软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(3):454−464 [doi: 10.3724/SP.J.1001.2013.04238]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于联合概率矩阵分解的上下文广告推荐算法*

涂丹丹, 舒承椿, 余海燕

(中国科学院 计算技术研究所,北京 100190)

通讯作者: 涂丹丹, E-mail: tudandan@software.ict.ac.cn

摘 要: 上下文广告与用户兴趣及网页内容相匹配,可增强用户体验并提高广告点击率.而广告收益与广告点击率直接相关,准确预测广告点击率是提高上下文广告收益的关键.目前,上下文广告推荐面临如下问题:(1) 网页数量及用户数量规模很大;(2) 历史广告点击数据十分稀疏,导致点击率预测准确率低.针对上述问题,提出一种基于联合概率矩阵分解的因子模型 AdRec,它结合用户、广告和网页三者信息进行广告推荐,以解决数据稀疏时点击率预测准确率低的问题.算法复杂度随着观测数据数量的增加呈线性增长,因此可应用于大规模数据.

关键词: 推荐算法;联合概率矩阵分解;上下文广告;准确率;数据稀疏

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 涂丹丹,舒承椿,余海燕.基于联合概率矩阵分解的上下文广告推荐算法.软件学报,2013,24(3):454−464.http://www.jos.org.cn/1000-9825/4238.htm

英文引用格式: Tu DD, Shu CC, Yu HY. Using unified probabilistic matrix factorization for contextual advertisement recommendation. Ruanjian Xuebao/Journal of Software, 2013,24(3):454−464 (in Chinese). http://www.jos.org.cn/1000-9825/4238.htm

Using Unified Probabilistic Matrix Factorization for Contextual Advertisement Recommendation

TU Dan-Dan, SHU Cheng-Chun, YU Hai-Yan

(Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100190, China)

Corresponding author: TU Dan-Dan, E-mail: tudandan@software.ict.ac.cn

Abstract: Combining user interests with visited web page contents to perform contextual advertising enhances the user experience and adds more ad clicks, increasing revenue. The key issue is to improve the prediction accuracy of click rates for advertisements. The crucial challenges of the advertisement recommendation algorithm are the scalability on large number of users and web page contents, and the low prediction accuracy resulting from data sparsity. When data is large and sparse, the accuracy and efficiency of the traditional recommendation algorithms is poor. This paper proposes a factor model called AdRec. Based on the Unified Probability Matrix Factorization (UPMF), the model addresses the data sparsity problem by combining features of users, advertisements and web page contents to predict the click rate with higher accuracy. In addition, the computational complexity of our algorithm is linear with respect to the number of observed data, and scalable to very large datasets.

Key words: recommendation algorithm; unified probabilistic matrix factorization; contextual advertising; accuracy; sparse data

在线广告(online advertising)已经成为三大主要广告投放方式(电视、报纸、互联网)之一.互联网在线广告收益与点击率密切相关,增加广告点击率是提高广告收益的有效途径之一.研究表明,网页上投放的广告与网页内容越相关,广告被点击的概率越高[1].此外,推荐与用户兴趣相关的广告也可以增加广告的点击率[2].也就是说,

* 基金项目: 国家自然科学基金(60873243)
收稿时间: 2011-04-01; 定稿时间: 2012-04-01

向用户推荐其感兴趣且与浏览网页内容相关的广告,可以提高广告的点击率.

本文主要研究上下文广告推荐(contextual advertising,简称 CA)问题,这是一种在被访问的网页上向用户推荐与网页内容相关广告的方法.通常,广告主为每一次广告点击支付一定金额,广告收益与广告点击次数成正比.上下文广告推荐的关键问题是预测广告在特定网页上被用户点击的概率.上下文广告推荐面临如下挑战:

(1) 网页数量及用户数量规模很大;

(2) 广告点击数据十分稀疏.

目前,上下文广告领域相关工作主要分为两类:一类是基于广告与网页内容相关度信息,以及用户点击反馈信息计算广告与网页匹配度[3−6];另一类将上下文广告问题转化为网页内容(查询)在待投放广告集合(文档集)中的信息检索问题[7].传统算法存在以下缺陷:

(1) 仅考虑两两关系.例如,仅了考虑广告和网页内容之间的相似度,忽略了用户的其他信息,算法缺乏灵活性;

(2) 广告点击数据越稀疏,推荐准确率越低.在实际应用中,由于长尾特性导致的数据稀疏性,使得这些推荐算法的准确度降低.

本文提出一种基于联合概率矩阵分解(unified probabilistic matrix factorization,简称 UPMF)的上下文广告推荐的新算法 AdRec.该算法结合与广告点击相关的多方面信息进行广告推荐:用户浏览网页信息,用户点击广告信息,网页与广告内容相似性和广告投放在目标网页上的点击率.其中,用户浏览网页信息与用户点击广告信息矩阵体现用户兴趣,而网页与广告内容相关性和广告投放在目标网页上的点击率信息矩阵用于计算广告与网页关联度.本文采用正则化(regularization)的 UPMF 方法,同时最优化分解 3 个矩阵,求解用户、广告、网页内容的隐含特征向量,并以此进行广告点击率的预测.

实验结果表明,AdRec 算法具有如下优点:

(1) 当数据稀疏时,采用正则化的 UPMF 分解,广告推荐的准确度较高;

(2) 结合丰富的用户浏览网页及点击广告历史信息增加了算法灵活性;

(3) 经算法复杂度分析,此算法可应用于大规模数据.

1 相关工作

上下文广告目前支撑大部分的网络生态系统.用户体验和广告收益(由网站发布者和广告网络共享)依赖于投放的广告与用户兴趣及网页内容相关度.目前,上下文广告相关工作主要分为两类:一类是基于广告与网页相关度与用户点击反馈计算广告与网页匹配程度[3−6];另一类将上下文广告问题转化为文件检索问题[7],广告类比于文档,网页内容类比于查询关键词.

