一种数据递增式的混合推荐方法①

陈洪涛[1], 肖如良[1], 林丽玉[1], 颜杰敏[2], 蔡声镇[1]

[1](福建师范大学 软件学院, 福州 350108)
[2](73683 部队, 福州 350003)

摘 要: 推荐系统由于较大的训练数据量和推荐算法较高的复杂度, 其推荐的更新周期往往较长. 然而系统上的数据时刻都在增长, 更新推荐期间会产生大量数据, 这些新数据对下一刻的推荐有较大的利用价值, 系统却无法及时利用起来. 为了能及时的利用这些新数据来提高推荐系统的推荐质量, 提出一种数据递增式的混合推荐方法. 该模型主要分为离线计算模块和在线推荐模块, 离线模块用于计算出个性化推荐列表, 在线推荐模根据递增的实时数据维护一个流行趋势动量表, 然后结合两个模块的结果给出匿名推荐或者个性化推荐. 实验证明, 该方法简单、有效、可行, 能较好的改善推荐系统性能.

关键词: 推荐系统; 更新周期; 递增数据; 流行趋势动量; 混合推荐

New Approach to Hybrid Recommendation Based on Incremental Data

CHEN Hong-Tao[1], XIAO Ru-Liang[1], LIN Li-Yu[1], YAN Jie-Min[2], CAI Sheng-Zhen[1]

[1](Faculty of Software, Fujian Normal University, Fuzhou 350108, China)
[2](73683 Army, Fuzhou 350108, China)

Abstract: Due to the large amount of training data and the high complexity of its recommend algorithm, the updating cycle of recommendation system tend to be long. However, the data on the system is growing all the time, and a lot of data is produced during the cycle, which is useful for the recommendation of next moment, and recommendation system can't use these data in time. In order to use these data in time to improve the quality of recommendation system, a new approach to hybrid recommendation based on incremental data was proposed. The approach mainly divided recommendation into offline and online module, the offline module is used to produce the personalized recommendation list, while the online recommendation module maintains a list of popular trend momentum based on real-time and incremental data. Then, combining with the results of the two modules, based on which give users anonymous or personalized recommendation. Experiments show that the approach is simple, effective, feasible, and can improve the performance of recommendation system better.

Key words: recommendation system; updating cycle; incremental data; popular trend momentum; hybrid recommendation

1 引言

大数据不仅意味着海量的数据, 同时也关联了技术的巨大发展与创新. 而如今, 我们已经步入了大数据时代, 随着计算机应用的普及以及互联网技术的发展, 网络上的信息正在成指数增长, 为了迎合用户面对海量数据时快速找到喜欢物品的需求, 推荐系统应运而生. 几乎所有的推荐系统都是由前台展示页面、后台日志系统以及推荐算法系统三部分组成, 其中核心部分是推荐算法. 传统的推荐算法[1]主要包括协同过滤推荐算法[2], 基于内容的推荐算法[3,4], 基于网络结

① 基金项目:教育部规划基金(11YJA860028);福建省自然科学基金(2013J01219)
收稿时间:2014-02-24;收到修改稿时间:2014-03-17

构的推荐算法[5,6]以及混合推荐算法[7,8]. 近年来, 人们逐渐从信息匮乏的时代走向了信息过载的时代[9,10], 传统的推荐算法也渐渐暴露出各自的缺陷[11]. 协同过滤算法受限于数据稀疏, 冷启动以及可扩展性等问题[5]; 基于内容的推荐算法不可避免的地因内容特征信息获取技术的局限性而使其在很多时候效果不如协同过滤算法[3,6]; 基于网络结构的很多推荐算法具有较高的复杂度并且也存在冷启动问题[3,9]; 混合推荐算法虽然综合其他算法做推荐, 相互之间取长补短, 但仍无法从根本上解决问题, 同时使得算法的复杂度较高[3]. 与此同时, 推荐系统也越来越重视用户登录之前的匿名推荐. 匿名推荐研究中[12,13], Knuchel JP 等人以及 Stritt M 等人分别提出了结合学习组件和知识挖掘的匿名推荐系统和基于属性分类的匿名推荐系统, 并分别在不可预见性场景和多噪声场景的推荐中取得较好的效果. 然而上述匿名推荐算法都有较高的复杂度, 并且没有个性化推荐, 在现实应用中有较大的局限性. 面对传统推荐算法的缺陷, 近年来的研究在针对上述问题上取得了一定的成果, 然而, 目前的推荐系统或多或少仍存在下述的二个问题.

