文章编号:1000-5641(2013)03-0015-11

基于概率图模型的互联网广告点击率预测

岳 昆, 王朝禄, 朱运磊, 武 浩, 刘惟一
(云南大学 信息学院计算机科学与工程系,昆明 650091)

摘要: 点击率预测可以提高用户对所展示互联网广告的满意度,支持广告的有效投放,是针对用户进行广告的个性化推荐的重要依据.对于没有历史点击记录的用户,仍需对其推荐广告,预测所推荐广告的点击率.针对这类用户,以贝叶斯网这一重要的概率图模型,作为不同用户之间广告搜索行为的相似性及其不确定性的表示和推理框架,通过对用户搜索广告的历史记录进行统计计算,构建反映用户间相似关系的贝叶斯网,进而基于概率推理机制,定量度量没有历史点击记录的用户与存在历史点击记录的用户之间的相似性,从而预测没有历史点击记录的用户对广告的点击率,为广告推荐提供依据.通过建立在KDD Cup 2012-Track 2的Tencent CA训练数据集上的实验,测试了方法的有效性.

关键词: 计算广告; 点击率; 个性化推荐; 贝叶斯网; 概率推理
中图分类号: TP311 **文献标识码**: A **DOI**:10.3969/j.issn.1000-5641.2013.03.002

Click-through rate prediction of online advertisements based on probabilistic graphical model

YUE Kun, WANG Chao-lu, ZHU Yun-lei, WU Hao, LIU Wei-yi
(*Department of Computer Science and Engineering, School of Information Science and Engineering, Yunnan University, Kunming* 650091, *China*)

Abstract: CTR (Click-Through Rate) prediction can be used to improve users' satisfaction with respect to the presented online advertisements (ads) and support effective advertising. CTR prediction is the basis for personalized recommendation of online ads. It is also necessary to recommend ads and predict their CTRs for the users that have no historical click-through records. In this paper, we adopted BN (Bayesian network), an important probabilistic graphical model, as the framework for representing and inferring the similarity and the corresponding uncertainty of the behaviors in ad search of different users. First, we constructed the BN to reflect the similarity between users by means of statistic computations on the historical records of user's ad search. Then, we measured the behavior similarity between the users with click-through records and those without records quantitatively based on the mechanism of BN's probabilistic infer-

收稿日期:2013-03
基金项目:国家自然科学基金(61063009, 61163003, 61232002);国家教育部博士点基金新教师类课题(20105301120001);云南省中青年学术与技术带头人后备人才培养计划(2012HB004);云南省教育厅科研基金重点项目(2011Z015)
通信作者:岳昆,男,博士、副教授,研究方向为数据与知识工程. E-mail: kyue@ynu.edu.cn.

ences. Consequently, we predicted the CTRs of ads with respect to the users without historical click-through records, in order to provide a metric for ad recommendation. We made experiments on the training data of Tencent CA from KDD Cup 2012-Track 2 and tested the effectiveness of our methods.

Key words: computing advertising; click-through rate; personalized recommendation; Bayesian network; probabilistic inference

0 引 言

计算广告[1]是根据给定的用户和网页内容,通过计算得到与之最匹配的广告并进行精准定向投放的一种广告投放机制.计算广告通过提高广告定向投放的精度,使广告从骚扰信息变为有用信息,并且通过推动第三方付费模式的发展使互联网广告的产业环境日臻完善.当前互联网广告的新变化,使得计算广告成为国内外研究的热点,但目前仍面临着广告投放的精确化和个性化等方面的问题.

互联网广告通常的运营模式为关键词竞拍,广告主支付获得这些关键词的费用,其主要付费方式为按点击付费(Pay Per Click,PPC).广告主的付费为每次点击费用(Cost Per Click,CPC)[1];广告的热门程度用点击率(CTR)描述;而广告媒介的收益则是 CTR×CPC.由于用户点击广告的可能性按广告的排放位置快速递减,最高可达 90%,广告媒介要获得最大的收益,就需要把 CTR 高的广告投放在靠前的位置,那么只有对 CTR 进行准确的预测,才能达到这一目的.这不仅可以提高广告媒介的收益,而且可以提高用户对所展示广告的满意程度.因此,CTR 预测是计算广告领域中的一个关键问题,也正是本文研究的目标所在.

