类别寻找用户或项目的最近邻作推荐. 常用的聚类方法可以分为: 基于密度的方法、基于划分的方法、基于神经网络的方法、基于模糊聚类的方法以及层次聚类的方法. 在聚类分析模型中只将目标用户或项目与预先聚类好的簇进行比较, 而不必再与整个数据集进行比较, 所以在系统的实时性和可扩展性等问题上要比传统的协同过滤方法优越^[13].

(3) 基于降维技术的协同过滤. 随着用户和项目数量大规模增加, 可用于预测评分的数据维度随之增加, 维数灾难^[14]所带来的各种问题将无法避免, 例如在高维空间上搜索最近邻居将非常困难并造成系统性能下降. 因此不少学者引入降维技术把高维数据映射到低维空间中, 从而提高系统的伸缩性. 常用的降维技术有奇异值分解 (singular value decomposition)、概率潜在语义分析 (probabilistic la-

tent semantic analysis)、主成分分析 (principle component analysis)、最大边际矩阵分解 (maximum margin matrix factorization) 等.

(4) 基于图模型的协同过滤. 基于图模型的协同过滤是以图论作为理论基础, 将用户的历史行为数据用图的形式表示, 用户、项目和标签等信息可以看作图结构中不同类型的节点, 分析不同节点之间的关系, 构造关系网络图. 常用的基于图模型的协同推荐算法有二分图和三分图. 前者节点有2类, 分别代表用户和项目, 节点间的边代表用户对项目的评分; 后者比前者多了标签节点^[15]. 如图3A所示, 该二分图有3个用户节点和4个项目节点, 边的权值为用户对项目的评分. 三分图 (图3B) 多了6个标签节点, 用户3-项目2-标签1之间的边代表用户3对项目2打了标签1.

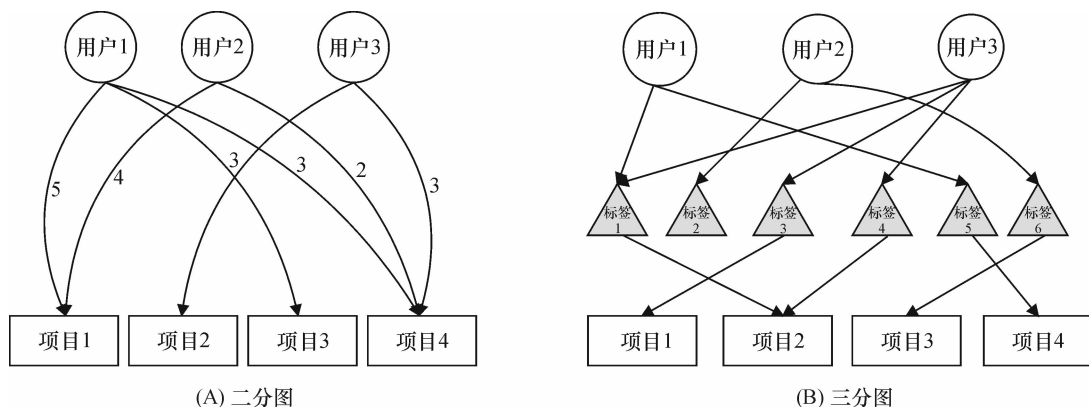


图3 基于图模型的协同过滤

Figure 3 Graph-based collaborative filtering

根据计算原理不同, 可以分为基于物资扩散、热传导和资源分配的二分图/三分图协同推荐. 在基于物资扩散的二分图协同推荐中, 用户相似度为

$$s_{uv} = \frac{1}{k(u)} \sum_{i \in I} \frac{r_{ui} \times r_{vi}}{k(i)}, \quad (7)$$

其中 $k(u)$ 表示用户 u 的度, $k(i)$ 表示项目 i 的度, $r_{ui}=1$ 表示用户 u 对项目 i 进行了评分, 否则 $r_{ui}=0$. 基于物资扩散的三分图协同推荐增加了用户到标签的计算:

$$s_{uv} = \frac{1}{k(u)} \left(\alpha \cdot \sum_{i \in I} \frac{r_{ui} \times r_{vi}}{k(i)} + (1 - \alpha) \sum_{t \in T} \frac{r_{ut} \times r_{vt}}{k(t)} \right), \quad (8)$$

其中 $k(t)$ 表示标签 t 的度, $\alpha \in (0, 1]$ 是随机参数, 当 $\alpha=0$ 或者 $\alpha=1$ 时, 该算法将退化到纯粹的基于项目-标签或用户-项目的二分图算法.