1) 不支持基于数据递增式更新推荐[14]. 在数据过载的今天, 推荐系统上的数据时刻都在增长, 推荐物品理论上应该是按照毫秒为单位在持续动态的变化, 用户也希望在对系统回馈之后能很快的得到实时更新的结果. 然而, 目前的推荐算法一般都需要一系列基于静态训练集上的复杂工作, 而实时的动态跟踪这些快速膨胀的数据需要不停的完全重建模型, 这种做法显然不可行, 因为计算量太大, 也不可能达到即时回馈用户的目标.

2) 缺乏较好的匿名推荐. 大多数推荐系统都只着重于个性化推荐, 却忽略了用户在登录之前的匿名推荐的作用. 一方面, 用户必须进入门户网站并登录系统以后才能对其进行个性化推荐, 若在用户登录之前就能推荐给他喜欢的物品, 那么用户将更有欲望登录系统寻找自己喜爱的物品; 另一方面, 当未注册的陌生用户来到此系统时, 准确的匿名推荐也能留给他们良好的印象. 因此在用户登录系统之前对其进行效果较好的匿名推荐同样有重要意义. 有些推荐系统虽然有匿名推荐部分, 但通常用来做匿名推荐的如 Most-popular 等算法, 其推荐效果不佳.

针对以上问题, 本文提出一种基于数据递增变化的匿名与个性化混合推荐模型 APRS(Anonymous and Personalized Recommendation based on Incremental Data). 该模型分为离线计算部分和在线推荐部分, 其中离线部分主要包括个性化推荐算法, 计算物品的个性化推荐值, 在线部分维护一个流行趋势动量表, 用于追踪系统数据的实时变化. 该模型不但实现了基于数据递增式更新推荐, 而且也实现了匿名推荐和个性化推荐一体化. 通过实验可知, 该模型下的匿名推荐和个性化推荐相均有较好的推荐效果.

2 数据递增式混合推荐模型APRS

本文提出的基于数据递增变化的匿名与个性化混合推荐模型主要分为离线计算模块和在线推荐模块(如图 1). 其中离线模块主要利用个性化推荐算法计算个性化推荐度量值矩阵, 类似于传统推荐模型, 该模块更新较慢, 无法适应递增的数据. 而在线模块利用最近时间段的数据计算一个物品流行趋势动量表并高频率的更新, 以此来追踪快速膨胀的数据, 实时的预测并更新下一刻物品热度走势, 弥补了离线模块的不足. 若有“陌生用户”来访系统, 则根据流行趋势动量表来给出最有“潜力”的匿名推荐, 若用户登录系统, 则提取离线模块计算结果中的个性化推荐度量值矩阵中对应该用户的所有物品推荐度量值, 组成一个推荐列表, 然后根据每个物品的推荐值以及实时更新的流行趋势动量表综合分析给出有效的推荐.

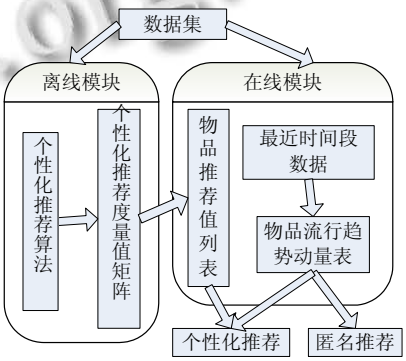

图 1 基于数据递增变化的匿名与个性化混合推荐模型

2.1 离线推荐模块

离线模块主要是由离线个性化推荐算法以及个性化推荐度量值矩阵构成. 离线模块的工作主要分为如下三个步骤: ①首先读取全部数据集记录; ②利用个性化推荐算法得到每个用户对应每个物品的推荐值; ③

最后将②中计算得出的结果规约处理, 然后存放在个性化推荐度量值矩阵 P 中. 当需要对某用户进行个性化推荐时, 将 P 中对应的该用户个性化推荐信息传递到在线模块中的物品推荐值列表 V 中. 然而由于数据量的巨大以及个性化推荐算法较高的复杂度, 矩阵 P 只能以较长时间段 T 为周期更新.