文献[3−6]将广告和网页内容映射到相同的向量空间,广告和网页内容分别用特征向量表示.将广告与网页匹配问题转化为寻找与网页向量最相似的广告向量问题.文献[3−5]中各方法均根据计算网页向量与广告向量相似度衡量网页与广告匹配程度,但均未结合广告点击数据.文献[6]提出一种将相关度与点击反馈相结合以预测上下文广告点击率的方法,基于 logistic 回归模型,将点击反馈与决定相关度的网页和广告以语义信息相结合,学习得到更多的参数以扩展广告-网页打分模型.文中提出了多个直观模型,根据网页及广告中共同出现的关键词估计点击率,其中最复杂的模型不仅考虑了关键词的位置(标题、主体等),同时也考虑了关键词的同义词、关键词的 tf-idf 值及相关度值等.文献[7]中的关键词抽取系统,利用各种特征确定网页中关键词的重要性以用于广告匹配.系统的训练集为人工标注过的网页集合.学习算法考虑基于 tf-idf、HTML 元数据、查询日志等特征以计算词组重要性.

而传统的推荐算法主要分为两类:基于内容过滤推荐算法(CBF)[8,9]和基于协同过滤推荐算法(CF)[10−17].CBF 算法主要利用信息检索或信息过滤技术,根据推荐项目(item)的内容信息和用户配置文件的相关性向目标用户推荐相关项目.CF 推荐主要分为两类:基于记忆的(memory-based)方法[10−13]和基于模型的(model-based)方法[14,15].基于内容过滤推荐算法通常存在无法灵活结合多方面有用信息(例如用户兴趣等),而协同过滤算法依

赖于显示或隐式评分数据.推荐算法通常都面临着冷启动(如何对新用户进行推荐和如何推荐新项目给用户)、数据稀疏性、算法可扩展性等问题.

近年来,矩阵分解算法(matrix factorization,简称 MF)已逐渐应用于推荐系统[16−18],传统的矩阵分解算法有:奇异值分解(singular value decomposition,简称 SVD)[16]、非负矩阵分解(non-negative matrix factorization,简称 NMF)[17]、概率矩阵分解(probabilistic matrix factorization,简称 PMF)[18]等.这些算法的共同点是,通过将一个高维的矩阵分解为两个或多个低维矩阵的乘积实现维度规约,方便于在一个低维空间研究高维数据的性质.但 SVD 允许分解后结果出现负值,从计算角度看这是正确的,但就应用角度看负值是没有实际意义的.NMF 与 SVD 的不同之处在于,它将原始给定的非负矩阵近似分解为两个非负矩阵的乘积,即 NMF 保证分解所得矩阵的每一元素都是正值.NMF 分解具有局部性、部分表达等优于 SVD 算法的独特性质,具有直观的物理含义.Salakhutdinov 和 Mnih[18]于 2008 年在 NIPS 上提出概率矩阵分解(probabilistic matrix factorization,简称 PMF)算法,探讨了 MF 更深层的概率解释.

但是,以上矩阵分解方法只能结合两方面信息进行两维分解.文献[20]提出基于联合概率矩阵分解(UPMF)方法,并把该方法应用于广告推荐领域.实验结果表明,此方法比传统的协同过滤方法的推荐效果更好.本文把 UPMF 方法首次应用于上下文广告推荐,它结合三方面的信息进行矩阵分解.实验结果表明,该算法在数据稀疏时有更好的准确率,方便利用多方面信息,而且复杂度不高,适合处理大规模数据.

2 问题的定义

为了便于形式化,本文的符号标记见表 1.

Table 1 Notation

表 1 符号表

符号	解释
$US=\{u_1,u_2,\ldots,u_m\}$	用户集合,共有 m 个用户
$WS=\{w_1,w_2,\ldots,w_n\}$	网页集合,共有 n 个网页
$AS=\{a_1,a_2,\ldots,a_o\}$	广告集合,共有 o 个广告
$U\in\mathbb{R}^{l\times m}$	用户潜在特征矩阵
$W\in\mathbb{R}^{l\times n}$	网页潜在特征矩阵
$A\in\mathbb{R}^{l\times o}$	广告潜在特征矩阵
$l\in\mathbb{R}$	潜在特征空间维数
$B=\{b_{ij}\},B\in\mathbb{R}^{m\times n}$	B 为用户-网页点击矩阵
$C=\{c_{ik}\},C\in\mathbb{R}^{m\times o}$	C 为用户-广告点击矩阵
$R=\{r_{jk}\},R\in\mathbb{R}^{n\times o}$	R 为网页-广告关联度矩阵

上下文广告推荐是在用户访问的网页上推荐与网页内容相关的广告.为了最大限度提高广告收益,推荐算法的关键问题是尽量推荐点击率高的广告.而现有工作仅使用有限信息预测广告点击率,例如,利用网页内容与广告相关度信息.由于长尾特性导致广告点击数据稀疏,从而使得推荐算法预测准确率不高.

为此,本文试图结合与广告点击相关的多方面信息预测广告点击率,主要是用户访问网页信息、用户点击广告信息和广告与网页关联度信息.当用户浏览网页时,向目标用户推荐符合其兴趣爱好且与网页内容相关的高点击率广告.

通常,用户访问网页信息、用户点击广告信息和广告与网页关联度信息分别表示成用户-网页访问矩阵 B、用户-广告点击矩阵 C 和广告-网页关联度矩阵 R.

用户浏览网页时点击广告的概率由用户对广告感兴趣程度、用户对于网页感兴趣程度及广告与网页的关联度决定.为了减少噪声数据的影响,广告点击概率的计算采用用户、广告、网页的隐含特征向量计算得到.具体地,用户对网页的感兴趣程度由用户隐含特征向量与网页隐含特征向量的内积得到,用户对广告的感兴趣程度由用户隐含特征向量与广告隐含特征向量的内积得到,广告与网页的关联程度由广告隐含特征向量与网页

隐含特征向量的内积得到.