2.2.3 基于知识的推荐 很多推荐算法都无法解决冷启动问题, 因此不少专家提出基于知识的推荐 (knowledge-based recommendation), 希望利用用户

的需求爱好、产品知识和功能知识来为用户推荐项目, 该方法在某种程度可以看成是一种推理 (inference) 技术. 该方法可以分为3类: 基于约束的推荐 (constraint based recommendation)^[16]、基于实例的推荐 (case based recommendation)^[17] 和基于知识推理的推荐系统. 由于不需要用户的历史行为数据, 所以不存在冷启动问题, 但也存在不少问题, 例如如何获取、如何表示领域知识以及在推荐系统中采用哪种交互模式等.

2.2.4 混合的推荐 组合推荐 (hybrid recommendation) 是多种推荐算法的组合, 希望通过不同推荐算法的组合达到保留优点避免缺点的应用效果, 在组合方式上, 主要有7种组合思路^[18]: 加权融合、切换、混合、特征组合、级联、特征扩充及元层次混合. 这7种思路又分成3种基本设计方式: 整体式、流水线式和并行式. 虽然理论上存在多种组合方法, 但不同的组合思路适用于不同的应用场景. 按推荐算法组合发生的阶段及融合的程度可以把组合推荐分为

前融合、中融合和后融合^[19-21]:(1)前融合:对各种推荐算法进行直接融合,统一在一个框架模型里,从算法层次看是深度融合;(2)中融合:在以某种推荐算法为主的框架模型上,融合另一种推荐算法的部分特征;(3)后融合:直接将多种推荐算法各自计算得到的结果进行融合.

常用的混合推荐有:(1)将协同推荐和基于内

容推荐单独运行的结果进行组合推荐;(2)将基于内容推荐的特征融合到某种协同推荐中;(3)建立具有基于内容推荐和协同推荐特点的预测模型.

各种推荐算法都存在自身的优、缺点,有各自适用的环境和数据源,在推荐的效果方面也存在差异(表 1).

表 1 常用协同过滤算法及其优缺点
Table 1 The advantages and disadvantages of collaborative algorithms

推荐算法分类	代表性算法	主要优点	主要缺点
基于内容的推荐	文本推荐方法	(1)可处理新项目的冷启动问题; (2)解决评分稀疏问题; (3)能为具有特殊兴趣爱好的用户进行推荐; (4)能推荐新产品和非流行产品; (5)可以直观解释推荐的原因.	(1)无法有效处理非结构化数据,例如音频、视频等多媒体信数据; (2)缺乏对用户潜在的挖掘能力; (3)存在新用户出现时的冷启动问题; (4)自我学习能力较差.
基于内存的协同推荐	(1)基于近邻的推荐; (2)基于用户/项目 Top-N 推荐.	(1)适合处理复杂的非结构数据; (2)有效发现新颖性高的项目; (3)可提供推荐解释.	(1)存在冷启动问题; (2)数据规模扩大,数据稀疏严重时,推荐质量和效率都下降; (3)依赖历史行为数据.
基于模型的协同推荐	(1)贝叶斯网络; (2)聚类(k-means、神经网络); (3)降维技术(SVD、LDA、pLSA、PCA 等); (4)图模型.	(1)有效缓解了数据的稀疏性问题; (2)增强了系统的扩展性; (3)提高了预测精度.	(1)建模复杂且需要周期性更新; (2)降维会造成部分信息损失,难以提供推荐解释.
基于知识的推荐	(1)基于约束的推荐; (2)基于实例的推荐; (3)基于知识推理的推荐.	(1)不依赖用户的历史行为数据; (2)解决冷启动问题.	(1)推荐质量取决于知识获取和质量; (2)推荐是静态的.
混合式协同过滤	(1)基于内容的混合式协同过滤; (2)基于内存和基于模型的混合式协同过滤.	(1)混合推荐可以取长补短; (2)利用内容分析可处理新项目或新用户问题.	(1)方法的选取与组合顺序; (2)对不同方法所得结果的权重分配.