近年来,CTR 预测被学术界广泛关注,已有许多的研究成果.例如:Regelson 等[2]提出基于分层聚类的方法,用以解决用户历史数据不足时的 CTR 预测问题;Agarwal 等[3]基于层次概念,先对广告进行分层,进而预测广告的 CTR;Richardson[4]和 Chakrabarti[5]分别提出了基于 Logistic 回归(Logistic Regression)的 CTR 预测方法.然而,上述工作仅针对广告本身来预测其 CTR,即认为 CTR 的预测只与广告本身的内容和用户查询内容有关,并没有考虑针对于不同用户的精确个性化推荐.

事实上,考虑广告的精确个性化推荐,CTR 的预测与用户和广告都有关系,基于行为定向(Behavioral Targeting, BT)通过获取用户的偏好,有针对性地为用户推荐广告,从而提高在线广告的 CTR[6-9],这也是一类有代表性的工作.这一思想为本文的研究提供了借鉴和参考.特别地,对于没有直接历史点击记录的用户,仍需要考虑对其推荐广告,并预测其对广告的 CTR,从而有效地为其推荐个性化的广告.当用户对特定广告没有点击记录时,如何通过分析该用户的历史行为来预测将要推荐的广告的 CTR? 在实际的互联网广告应用中,用户历史点击记录往往比较稀疏,如何在这样的情形下基于用户偏好来预测广告的 CTR? 前述基于 BT 提高 CTR 的方法仍未考虑不同用户之间在广告搜索行为方面的相似性,也没有解决不存在点击记录的用户对广告 CTR 预测的问题.

同时,我们注意到,作为 CTR 预测基础的广告点击或搜索行为,相应的用户历史记录,以及用户之间行为的相关性等,都广泛存在不确定性;而目前的工作不能较好地反映广告搜

索或点击等用户行为中所蕴含的用户相似性及其不确定性，并对其进行推理计算，因此未能将用户偏好及行为的相似性用于广告的 CTR 预测分析中.

作为一种重要的概率图模型，贝叶斯网(BN)[10,11]是不确定性知识表示和推理的有效框架，具有坚实的概率论理论基础和广泛的应用，基于 BN 可有效地描述随机变量或对象属性间的相关性和相互依赖. BN 以有向无环图(Directed Acyclic Graph，DAG)模型表示这种依赖关系，每个结点的条件概率参数表(Conditional Probability Table，CPT)定量地描述变量间的依赖关系. 近年来，基于概率计算的思想已被用于广告的 CTR 预测中，例如：Dembczyński 等[12]给出了基于极大似然假设的 CTR 预测方法；Graepel 等[13]提出了一种实现二分预测(Binary Prediction)的算法，称为在线贝叶斯概率回归(Online Bayesian Probability Regression)算法，用于赞助商搜索广告情形下的 CTR 预测；Chapelle 等[14]考虑到 CTR 预测的动态性要求，提出了基于动态贝叶斯网(Dynamic BN，DBN)的 CTR 预测模型，通过对用户的点击过程进行建模，分别估计出文档的观察相关性与实际相关性；文献[15]将 BN 与小世界网络相结合用于建立混合推荐系统. 但是，这些方法都没有涉及基于 BN 表示用户间的相似性，并进行演绎推断.

从前述分析不难看出，用户之间由于搜索广告而存在着行为上的相似性，并且这种相似性具有不确定性. 因此，本文仍以 BN 作为反映广告历史搜索行为中用户相似关系的基本框架，利用 BN 的概率推理算法定量地分析任意用户之间的相似性，从而针对特定的广告，获得没有历史点击记录的用户与存在历史点击记录的用户之间存在的广告搜索行为的相似性，进而预测没有点击记录的用户对广告的 CTR.

具体而言，本文的研究主要包括以下三方面.

(1) 用户相似模型的构建. 为了构建反映用户在广告搜索行为方面的相似关系，本文基于 BN 给出了用户相似模型的概念，称为用户贝叶斯网(User BN，UBN)，并针对 DAG 构建这一 BN 构建中的关键和难点，通过对用户搜索广告的历史记录进行统计计算，给出构建 UBN 的方法.