通过离线模块可以计算出较精准的个性化推荐. 然而由于数据迅速的增长, 周期 T 内产生的数据量不容忽视, 同时新数据有更高的价值, 离线模块却无法利用起来, 推荐实时性不好. 为了弥补离线模块的缺陷, 实现基于数据递增的推荐, 设计了在线模块.

2.2 基于数据递增式的在线推荐模块

物品的流行趋势总是随着时间的变化而变化, 通过物品近阶段被访问信息的频数以及其变化情况, 可以预测出下阶段该物品的流行趋势, 较好的抓住物品受欢迎程度 F 以及其即将变化的趋势 *Trend* 等信息. F 可以用当前时间物品的平均访问量来度量, 而 *Trend* 的计算是基于 F 的, 可以由当前某时间点为基点的短时间段物品平均访问量 F_s 减去长时间段物品平均访问量 F_l 来衡量. 为了将物品的这种受欢迎程度 F 以及其变化趋势 *Trend* 两种因素进行融合量化, 在计算物品 *Trend* 时, 将 F_l 上添加一个指数控制参数 $\eta(\eta\in(0,1])$, 如此可以得到一个能较好预测下一刻物品访问率的融合变量, 称之为瞬时流行趋势动量 *PPTM*. 当 η 越大, *PPTM* 的计算中, 因素 *Trend* 占比重越大, 反之则 F 占比重越大. 然而 *PPTM* 的计算只是基于某时间点, 为了使结果更可靠, 均匀的取最近时间段多个时间点计算 *PPTM* 并取均值, 得到物品的流行趋势动量 *PTM*. 计算 *PTM* 具体步骤如下:

①读取以当前时刻为时间点的最近时间段数据, 最近时间段长度为离线推荐模块更新周期 T;

②将时长 T 为的最近时间段均匀划分为多个时间片, 按照逆时顺序依次命名为 t_1, t_2, t_3……, 物品 j 在时间片 t_i 里的被访问数用 $f(j,t_i)$ 表示;

③确定长时间段和短时间段的滑动窗口大小 l 和 s(其中 $0<s<l<T$), 计算物品 j 在时间片 t_i(其中 $1\le i\le T-l+1$)末尾时间点处的瞬时流行趋势动量 *PPTM*, 如公式(1)所示;

$$PPTM(j,t_i)=\frac{p\sum_{k=i}^{s+i-1}f(j,t_k)}{s}-\left(\frac{p\sum_{k=i}^{l+i-1}f(j,t_k)}{l}\right)^{\eta} \tag{1}$$

其中 p 用来控制 *PPTM* 的计算精度, 其大小由数据集稀疏性而定, 变量 $\eta\in(0,1]$, 该变量用来调整物品当前访问量和访问量变化趋势两个因素对结果的影响权重.

④计算物品 j 的流行趋势动量 *PTM*, 如公式(2)所示.

$$PTM(j)=\frac{\sum_{k=1}^{T-l+1}PPTM(j,t_k)}{T-l} \tag{2}$$

将所有物品的流行趋势动量汇总起来即可得到流行趋势动量表. 由于流行趋势动量表计算简单, 复杂度较低, 同时用于计算的数据量少, 因此可以实现高频率快速更新, 能够及时的利用产生的新数据来做计算, 有较强的实时性. 另外, *PTM* 能够实时的反映物品流行的变化趋势, 对提高推荐效果有重要意义.

2.3 数据递增式混合推荐

对于任何物品, 该物品的流行趋势动量 *PTM* 越大, 说明其下一刻被访问的可能性越大, 被推荐的概率也越大, 因此适合用来做匿名推荐. 当用户未登录系统时, 可为匿名用户提供 *PTM* 值较大的物品的推荐.