3 评测的方法及指标

3.1 评测方法

评价推荐系统的方法一般分为 3 种:离线实验 (offline experiment)、在线实验 (online experiment) 和用户调查 (user study)^[20]. 离线实验是将用户数据集分为训练集和测试集,训练集用于用户兴趣模型构造,测试集用于对模型性能测试. 该方法利用用户的历史数据进行测试,不需要真正用户的参与,比在线实验的成本低,适用于对不同推荐算法的比较和过滤,然而由于缺乏实际用户参与,算法的预测质量难以保证. 因此当完成离线实验后,还需要进行在线实验. 在线实验是根据用户在线实时反馈或事后问

卷调查等结果来衡量推荐系统的表现^[20]. 用户调查是让测试用户来完成被测试的推荐系统的一系列任务并对测试用户的行为进行记录. 优点在于测试到推荐系统对用户行为的影响,体现用户主观的指标以及出现错误后容易弥补,但是该方法由于成本高和需要大量人员参与,从而让测试的范围和次数受到了限制.

3.2 评测指标

任何推荐算法都需要通过评测,这样才能评估它的推荐质量. 常用的推荐评测指标有:用户满意度 (user preference)、预测准确度 (accuracy)、覆盖率 (coverage)、多样性 (diversity) 和新颖性 (novelty) 等.

3.2.1 用户满意度 作为评测推荐系统的最重要指标之一,用户满意度描述用户对推荐结果的满意

程度.一般可采用调查问卷或用户在线行为数据分析的形式来获取用户满意度.

3.2.2 预测的准确度 预测准确度用于衡量推荐算法预测用户对项目兴趣偏好的能力,主要是采用统计学的方法来量化推荐系统产生的预测评分与实际评分之间的误差.常用的预测准确度分为评分预测准确度、使用预测准确度和排序准确度.

常用评分预测有平均绝对误差 (Mean Absolute Error, MAE)、归一化平均绝对误差 (Normalized Mean Absolute Error, NMAE)、均方根误差 (Root Mean Squared Error, RMSE)、归一化均方根误差 (Normalized Root Mean Squared Error, NRMSE)^[22].

平均绝对误差 (MAE) 是最常用的评估方法之一,它是通过统计预测值和真实评分值之间的绝对误差值得到的,计算公式为:

$$MAE = \frac{\sum_{u,i \in T} |r_{ui} - \hat{r}_{ui}|}{|n|}, \quad (9)$$

其中 n 为评分的总数, r_{ui} 为用户 u 对项目 i 的实际评分, \hat{r}_{ui} 为用户 u 对项目 i 的预测评分. MAE 数值与推荐的预测准确性呈反比. NMAE 是 MAE 在评分值的范围内归一化后的结果,着重于较大的偏差.

均方根误差 (RMSE) 与 MAE 相比, RMSE 加大了对预测不准的用户项目评分的惩罚,因而对系统的评测更加苛刻. 其计算公式如下:

$$RMSE = \sqrt{\frac{\sum_{u,i \in T} (r_{ui} - \hat{r}_{ui})^2}{n}}. \quad (10)$$

Top-N 推荐属于使用预测准确度,由推荐网站采用个性化的推荐列表为用户提供推荐. 这类推荐的重点在于分类准确率,常用分类衡量指标有准确率 (precision) 和召回率 (recall).

3.2.3 覆盖率 覆盖率^[22]体现的是推荐系统挖掘长尾的能力,主要通过研究推荐对象在推荐列表中出现的次数的分布来描述,适用于那些需要为用户找出所有感兴趣商品的系统. 常用覆盖率有预测覆盖率 (prediction coverage)、带权预测覆盖率 (weighted prediction coverage)、用户覆盖率 (user coverage)、种类覆盖率 (catalog coverage) 和带权种类覆盖率 (weighted catalog coverage) 等^[23].

预测覆盖率表示系统可以预测评分的项目占所有项目的比例,该指标依赖于推荐算法和输入值,定义为:

$$Predictioncoverage = \frac{|I_p|}{|I|}, \quad (11)$$

其中 I_p 表示系统可以预测评分的项目集, I 为所有项目集.