形式化地,用户 u_i 浏览网页 w_j 时,点击广告 a_k 的概率用实数 y_{u_i,w_j,a_k} 表示,定义如下:

$$y_{u_i,w_j,a_k} := h(U_i^T W_j, U_i^T A_k, W_j^T A_k) \tag{1}$$

其中,U_i 为用户 u_i 的隐含特征向量,W_j 为网页 w_j 隐含特征向量,A_k 为广告 a_k 的隐含特征向量;$U_i^T W_j, U_i^T A_k, W_j^T A_k$ 用于计算用户 u_i 对网页 w_j 的感兴趣程度、用户 u_i 对广告 a_k 的感兴趣程度及广告 a_k 与网页 w_j 的关联程度;$h(\cdot)$是参数为 $U_i^T W_j, U_i^T A_k$ 和 $W_j^T A_k$ 的函数.

当用户 u 正在访问网页 w 时,上下文广告的 Top-N 推荐列表可定义如下:

$$Top(u,w,N) := \arg\max_{a_k \in AS}^{N} y_{u_i,w_j,a_k} \tag{2}$$

3 AdRec 模型框架

本文提出一种基于联合概率矩阵分解(UPMF)的上下文广告推荐算法 AdRec,该算法主要由以下 3 个部分组成:

(1) 求解隐含特征向量.该算法以最大化联合后验概率为目标函数,基于梯度下降法方法学习得到用户隐含特征向量、网页隐含特征向量和广告隐含特征向量;

(2) 对给定的用户和网页计算广告集合中广告点击率.利用逻辑斯蒂函数(logistic function),将用户对广告感兴趣程度、用户对网页感兴趣程度及广告与网页关联度的线性组合 $\beta \cdot U_i^T W_j + \gamma \cdot U_i^T A_k + (1-\beta-\gamma) W_j^T A_k$ 映射到[0,1],最终计算得到的结果为广告点击率;

(3) 推荐 Top-N 广告.根据估计的广告点击概率对广告排序,在给定网页上推荐排前 N 名的广告.

本节首先介绍如何计算获得用户-网页访问矩阵、用户-广告点击矩阵、广告-网页关联度矩阵,最后详细介绍求解用户隐含特征向量、网页隐含特征向量和广告隐含向量矩阵的 UPMF 方法.

3.1 用户-网页访问矩阵

表 1 中,B 表示 m 个用户访问 n 个网页的用户-网页访问矩阵.B 中的元素 $b_{ij}(b_{ij} \in [0,1])$表示用户 u_i 对于网页 w_j 感兴趣程度.显然,用户浏览网页的次数越多,表明用户对此网页内容越感兴趣.b_{ij} 可由公式(3)计算得到:

$$b_{ij}=g(f(u_i,w_j)) \tag{3}$$

其中,$g(\cdot)$是逻辑斯蒂函数,用于归一化;$f(u_i,w_j)$表示用户 u_i 浏览网 w_j 的次数.

3.2 用户-广告点击矩阵

表 1 中,C 表示用户-广告点击矩阵,矩阵中元素 c_{ik} 表示用户 u_i 对广告 a_k 的感兴趣程度.很显然,用户点击广告,表明用户对此广告感兴趣.c_{ik} 由公式(4)计算得到:

$$c_{ik}=g(f(u_i,a_k)) \tag{4}$$

其中,$g(\cdot)$是逻辑斯蒂函数,用于归一化;$f(u_i,a_k)$表示用户 u_i 点击广告 a_k 的次数.

3.3 广告-网页关联度矩阵

表 1 中,R 表示广告-网页关联度矩阵.R 中元素 r_{jk} 表示网页 w_j 与广告 a_k 的关联度.同一广告在不同网页上显示时具有不同的点击率,并且,广告和网页内容越相关,广告被点击的可能性越大.本文将广告与网页内容相关性以及广告在网页上的点击率组合,计算得到网页与广告的关联度.将广告与网页内容相关性以及广告点击率结合可以增加广告与网页关联度计算的准确率.

r_{jk} 计算公式如下:

$$r_{jk}=\alpha \cdot d_{jk}+(1-\alpha) \cdot h_{jk} \tag{5}$$

其中,d_{jk} 为网页 w_j 与广告 a_k 的相似度,由概率潜在语义分析(probabilistic latent semantic analysis,简称 PLSA)方法计算得到;h_{jk} 为广告 a_k 投放在网页 w_j 上的点击率,等于广告 a_k 投放在网页 w_j 上时被点击的次数除以广告 a_k

在网页 w_j 上总的投放次数.

3.4 AdRec模型

用户对网页的访问历史和对广告的点击历史均能反映用户的兴趣及偏好,而广告点击率与用户兴趣及广告与网页相关度密切相关.本文提出一种基于 UPMF 的 AdRec 模型,将用户兴趣以及网页与广告关联度信息相结合.AdRec 模型的图形表示如图 1 所示,用户访问网页与用户点击广告信息共享用户隐含特征向量 U_i,广告点击信息与广告与网页关联度信息共享广告隐含特征向量 A_k.