(2) UBN 的近似概率推理和 CTR 预测. 针对 CTR 预测问题，本文基于 Gibbs 采样[16-17]给出 UBN 的近似概率推理算法，并利用近似推理算法用来高效地发现相似用户，进而预测没有点击记录的用户对广告的 CTR.

(3) 实验测试. 基于 KDD CUP 2012-Track 2[18]中的测试数据，我们实现并测试了本文提出的 UBN 构建、近似推理及 CTR 预测方法. UBN 构建及概率推理的效率、CTR 预测结果准确性等方面的实验结果表明，本文基于概率图模型的 CTR 预测方法具有一定的可行性.

本文第 1 节给出 UBN 的定义及其构建方法；第 2 节给出 UBN 的近似推理算法和相应的 CTR 预测方法；第 3 节给出实验结果和性能分析；第 4 节总结全文并展望未来的工作.

1　用户相似模型

搜索引擎这一广告媒介通过广告主竞拍得到的关键词来标识广告，它把用户输入的广告查询内容作为一次广告搜索行为，搜索行为用关键词集来表达. 基于此，下面给出用户、广告搜索行为和关键词的概念.

定义 1　假设 $U=\{U_1, U_2, \cdots, U_n\}$，$K=\{K_1, K_2, \cdots, K_m\}$ 和 $B=\{B_1, B_2, \cdots, B_n\}$

分别为用户、搜索关键词和搜索行为的集合,其中 $B_i=\{B_{i1}, B_{i2}, \cdots, B_{il_i}\}$,$B_{ij}\in K$,$1\leqslant i\leqslant n$,$1\leqslant j\leqslant l_i$,$B_{ij}$ 表示用户 U_i 进行一次搜索时对应的关键词,称为搜索行为. 用户 U_i 的搜索内容为一个二元组 $C_i=\langle U_i, M_i\rangle$,$M_i=\{B_{ij}\mid i=1, 2, \cdots, l_i\}$,描述用户与关键词之间的关系.

例 1 用户 U_i 的搜索行为集合为 $B_i=\{B_{i1}, B_{i2}, B_{i3}, B_{i4}\}$. 其中 B_{i1}、B_{i2}、B_{i3} 和 B_{i4} 对应的关键词分别为 K_1、K_2、K_3 和 K_4,则用户 U_i 的搜索内容为 $C_i=\langle U_i, \{K_1, K_2, K_3, K_4\}\rangle$.

1.1 用户相似模型的定义

贝叶斯网是一个 DAG $G=(V, E)$,随机变量集 V 构成 G 中的结点,每个结点对应一个随机变量,结点的状态对应随机变量的值,E 中的有向边表示结点间的条件依赖关系. 如果存在从结点 X 指向结点 $Y(X\neq Y; X, Y(V)$ 的有向边,称 X 是 Y 的一个父结点,变量 X 在图 G 中的父结点集用 $Pa(X)$. 每个结点 X 都有一张 CPT,用以量化父结点集对该结点的影响. 基于 BN 的基本概念和定义 1,下面给出 UBN 的概念,作为 CTR 预测的模型基础.

定义 2 UBN 用一个二元组 $G=(G_U, P)$表示,其中:

(1) $G_U=(U, E)$为 UBN 的 DAG 结构,$U=\{U_1, U_2, \cdots, U_n\}$为 G_U 的结点集,每个结点对应一个存在搜索内容的用户,有向边集 E 表示用户间的相似关系. U_i 取值为 1 或 0,分别表示 U_i 是否存广告搜索行为;若存在有向边 $U_j(U_i$,则称 U_j 是 U_i 的一个父结点,U_i 的父结点集表示为 $Pa(U_i)$.

(2) $P=\{p(U_i\mid Pa(U_i))\mid U_i\in U)\}$为条件概率分布的集合,由各结点 CPT 中各概率参数值构成,$p(U_i\mid Pa(U_i))$表示结点 U_i 在其父结点的影响下的条件概率,用来描述 $Pa(U_i)$ 的状态对 U_i 的状态的影响.

1.2 用户相似模型的构建

DAG 的构建是 BN 构建的关键和难点[11],基于 DAG 可容易地计算各结点的 CPT. 因此,本节以 DAG 的构建作为重点,讨论基于对历史搜索记录的统计计算构建 UBN 的方法.