带权预测覆盖率^[23]主要考虑了项目的效用性,该效用性主要从准确性、新颖性和推荐有效性等方面获取. 用 $r(i)$ 代表项目 i 的效用值,定义为:

$$Weightedpredictioncoverage = \frac{\sum_{i \in I_p} r(i)}{\sum_{j \in I} r(j)}. \quad (12)$$

用户覆盖率 (UCOV) 表示推荐算法能为多大比例用户计算推荐结果的能力,与准确性指标结合使用,常用于分析新用户对系统影响.

3.2.4 多样性、新颖性和惊喜度 一个好的推荐系统应该能为用户提供多样化的选择,即推荐结果应具有多样性. 在推荐系统中,多样性可以从个体层面和总体层面进行评估. 个体层面主要考虑的是推荐系统对单用户推荐项目的多样性,主要计算推荐项目差异平均值;总体层面则是衡量推荐系统对所有用户推荐不同项目的能力^[24].

新颖性^[25]和惊喜度是最近 2 年推荐系统领域比较关注的指标. 新颖的推荐是指给用户推荐那些他们以前没有听说过的物品. 新颖度是为了度量推荐列表中物品对于用户的新颖程度. 某种程度上,惊喜度同时兼顾了新颖性和准确度.

4 推荐算法的常用数据集

随着推荐系统的广泛应用,目前大部分推荐算法所用数据集都是来源现实的推荐系统,每年数据挖掘与知识发现竞赛都采用企业提供的实际应用数据集. 但有时候为了测试一些特定应用领域的算法需要采用人工数据集,但该类数据应用具有局限性. 目前在推荐算法研究中,已经有很多被研究人员公认的标准测试数据集并且成为推荐算法性能测试的基础和依据. 常用公认的标准测试数据集有:

(1) 腾讯微博数据集. 腾讯微博数据集是由 2012 年的数据挖掘与知识发现竞赛 (Kdd-cup) 所提供的. 该数据集是从腾讯 4.25 亿微博用户中的 50 天数据采样得到,包含有 200 多万活跃用户、6 千被推荐用户或信息源、3 亿多条推荐记录及其 300 多万个收听动作、7 000 多万条训练记录、3 000 多万条测试记录. 该数据集的规模超过以往的 KDD Cup 比赛.

(2) Netflix 数据集^[26]. 该数据集来自著名的电影网站 Netflix. 包含 480 189 位用户对 17 770 部电影的 100 480 507 条评分记录,与 MovieLen 评分不同,其数据为区间 $[1, 5]$ 的离散整数值,这是目前规模最大的电影评分数据集. 由于 Netflix 比赛已经结束,该数据集已不对外公开.

(3) CiteULike 数据集^[27]. CiteULike 是由著名的

施普林格出版社 (Springer) 提供的一个免费协助用户存储、管理和分享学术文章的网站,用户可以收藏自己喜欢的论文并给它们打上标签. CiteULike 公布了一个包含了用户收藏论文及给论文打标签的数据集. 这个数据集包含 52 689 个用户,1 793 954 篇论文以及 2 119 200 个用户和论文之间的关系.

表 2 推荐系统评价标准分类

Table 2 Evaluation criteria of recommendation systems

评价指标		评测方法		
		离线	在线	用户调查
用户满意度	用户对推荐系统的满意程度,主要通过问卷调查或者在线统计一些用户行为获得.		√	√
准确度	预测评分准确度:包括平均绝对误差、平方根误差、均方根误差等;与推荐列表长度无关;适用于评分预测的系统. TOP-N 推荐:包括准确率、召回率;依赖于推荐列表长度;不适用于没有明确二分喜好的系统.	√		
覆盖率	该指标能够反映推荐系统对项目长尾的挖掘能力. 可分为预测覆盖率、带权预测覆盖率、用户覆盖率、推荐覆盖率和种类覆盖率等;后 2 种覆盖率依赖于推荐列表长度;覆盖率应结合准确度指标一起考虑.	√	√	√
多样性	推荐列表应尽量覆盖用户的不同兴趣领域,分为用户间的多样性和用户内的多样性;依赖于推荐列表长度.		√	√
新颖度	以用户以往的兴趣爱好为参考,推荐项目是用户之前不知道的但能让用户满意的.		√	√
惊喜度	它是衡量推荐系统能令用户惊喜的程度,不依靠用户以往的兴趣爱好,但推荐项目还是让用户满意的.		√	√

(4)Yahoo! 音乐数据集^[28]. Yahoo! 音乐数据集包含了用户对单曲、专辑、歌手以及流派等不同的音乐元素的评分. 评分区间是 0 ~ 100 之间的整数,总共涉及 1 000 990 个用户、624 961 个音乐元素以及 262 810 175 条评分记录.