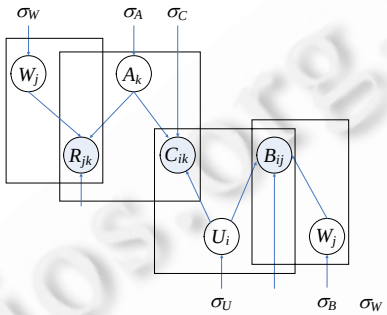


Fig.1 AdRec graphical model

图 1 AdRec 图模型

AdRec 模型基于以下假设:

(1) 假设 U_i,W_j,A_k 先验服从正态分布且相互独立,即:

$$p(U\mid\sigma_U^2)=\prod_{i=1}^{m}N(U_i\mid 0,\sigma_U^2 I) \tag{6}$$

$$p(W\mid\sigma_W^2)=\prod_{j=1}^{n}N(W_j\mid 0,\sigma_W^2 I) \tag{7}$$

$$p(A\mid\sigma_A^2)=\prod_{k=1}^{o}N(A_k\mid 0,\sigma_A^2 I) \tag{8}$$

(2) 在给定用户 u_i、网页 w_j 的隐含特征向量(维数为 l)U_i,W_j 后,用户 u_i 对网页 w_j 的感兴趣程度 b_{ij} 满足均值为 $g(U_i^TW_j)$,方差为 σ_B^2 的正态分布且相互独立.用户-网页浏览矩阵 B 的条件概率分布如下:

$$p(B\mid U,W,\sigma_B^2)=\prod_{i=1}^{m}\prod_{j=1}^{n}[N(b_{ij}\mid g(U_i^TW_j),\sigma_B^2)]^{I_{ij}^B} \tag{9}$$

其中,I_{ij}^B 是指示函数,当用户 u_i 访问过网页 w_j,$I_{ij}^B=1$;否则,$I_{ij}^B=1$.$g(x)=1/(1+e^{-x})$ 是逻辑斯蒂函数,将 $U_i^TW_j$ 值映射到[0,1].UPMF 引入概率思想,矩阵中各元素的值应属于[0,1].

(3) 用户 u_i 对广告 a_k 的感兴趣程度 c_{ik} 满足均值为 $g(U_i^TA_k)$,方差为 σ_C^2 的正态分布且相互独立.用户-广告点击矩阵 C 的条件概率分布如下:

$$p(C\mid U,A,\sigma_C^2)=\prod_{i=1}^{m}\prod_{k=1}^{o}[N(c_{ik}\mid g(U_i^TA_k),\sigma_C^2)]^{I_{ik}^C} \tag{10}$$

其中,I_{ik}^C 是指示函数,当用户 u_i 点击过广告 a_k 时,$I_{ik}^C=1$;否则,$I_{ik}^C=0$.

(4) 网页 w_j 与广告 a_k 的关联度 r_{jk} 满足均值为 $g(W_j^TA_k)$,方差为 σ_R^2 的正态分布且相互独立.广告-网页关联度矩阵 R 的条件概率分布如下:

$$p(R\mid W,A,\sigma_R^2)=\prod_{j=1}^{n}\prod_{k=1}^{o}[N(r_{jk}\mid g(W_j^TA_k),\sigma_R^2)]^{I_{jk}^R} \tag{11}$$

其中, I_{jk}^R 是指示函数,当网页 w_j 与广告 a_k 有关联(即 r_{jk}>0)时, $I_{jk}^R=1$;否则, $I_{jk}^R=0$.

由图 1 可以推导出 U,W,A 的后验分布函数.后验分布函数的 log 函数见公式(12):

$$\begin{aligned}\ln p(U,W,A\mid B,C,R,\sigma_A^2,\sigma_W^2,\sigma_U^2,\sigma_R^2,\sigma_B^2,\sigma_C^2)=&-\frac{1}{2\sigma_B^2}\sum_{i=1}^{m}\sum_{j=1}^{n}I_{ij}^B(b_{ij}-g(U_i^TW_j))^2-\\ &\frac{1}{2\sigma_C^2}\sum_{i=1}^{m}\sum_{k=1}^{o}I_{ik}^C(c_{ik}-g(U_i^TA_k))^2-\frac{1}{2\sigma_R^2}\sum_{j=1}^{n}\sum_{k=1}^{o}I_{jk}^R(r_{jk}-g(W_j^TA_k))^2-\frac{1}{2\sigma_U^2}\sum_{i=1}^{m}U_i^TU_i-\\ &\frac{1}{2\sigma_W^2}\sum_{j=1}^{n}W_j^TW_j-\frac{1}{2\sigma_A^2}\sum_{k=1}^{o}A_k^TA_k-\sum_{i=1}^{m}\sum_{j=1}^{n}I_{ij}^B\ln\sigma_B-\sum_{i=1}^{m}\sum_{k=1}^{o}I_{ik}^C\ln\sigma_C-\\ &\sum_{j=1}^{n}\sum_{k=1}^{o}I_{jk}^R\ln\sigma_R-l\cdot\sum_{i=1}^{m}\ln\sigma_U-l\cdot\sum_{j=1}^{n}\ln\sigma_W-l\cdot\sum_{k=1}^{o}\ln\sigma_A+\mathcal{C}\end{aligned}\tag{12}$$

其中,$\mathcal{C}$是常量.最大化公式(12)可视为无约束优化问题,最小化公式(13)等价于最大化公式(12):

$$\begin{aligned}E(U,W,A,B,C,R)=&\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}I_{ij}^B(b_{ij}-g(U_i^TW_j))^2+\frac{\theta_C}{2}\sum_{i=1}^{m}\sum_{k=1}^{o}I_{ik}^C(c_{ik}-g(U_i^TA_k))^2+\\ &\frac{\theta_R}{2}\sum_{j=1}^{n}\sum_{k=1}^{o}I_{jk}^R(r_{jk}-g(W_j^TA_k))^2+\frac{\theta_U}{2}\sum_{i=1}^{m}U_i^TU_i+\frac{\theta_W}{2}\sum_{j=1}^{n}W_j^TW_j-\frac{\theta_A}{2}\sum_{k=1}^{o}A_k^TA_k\end{aligned}\tag{13}$$