由 *PTM* 的意义可知, 该变量对提升个性化推荐质量同样也有重要意义. *PTM* 是利用实时产生的数据计算得到的, 并且不断的迅速更新的变量. 如果将该变量与离线模块的个性化推荐融合起来做推荐, 那么就可以实时的将产生的新数据用于推荐中, 实现了基于数据递增式推荐. 不断的将新产生的数据及时的用于个性化推荐中, 可以提高个性化推荐的质量, 因此可将离线模块得到的个性化推荐值同物品的流行趋势动量进行加权融合, 得到最终的个性化推荐值, 并以此为根据做个性化推荐. 用户 u 对物品 j 的最终推荐值 *Value* 可用公式(3)来计算.

$$Value(u,j)=\mu\cdot V(u,j)+(1-\mu)\cdot PTM(j) \tag{3}$$

其中, 参数 $\mu(\mu\in[0,1])$是控制离线模块和在线模块计算结果对最终结果影响的权重因子, $V(u,j)$表示通过离线模块计算得到的用户 u 对物品 j 的个性化推荐值. μ 越大则推荐的离线权重越大, 反之则越小. 当用户登录系统以后, 可为用户提供 *Value* 值较大的物品进行个性化推荐.

3　实验结果与分析

本文提出的推荐模型对实验数据集有这样的要求, 数据集上面的记录必须有时间信息, 并且数据集是通过长期收集而来. 因此选用符合要求的 Netflix 数据集, 该数据集是 Netflix Prize 竞赛数据集, 包含了从 1998 年 10 月到 2005 年 12 月时间范围内 480189 个有唯一 ID 标志的用户对 17770 有唯一 ID 标志的电影的 100480507 条评分记录, 每条记录都有相应的评分时间. 为了提高实验效率, 先对数据做预处理. 为了不破坏数据集的整体特性, 随机选取 4000 个电影和 6000 个用户, 并按顺序更新 ID. 最后将日期字段更新为从 1998 年 1 月 1 日到评分记录产生之日之间的天数, 然后按照原格式存起来. 预处理后的数据集包括 987016 条记录, 最后一个评分记录在 1998 年 1 月 1 日之后的第 2922 天.

3.1 实验及数据集

为了使离线模块的个性化推荐计算有充分的数据, 同时又不使在线模块训练集过于稀疏, 采用数据集中时间顺序的中后期某时间点进行实验, 并假设离线模块的更新周期 T 为 30 天, 第 2832 天是离线模块更新日期, 2832 天到 2862 天之间是离线模块某一更新推荐周期. 为了体现混合模型的数据递增式推荐, 分别在该周期起始点、中间点和末尾点三个时间点做推荐, 因此实验数据按照上述要求分为三组, 依次对应周期的起始点、中间的和末尾点(如表 1).

表 1　分组情况(0-2832 表示 0 天到 2832 天内的所有数据)

	第一组	第二组	第三组
离线模块训练集	0-2832	0-2832	0-2832
在线模块训练集	2802-2832	2817-2847	2832-2862
测试集	2832-2847	2847-2862	2862-2877

按照表 1 所示, 在计算 PTM 时, 按照一天为单位时间片, 以 2832 天为起始点的下一个周期 T 划分为 t_1-$t30$ 这 30 个时间片. 最近短时间段 s = 10 天, 最近长时间段 l = 20 天, 鉴于数据的稀疏性, 取参数 p 为 10, 离线模块的个性化推荐算法用常见的基于用户的协同过滤算法 U-CF,基于物品的协同过滤算法 I-CF,基于图模型的算法 Graph 以及基于随机游走的 P-Rank 算法分别实验. 所有的实验均按照 N=10 的 Top-N 推荐来进行. 最后给出每种算越高表明推荐算法的综合性能越好, 实验以此为评估法的准确率和召回率的调和平均值(公式(4)), 该值标准.

$$F_{measure}=\frac{2\cdot precision\cdot recall}{precision+recall} \tag{4}$$

3.2 实验结果与评价

首先分析公式⑶中数参数 η 值. 在各组实验数据集下取不同的 η 值进行匿名推荐, 匿名推荐的效果越好, 说明最近时段数据变化情况追踪的越好, 那么考虑最近时段数据变化情况的混合推荐效果就越好. 因此匿名推荐结果的准确率和召回率调和平均值 $F_{measure}$ 越大, 说明此时 η 值取值越合适. 取 μ 值为 0, 实验结果如图 2 所示.