5 总结与展望

推荐系统可应用于不同领域,在海量个性化需求的驱动下,随着云计算、个性化服务、人工智能、决策科学和信息检索等领域发展的推动,其发展应用迅速并取得很好的研究成果. 但也存在下面几个问题,而这些问题也是未来研究的热点^[20,29-32].

(1)上下文感知推荐系统. 上下文包含多种多样的信息,例如时间、位置和情感等. 因此上下文感知推荐系统的主要任务就是如何将上下文信息应用到推荐算法当中,从而提高推荐的精确度和用户满意度.

(2)推荐的实时性研究. 由于数据量的庞大,推荐系统的推荐精度和实时性一直都是一个矛盾. 目前大部分系统采用离线计算,某种程度上是以推荐质量为代价的,因此如何有效提高推荐系统的推荐质量并兼顾系统的实时性,需要做进一步深入的研究.

(3)稀疏性和冷启动研究. 这 2 个问题一直是

推荐系统的研究难点,它们可以导致协同过滤模式的运行效率和推荐精确度较低. 尽管很多学者对该问题进行研究并提出解决办法,但效果并不十分显著,还需要对其进行研究.

参考文献:

[1] Pazzani M J, Billsus D. Content-based recommendation systems[M] // Brusilovsky P, Kobsa A, Nejdl W. The Adaptive Web. Berlin, Heidelberg: Springer - Verlag, 2007,4321:325 - 341.

[2] Melville P, Mooney R J, Nagarajan R. Content-boosted collaborative filtering for improved recommendations[C] //Proceeding of the 18th national conference on artificial intelligence. Edmonton: AAAI Press, 2002: 187 - 192.

[3] Adomavicius G, Tuzhilin A. Toward the next generation recommender systems: A survey of the state-of-the-art and possible extensions[J]. IEEE Trans on Knowledge and Data Engineering,2005,17(6):734 - 749.

[4] Belkin N, Croft B. Information filtering and information retrieval[J]. Communications of the ACM,1992,35(12): 29 - 37.

[5] Shardanand U, Maes P. Social information filtering: Algorithms for automating ‘Word of Mouth’[C] //Proceedings of the ACM SIGCHI conference on human factors in computing systems. Denver: ACM Press, 1995:210 - 217.

[6] Breese J, Hecherman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[C] //Pro-