其中, $\theta_C=\frac{\sigma_B^2}{\sigma_C^2},\theta_R=\frac{\sigma_B^2}{\sigma_R^2},\theta_U=\frac{\sigma_B^2}{\sigma_U^2},\theta_W=\frac{\sigma_B^2}{\sigma_W^2},\theta_A=\frac{\sigma_B^2}{\sigma_A^2}$.公式(13)的局部最小值可由梯度下降法求得.参数 U_i,W_j,A_k 的梯度下降公式见公式(14)~公式(16):

$$\frac{\partial E}{\partial U_i}=\sum_{j=1}^{n}I_{ij}^B(g(U_i^TW_j)-b_{ij})g'(U_i^TW_j)W_j+\theta_C\sum_{k=1}^{o}I_{ik}^C(g(U_i^TA_k)-c_{ik})g'(U_i^TA_k)A_k+\theta_UU_i\tag{14}$$

$$\frac{\partial E}{\partial W_j}=\sum_{i=1}^{m}I_{ij}^B(g(U_i^TW_j)-b_{ij})g'(U_i^TW_j)U_i+\theta_R\sum_{k=1}^{o}I_{jk}^R(g(W_j^TA_k)-r_{jk})g'(W_j^TA_k)A_k+\theta_WW_j\tag{15}$$

$$\frac{\partial E}{\partial A_k}=\sum_{i=1}^{m}I_{ik}^C(g(U_i^TA_k)-c_{ik})g'(U_i^TA_k)U_i+\theta_D\sum_{j=1}^{n}I_{jk}^R(g(W_j^TA_k)-r_{jk})g'(W_j^TA_k)W_j+\theta_AA_j\tag{16}$$

3.5 算法时间复杂度分析

梯度下降法的计算开销主要来自于目标函数 E 和对应的梯度下降公式.由于矩阵 B,C,R 很稀疏,计算公式(12)中目标函数时间复杂度为 $O(n_Bl+n_Cl+n_Rl)$,其中,n_B,n_C,n_R 分别表示矩阵 B,C,R 中非零元素个数.同理可以推导出计算公式(14)~公式(16)的时间复杂度.因此,每次迭代的总时间复杂度为 $O(n_Bl+n_Cl+n_Rl)$,即算法时间复杂度随 3 个稀疏矩阵中观测数据数量增加成线性增长,因此算法可应用于大规模数据.

4 实验结果及分析

4.1 实验设计

(1) 实验数据

本文实验的数据集来自某在线广告系统从 2010 年 8 月~10 月之间 3 个月的实际运行数据.原始数据集由 40 万用户的互联网访问日志、185 个广告的点击数据和广告投放数据 3 个部分组成.

借鉴文献[15]中的方法,我们从实验数据中随机抽取一部分为训练集 S_{train},另一部分作为测试集 S_{test}.我们分别从实验数据中抽取 99%,80%,50%,20%,10%作为训练数据,在数据稀疏程度不同时,比较算法的效果.

由于网页数量巨大,我们首先根据域名分类数据将网页分为约 3 000 类.通过数据预处理得到 4 个矩阵:用户-网页类浏览矩阵、用户-广告点击矩阵、网页类-广告内容相关性矩阵、网页类-广告同时出现时点击率矩阵.

(2) 评价指标

我们采用 F-measure 评价算法预测准确率.F-Measure 综合了信息检索领域中的查准率(precision)和查全率

(recall).*F*-measure 定义如下:

$$F(S_{test},N):=\frac{2\cdot Prec(S_{test},N)\cdot Rec(S_{test},N)}{Prec(S_{test},N)+Rec(S_{test},N)}=\frac{2}{\frac{1}{Prec(S_{test},N)}+\frac{1}{Rec(S_{test},N)}} \tag{17}$$

其中,

$$Prec(S_{test},N):=\frac{|\text{推荐广告集合}\cap\text{测试集中相同推荐情景下被点击的广告集合}|}{\text{单次推荐的广告个数}},$$

$$Rec(S_{test},N):=\frac{|\text{推荐广告集合}\cap\text{测试集中相同推荐情景下被点击的广告集合}|}{\text{测试集中相同推荐情景下被点击的广告个数}}.$$

从公式(17)可看出,*F*-measure 是查准率和查全率的倒数平均,这个指标并不掩盖查准率和查全率任何一方面特别的不足.*F*-measure 值越高,表明推荐算法的准确率越高.

(3) 实验环境

本实验硬件配置为 3 计算节点的集群系统,每个节点 CPU 主频 2.53GHz,主存 2GB,操作系统为 Linux (Ubuntu 9.10).实验数据的分析程序由 python 语言实现,运行在 3 个节点组成的 Hadoop 集群系统上.

(4) 实验设计

实验旨在解决以下 4 个问题:

(a) 隐含特征向量维数 l 对广告推荐准确率的影响;

(b) 模型参数 θ_C 和 θ_R 对推荐准确率的影响;

(c) AdRec 算法与现有的矩阵分解算法以及文献[18]中结合相关性与点击数据的上下文广告算法(简称 CRC)准确率比较;

(d) AdRec 算法时间效率分析.

在实验过程中,我们在训练集上尝试不同参数值,在测试集上实验得到 *F*-measure 值.经过反复测试我们发现,参数分别设为 $\theta_U=\theta_W=\theta_A=0.001$,$\theta_C=0.5$,$\theta_R=10$ 时,算法的效果最优.以下实验中若非特别说明,上述 5 个参数均设为最优值.实验中每次推荐 $N=8$ 个广告.