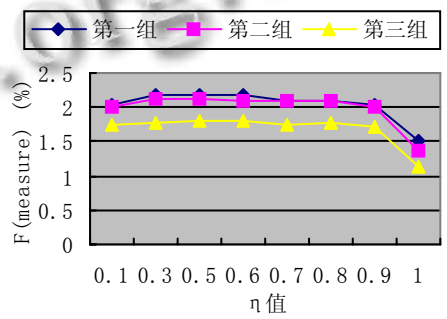

图 2　不同 η 值在各组数据集上匿名推荐结果

由图 2 可知当 0.1<η<0.3 时, $F_{measure}$ 缓慢增长, 当 0.3<η<0.6 时, $F_{measure}$ 保持平稳, 当 0.6<η<0.9 时, $F_{measure}$ 有减小的趋势,当 0.9<η<1.0 时, $F_{measure}$ 减小非常快. 综上所述, 可以取 η=0.5 为最优参数来进行后面的实验.

在离线模块训练集基础上进行离线模块的个性化推荐, 同时结合在线模块计算得出的结果, 利用公式⑶进行综合推荐, 用公式⑷的 $F_{measure}$ 来作为推荐效果度量. 权重因子 μ 分别取值范围为集合{0, 0.1, 0.2, …, 0.9, 1.0}. 四种离线算法的均值结果如表 2 所示.

表 2　不同 μ 值下的 $F_{measure}$ (%)平均值

	第一组	第二组	第三组
μ=0	2.1847	2.1155	1.7952
μ=0.1	2.4895	2.2897	1.9924
μ=0.2	2.7359	2.474	2.163
μ=0.3	2.7862	2.5257	2.2042
μ=0.4	2.7315	2.5585	2.2155
μ=0.5	2.7018	2.4582	2.1969
μ=0.6	2.636	2.3585	2.0814
μ=0.7	2.4951	2.2966	1.9393
μ=0.8	2.2612	2.0747	1.8052
μ=0.9	2.1109	1.9236	1.693
μ=1.0	1.7	1.5607	1.4101

表 2 给出了 4 个离线算法对应不同 μ 值在 3 组数据集上的 $F_{measure}$ 均值，由此表可以看出，调和均值 $F_{measure}$ 从第 1 组到第 3 组呈递减趋势. 在离线模块更新的周期起始点时推荐质量大于周期中间点的推荐质量，周期中间点的推荐质量大于周期末尾点的推荐质量，说明距离上一次离线模块更新时间越远，推荐效果越差，反应出了递增数据对推荐的影响较大. 当 μ=1.0 时，模型仅用离线模块做推荐(等同于传统推荐)，当 0<μ<1.0 时，是在线模块和离线模块结合的推荐，由表 2 可知融合了在线与离线模块的推荐质量好于传统推荐，体现出了及时的利用系统更新推荐期间产生的新数据来能提高推荐系统的推荐质量，也进一步说明本文基于数据递增式推荐的有效性. 为了说明模型的稳定性以及分析参数 μ 对推荐模型的影响，取每种算法在不同 μ 值下 3 组实验结果的 $F_{measure}$ 均值做分析，可以得到下图 3:

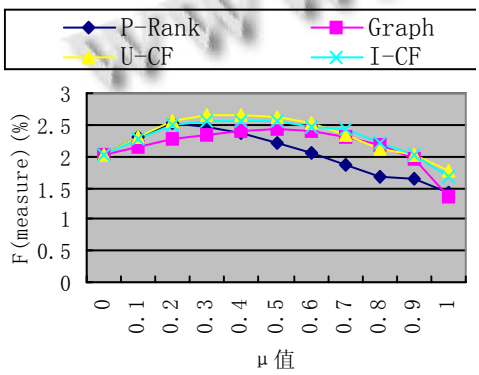

图 3 不同 μ 值在各组数据集上 $F_{measure}$ 的平均值结果

由图 3 可知，采用 4 种不同的离线算法进行实验，均有相同的趋势，说明了模型的稳定性. 当公式(3)中权重因子 μ 在 0.2 到 0.4 之间时，各算法可以取得较好的推荐效果. 当 μ=0 时，是由在线模块单独做推荐，此刻的推荐没有考虑个性化因素，属于匿名推荐；而当 0<μ<1 时，是离线模块和在线模块的混合个性化推荐；当 μ=1 时是由离线模块单独做推荐. 由于电影具有较强的实时性，在本实验中匿名推荐的有不低于单独离线模块的个性化推荐效果，并且混合个性化推荐效果好于离线模块和在线模块的单独推荐. 整体来看，个性化混合推荐的推荐效果要优于传统的推荐，匿名推荐也有较好的推荐效果.

本文提出的 APRS 混合推荐模型，不仅是基于传统推荐算法的，而且充分利用了具有较高价值的实时新数据来做推荐，因此推荐效果优于传统推荐算法. 与此同时，模型还可以利用实时新数据预测物品流行趋势，给出较好的匿名推荐，增强了推荐系统的推荐能力.

3 结论与未来的工作

本文提出一种基于数据递增变化的匿名与个性化混合推荐模型，引入一种能较好追踪物品流行趋势的变量，介绍了一种高效的在线推荐模型. 实现了基于数据递增式更新推荐，提高了推荐的实时性，同时也实现了匿名推荐和个性化推荐一体化. 与传统的推荐模型相比，本文提出的模型有较好的推荐效果.

在未来的工作中，还需要在以下两个方面进行更深入的研究: (1) 如何根据不同的场景去选择合适的离线模块算法; (2) 实现分场景和分时段的灵活推荐.

参考文献

1 刘建国,周涛,汪秉宏.个性化推荐系统的研究进展.自然科学进展,2009,19(1):1–15.

2 Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing, 2003, 7(1): 76–80.

3 Balabanovi M, Shoham Y. Fab: content-based, collaborative recommendation. Communications of the ACM, 1997, 40(3): 66–72.

4 Park HS, Yoo JO, Cho SB. A context-aware music recommendation system using fuzzy bayesian networks with utility theory. Fuzzy Systems and Knowledge Discovery. Springer Berlin Heidelberg, 2006: 970–979.

5 Huang Z, Zeng DD, Chen H. Analyzing consumer-product graphs: Empirical findings and applications in recommender systems. Management Science, 2007, 53(7).

6 Zhang YC, Blattner M, Yu YK. Heat conduction process on community networks as a recommendation model. arXiv Preprint arXiv: 0803.2179, 2008.

7 Schein AI, Popescul A, Ungar LH, Pennock DM. Methods and metrics for cold-start recommendations. Proc. of the 25th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM. 2002. 253–260.

8 Yoshii K, Goto M, Komatani K, Ogata T, Okuno HG. An efficient hybrid music recommender system using an

incrementally trainable probabilistic generative model. IEEE Trans. on Audio, Speech, and Language Processing, 2008, 16(2): 435–447.

9 Heylighen F. Complexity and information overload in society: Why increasing efficiency leads to decreasing control. Technological Forecasting and Social Change, 2004: 1–19.

10 Soucek R, Moser K. Coping with information overload in email communication: Evaluation of a training intervention. Computers in Human Behavior, 2010, 26(6): 1458–1466.

11 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.

12 Knuchel JP, Stoianovic N. A learning-based hybrid approach for anonymous recommendation. E-Commerce Technology, 2006. The 8th IEEE International Conference on and Enterprise Computing, E-Commerce, and E-Services, The 3rd IEEE International Conference on. IEEE, 2006. 33–36.

13 Stritt M, Tso KHL, Schmidt-Thieme L. Attribute aware anonymous recommender systems. Advances in Data Analysis. Springer Berlin Heidelberg, 2007. 497–504.

14 Luo X, Xia Y, Zhu Q. Incremental collaborative filtering recommender based on regularized matrix factorization. Knowledge-Based Systems, 2012, 27: 271–280.