- ceedings of the 14th conference on uncertainty in artificial intelligence (UAI'98). San Francisco: Morgan Kaufmann Publishers, 1998:43-52.
- [7] Sarwar B, Konstan J, Riedl J. Incremental singular value decomposition algorithms for highly scalable recommender systems[C]//Proceedings of the 5th international conference on computer and information science. Dhaka, Bangladesh, 2002.
- [8] Koren Y. Factorization meets the neighborhood: A multifaceted collaborative filtering model[C]//Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. Las Vegas, USA, 2008.
- [9] Su X, Khoshgoftaar T M. Collaborative filtering for multi-class data using belief nets algorithms[C]//Proceedings of the 18th IEEE international conference on tools with artificial intelligence. Arlington, USA, 2006.
- [10] Miyahara K, Pazzani M J. Collaborative filtering with the simple Bayesian classifier[C]//Proceedings of the 6th pacific rim international conference on artificial intelligence. Melbourne, Australia, 2000.
- [11] Connor M, Herlocker J. Clustering items for collaborative filtering[C]//Proceedings of the ACM SIGIR workshop on recommender systems. Berkeley, California, 1999.
- [12] Sarwar B, Karypis G, Konstan J, et al. Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering[C]//Proceedings of the 5th international conference on computer and information technology. Dhaka, Bangladesh, 2002.
- [13] Kohrs A, Merialdo B. Cluster for collaborative filtering application[C]//Proceedings of the international conference on computational intelligence for modelling control and automation. Amsterdam: IOS Press, 1999: 199-204.
- [14] 杨风召. 高维数据挖掘技术研究[M]. 南京:东南大学出版社, 2007.
- [15] Zhang Z K, Zhou T, Zhang Y C. Personalized recommendation via integrated diffusion on User-Item-Tag tripartite graphs[J]. Physica A, 2010, 389: 179-186.
- [16] Felfernig A, Kiener A. Knowledge-based interactive selling of financial services with FSAdvisor[C]//Proceedings of the 17th innovative applications of artificial intelligence conference (AAAI). Pittsburgh: AAAI Press, 2005: 1475-1482.
- [17] Bridge D, Göker M H, McGinty L, et al. Case-based recommender systems[J]. Knowledge Engineering Review, 2005, 20(3):315-320.
- [18] Robin B. Hybrid recommender systems: Survey and experiments[R]. Fullerton: California State University, 2003.
- [19] Claypool M, Gokhale A, Miranda T, et al. Combining content-based and collaborative filters in an online newspaper[C]//Proceedings of the ACM SIGIR '99 workshop on recommender systems: Algorithms and evaluation. Berkeley: ACM, 1999.
- [20] 项亮. 推荐系统实践[M]. 3版. 北京:人民邮电出版社, 2012:41-43.
- [21] 徐海玲, 吴潇, 李晓东, 等. 互联网推荐系统比较研究[J]. 软件学报, 2009, 20(2): 350-362.
- [22] Zhou T, Su R Q, Liu R R, et al. Accurate and diverse recommendations via eliminating redundant correlations[J]. New Journal of Physics, 2009, 11:123008-123026.
- [23] Ge M, Delgado-Battenfeld C, Jannach D. Beyond accuracy: Evaluating recommender systems by coverage and serendipity[C]//Proceedings of the fourth ACM conference on recommender systems. New York: ACM, 2010: 257-260.
- [24] Celma O, Herrera Venue P. A new approach to evaluating novel recommendations[C]//Proceedings of the 2008 ACM conference on recommender systems. New York: ACM, 2008: 179-186.
- [25] Adomavicius G, Kwon Y. Maximizing aggregate recommendation diversity: A graph-theoretic approach[C]//Proceeding of RecSys workshop on novelty and diversity in recommender systems. Chicago, USA, 2011:3-10.
- [26] Bennett J, Lanning S. The netflix prize[C]//Proceedings of KDD cup and workshop. San Jose: ACM, 2007.
- [27] Zlatić V, Ghoshal G, Caldarelli G. Hypergraph topological quantities for tagged social networks[J]. Physical Review E, 2009, 80:8pp.
- [28] Dror G, Koenigstein N, Koren Y, et al. The Yahoo! music dataset and KDD-cup'2011[DB/OL]. (2011-06-30)[2014-02-10]. <http://webscope.sandbox.yahoo.com/catalog.php?datatype=c>.
- [29] 马宏伟, 张光卫, 李鹏. 协同过滤推荐算法综述[J]. 小型微型计算机系统, 2009, 30(7): 1282-1288.
- Ma H W, Zhang G W, Li P. Survey of collaborative filtering algorithms[J]. Journal of Chinese Computer Systems, 2009, 30(7): 1282-1288.
- [30] 曾春, 邢春晓, 周立柱. 个性化服务技术综述[J]. 软件学报, 2002, 13(10): 1952-1961.
- Zeng C, Xing C X, Zhou L Z. A survey of personalization technology[J]. Journal of Software, 2002, 13(10): 1952-1961.
- [31] 夏培勇. 个性化推荐技术中的协同过滤算法研究[D]. 青岛:中国海洋大学, 2011.
- Xia P Y. Research on collaborative filtering algorithm of personalized recommendation technology[D]. Qingdao: Ocean University of China, 2011.
- [32] 任磊. 推荐系统关键技术研究[D]. 上海:华东师范大学, 2012.
- Ren L. Research on some key issues of recommender systems[D]. Shanghai: East China Normal University, 2012.