4.2 参数*l*对广告推荐准确率的影响

图 2~图 6 分别表示从实验数据中抽取 99%,80%,50%,20%,10%作为训练数据时,隐含特征向量维数 l 对矩阵分解算法准确率的影响.从图中可看出,随着隐含特征向量维数 l 的增加,本文提出的 AdRec 算法和其余 3 种矩阵分解算法(SVD,NMF,PMF)的 *F*-measure 值均有所提高.也就是说,增加隐含特征向量维数可以提高矩阵分解算法的准确率.但是,在增加隐含特征向量维数的同时也会减低算法的计算效率.仔细观察这 5 个图我们发现,l 取值为[0,30]时,随着隐含特征向量维数的增加,算法的准确率测量指标 *F*-measure 值增加约 0.07;当 $30<l<50$ 时,随着隐含特征向量维数的增加,算法的准确率测量指标 *F*-measure 值增加不到 0.01.对准确率和时间效率进行权衡,后面的实验我们分别取 $l=20$ 或 $l=30$ 时的实验结果.

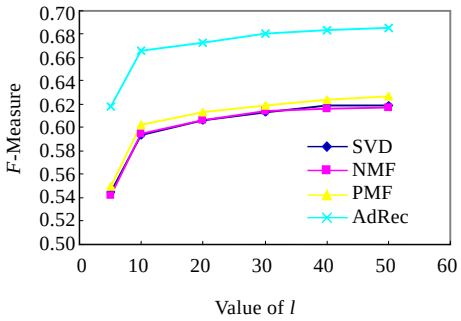


Fig.2 l impacts on accuracy of algorithm (99% as training data)

图 2 99%作训练数据时参数 l 对算法准确率影响

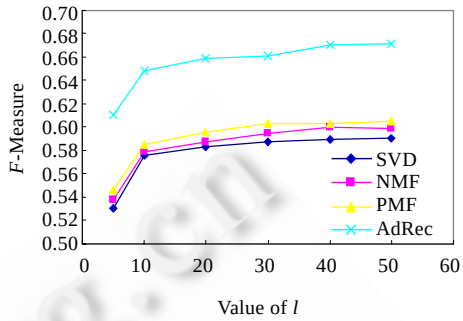


Fig.3 l impacts on accuracy of algorithm (80% as training data)

图 3 80%作训练数据时参数 l 对算法准确率影响

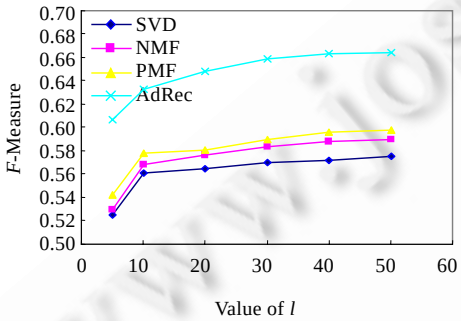


Fig.4 l impacts on accuracy of algorithm (50% as training data)

图 4 50%作训练数据时参数 l 对算法准确率影响

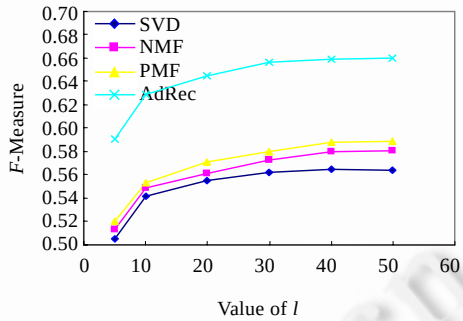


Fig.5 l impacts on Accuracy of Algorithm (20% as training data)

图 5 20%作训练数据时参数 l 对算法准确率影响

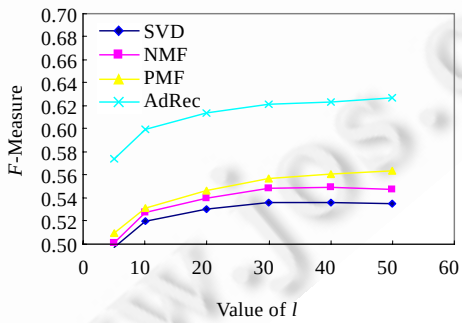


Fig.6 l impacts on accuracy of algorithm (10% as training data)

图 6 10%作训练数据时参数 l 对算法准确率影响

4.3 参数θ_C和θ_R对推荐准确率的影响

AdRec模型不仅利用用户点击广告及浏览网页的信息进行广告推荐,还可以基于联合矩阵分解方法UPMF将用户-网页浏览矩阵、用户-广告点击矩阵、广告-网页关联度矩阵结合起来.参数θ_C决定了用户-广告点击矩阵对算法效果的影响,而参数θ_R决定了广告-网页关联度矩阵对算法效果的影响.当θ_C和θ_R均设置为0时,意味着我们仅利用用户浏览网页信息.当θ_C或θ_R设为+∞时,意味着我们仅利用用户点击广告信息或广告-网页关联度信息.

我们分别测试两个参数对 AdRec 算法的影响.图 7 和图 8 显示了 l=20 和 l=30 时,参数 θ_C 对算法效果测量指标 F-measure 的影响.当测试参数 θ_C 对算法的影响时,我们将其余参数分别设为:$\theta_U=\theta_W=\theta_A$=0.001,$\theta_R$=10.图 9 和图 10 显示了 l=20 和 l=30 时,参数 θ_R 对算法效果测量指标 F-measure 的影响.当测试参数 θ_R 对算法的影响时,我们将其余参数分别设为 $\theta_U=\theta_W=\theta_A$=0.001,$\theta_C$=0.5.

仔细观察图 7~图 10 得出以下结论:参数 θ_C 和 θ_R 的值对于算法准确率的影响是较大的.也就是说,用户点击广告信息及广告-网页关联度信息可以提高算法的预测准确率.进一步观察发现,随着参数值的增加,算法的 F-measure 值先增加,但当 $\theta_C\in$[0.1,1],$\theta_R\in$[5,15]或 θ_R 大于某阈值时,F-measure 值逐渐减小.原因在于:仅利用用户点击广告信息或仅利用广告-网页关联度信息时,算法效果无法超越结合两方面信息所得的结果.当 $\theta_C\in$[0.1,1],$\theta_R\in$[5,15]时,算法的准确率最高.最优参数的选择范围较广,表明 AdRec 模型更容易训练.

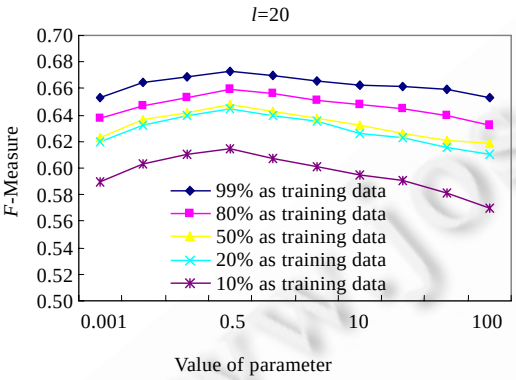


Fig.7 θ_C impacts on accuracy of algorithm (l=20)

图 7 l=20 时,参数 θ_C 对算法准确率影响

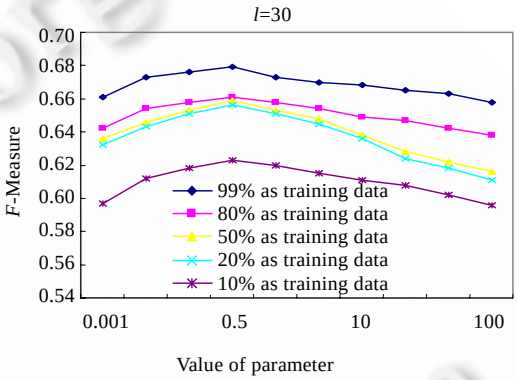


Fig.8 θ_C impacts on accuracy of algorithm (l=30)

图 8 l=30 时,参数 θ_C 对算法准确率影响

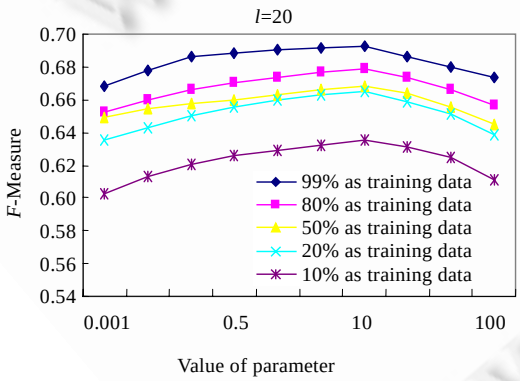


Fig.9 θ_R Impacts on Accuracy of Algorithm (l=20)

图 9 l=20 时参数 θ_R 对算法准确率影响

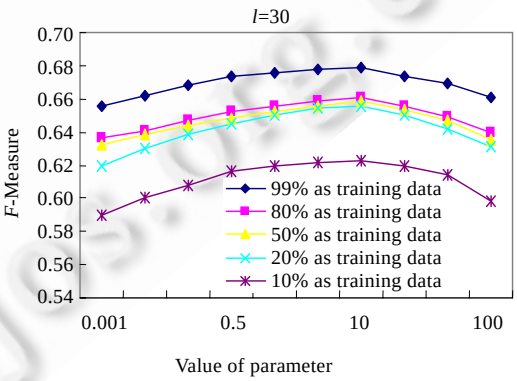


Fig.10 θ_R Impacts on Accuracy of Algorithm (l=30)

图 10 l=30 时参数 θ_R 对算法准确率影响

4.4 推荐质量分析

主要实验目的是比较本文提出的 AdRec 算法与奇异值分解(singular value decomposition,简称 SVD)、非负矩阵分解(nonnegative matrix factorization,简称 NMF)、概率矩阵分解(probabilistic matrix factorization,简称 PMF)算法以及 CRC 算法的预测效果.

我们利用不同稀疏度训练数据(99%,80%,50%,20%,10%)对以上算法进行实验.表 2 比较了训练数据稀疏程度不同时,5 种算法的 F-measure 值.见表 2,AdRec 算法推荐准确率比传统矩阵分解算法 SVD,NMF,PMF 以及 CRC 算法高.在数据稀疏的情形下,AdRec 算法推荐效果提高显著,其 F-measure 值比其他 4 种算法提高 3.8%~8.4%.

Table 2 Accuracy comparison for five recommendation algorithms

表 2 5 种算法推荐效果比较

训练数据集 (%)	CRC	*l*=20			
		SVD	NMF	PMF	AdRec
99	0.636	0.602	0.598	0.609	0.673
80	0.621	0.583	0.587	0.595	0.659
50	0.600	0.565	0.576	0.581	0.648
20	0.585	0.555	0.561	0.571	0.634
10	0.541	0.531	0.540	0.547	0.615

4.5 算法效率分析

算法时间复杂度分析表明,随着观测数据数量的增加,AdRec 算法的时间复杂度呈线性增长.因此,此算法可应用于大规模数据.虽然 AdRec 算法仅利用简单的梯度下降法求解目标函数,却十分高效.算法的收敛标准为公式(13)的值变化小于 0.001.图 11 显示了隐含特征维数 *l*=20 时,AdRec 算法时间效率.*l*=30 时结果类似.如图 11 所示,当使用 99%数据作为训练集时,每次迭代耗时不超过 0.5s,我们的算法仅需要不到 1 500 次迭代即可收敛,整个模型训练过程仅需约 12 分钟.当使用 10%的数据作为训练集时,模型训练过程仅需约 3 分钟.

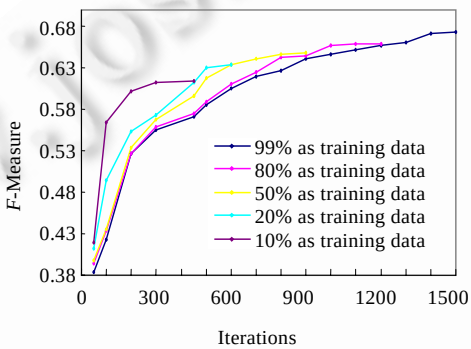


Fig.11 Time efficiency analysis for AdRec (*l*=20)

图 11 *l*=20 时,AdRec 时间效率分析

5 结束语

针对传统的上下文推荐算法通常对长尾特性的数据预测准确度不高的问题,本文提出了 AdRec 算法,利用 UPMF 方法进行上下文广告推荐.基于给用户推荐其感兴趣且与浏览网页内容相关的广告可提高广告点击率的假设,该方法结合用户浏览网页信息、用户点击广告信息和网页与广告关联度等信息.实验表明,AdRec 算法优于传统的上下文广告推荐算法.此外,算法复杂度分析表明,AdRec 算法可应用于大规模数据.在观测数据可表示成隐含特征的线性组合假设下,AdRec 使用隐含特征向量的内积之和来结合多方面信息,并使用逻辑斯蒂函数预测广告的点击概率.在下一步的研究工作中,我们将用高斯核或多项式核结合多个低维隐含特征向量,将其中任意两个特征向量的关系映射到非线性空间,从而进一步提高算法的性能.

References:

[1] Chatterjee P, Hoffman DL, Novak TP. Modeling the clickstream: Implications for Web-based advertising efforts. Marketing Science, 2003,22(4):520−541. [doi: 10.1287/mksc.22.4.520.24906]

[2] Wang C, Zhang P, Choi R, D'Eredita M. Understanding consumers' attitude toward advertising. In: Proc. of the 8th Americas Conf. on Information System. 2002. 1143−1148.

[3] Ribeiro-Neto B, Cristo M, Golgher PB, Moura ES. Impedance coupling in content-targeted advertising. In: Proc. of the SIGIR 2005. New York: ACM Press, 2005. 496−503. [doi: 10.1145/1076034.1076119]

[4] Lacerda A, Cristo M, Goncalves MA, Fan WG, Ziviani N, Ribeiro-Neto B. Learning to advertise. In: Proc. of the SIGIR 2006. New York: ACM Press, 2006. 549−556. [doi: 10.1145/1148170.1148265]

[5] Broder AZ, Fontoura M, Josifovski V, Riedel L. A semantic approach to contextual advertising. In: Proc. of the SIGIR. 2007. 559−566. [doi: 10.1145/1277741.1277837]

[6] Chakrabarti D, Agarwal D, Josifovski V. Contextual advertising by combining relevance with click feedback. In: Proc. of the 17th Int'l Conf. on World Wide Web (WWW 2008). Beijing: ACM Press, 2008. 417−426. [doi: 10.1145/1367497.1367554]

[7] Yih W, Goodman J, Carvalho VR. Finding advertising keywords on Web pages. In: Proc. of the 15th Int'l Conf. on World Wide Web (WWW 2006). New York: ACM Press, 2006. 213−222. [doi: 10.1145/1135777.1135813]

[8] Belkin N, Croft B. Information filtering and information retrieval. Communications of the ACM, 1992,35(12):29−37. [doi: 10.1145/138859.138861]

[9] Balabanovic M, Shoham Y. Fab: Content-based collaborative recommendation. Communications of the ACM, 1997,40(3):66−72. [doi: 10.1145/245108.245124]

[10] Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J. Grouplens: An open architecture for collaborative filtering of netnews. In: Proc. of the CSCW'94. 1994. [doi: 10.1145/192844.192905]

[11] Breese JS, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. In: Proc. of the Uncertainty in Artificial Intelligence (UAI'98). 1998.

[12] Deshpande M, Karypis G. Item-Based top-*N* recommendation. ACM Trans. on Information Systems, 2004,22(1):143−177. [doi: 10.1145/963770.963776]

[13] Sarwar B, Karypis G, Konstan J, Riedl J. Item-Based collaborative filtering recommendation algorithms. In: Proc. of the WWW 2001. 2001. 285−295. [doi: 10.1145/371920.372071]

[14] Hofmann T. Latent semantic models for collaborative filtering. ACM Trans. on Information Systems, 2004,22(1):89−115. [doi: 10.1145/963770.963774]

[15] Liu NN, Yang Q. Eigenrank: A ranking-oriented approach to collaborative filtering. In: Proc. of the SIGIR 2008. 2008. 83−90. [doi: 10.1145/1390334.1390351]

[16] Golub G, Kahan K. Calculating the singular values and pseudo-inverse of a matrix. Journal of the Society for Industrial and Applied Mathematics, 1965,2(2):205−224. [doi: 10.1137/0702016]

[17] Lee DD, Seung HS. Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems. 2000. 556−562.

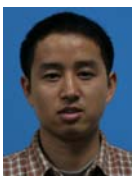
[18] Salakhutdinov R, Mnih A. Probabilistic matrix factorization. In: Advances in Neural Information Processing Systems. 2008,20(3): 451−432.

[19] Rendle S, Marinho LB, Nanopoulos A, Schmidt-Thieme L. Learning optimal ranking with tensor factorization for tag recommendation. In: Proc. of the 15th ACM SIGKDD. New York, 2009. 727−736. [doi: 10.1145/1557019.1557100]

[20] Ma H, Yang H, Lyu MR, King I. SoRec: Social recommendation using probabilistic matrix factorization. In: Proc. of the CIKM 2008. 2008. 931−940. [doi: 10.1145/1458082.1458205]

涂丹丹(1984－),女,江西南昌人,博士生,主要研究领域为个性化推荐,机器学习.
E-mail: tudandan@software.ict.ac.cn

余海燕(1974－),男,博士,副研究员,主要研究领域为云计算,服务计算,机器学习,数据挖掘.
E-mail: yuhaiyan@ict.ac.cn



舒承椿(1976－),男,博士,助理研究员,主要研究领域为云计算,机器学习.
E-mail: shuchengchun@software.ict.ac.cn