UBN 中有向边集 E 所描述的用户间的相似关系,包括如下两个方面的问题:① 用户间是否存在相似关系(是否有边相连);② 相似关系的指向(边的方向). 直观地,在特定的广告搜索历史记录中,搜索的关键词体现了用户的兴趣,并且不同用户同时搜索过的关键词越多,则这些用户的搜索行为就越相似.

针对问题①,对于任意两个用户,我们考虑这两个用户同时搜索过的关键词占他们搜索过的关键词的比例,该比例越高,则这两个用户就越相似;若该值高于相似度阈值,则在这两个用户对应结点之间存在一条以该值作为权重的无向边. 下面给出基于对历史搜索记录统计计算的用户间相似度计算方法,用户 U_i 与 U_j 的相似度用 $\mathrm{sim}(U_i, U_j)$表示:

$$\mathrm{sim}(U_i, U_j)=N(M_i\cap M_j)/N(M_i\cup M_j). \tag{1}$$

其中:U_i 和 M_i 由 $C_i=\langle U_i, M_i\rangle$描述,$M_i\cap M_j$ 表示用户 U_i 与 U_j 都搜索过的关键词集,$N(M_i\cap M_j)$为其中关键词个数;$M_i\cup M_j$ 表示用户 U_i 与 U_j 搜索过的的关键词集,$N(M_i\cup M_j)$为其中关键词个数;$N(M_i\cap M_j)$和 $N(M_i\cap M_j)$可简单地基于对 C_i 和 C_j 的统计计算得到.

设 ε 为给定相似度阈值,即当计算用户 U_i 与 U_j 的相似度 $\mathrm{sim}(U_i, U_j)>\varepsilon$ 时,认为用户 U_i 与用户 U_j 是相似的,即用户 U_i 与用户 U_j 对应结点间存在一条无向边.

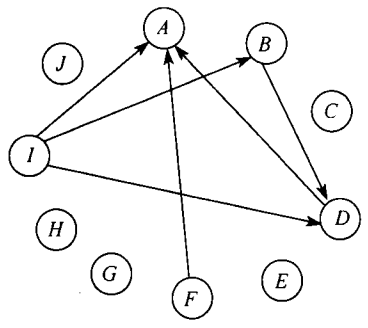
针对问题②,我们考虑任意两个有边相连的结点,这两个用户的搜索记录中,相同关键

词所占各自关键词的比例反映了他们共同兴趣所占个人兴趣的比例，该比例值高的用户所感兴趣的广告对于该比例值较低者也可能感兴趣；基于此可确定相似关系的指向，从而确定图中边的方向. 下面给出基于对历史搜索记录统计计算来判断用户间指向关系的方法，用户 U_i 对 U_j 兴趣的依赖度用 $D(U_i|U_j)$ 表示，用户 U_j 对 U_i 兴趣的依赖度用 $D(U_j|U_i)$ 表示：

$$D(U_i \mid U_j) = N(M_i \cap M_j)/N(M_j), D(U_j \mid U_i) = N(M_i \cap M_j)/N(M_i). \quad (2)$$

若 $P(U_i|U_j)>P(U_j|U_i)$，则表示 U_i 对 U_j 兴趣的依赖程度高于 U_j 对 U_i 兴趣的依赖程度，即 U_i 和 U_j 之间的无向边应由 U_j 指向 U_i，这意味着：U_j 点击过的广告 U_i 点击的可能性也较大.

例 2　基于以上相似度和兴趣依赖度两个度量标准，可得到 UBN 的 DAG 结构，进而可容易地计算 UBN 各结点的 CPT，从而最终得到 UBN. 一个简单的 UBN 示例如图 1 所示.



(a) UBN 的 DAG 结构

(a) DAG structure of UBN

B	D	F	I	A	
				0	1
0	0	0	0	1.0	0.0
0	0	0	1	0.0	1.0
0	0	1	0	0.0	1.0
0	0	1	1	0.0	1.0
0	1	0	0	0.0	1.0
0	1	0	1	0.0	1.0
0	1	1	0	0.0	1.0
0	1	1	1	0.0	1.0
1	0	0	0	0.0	1.0
1	0	0	1	0.0	1.0
1	0	1	0	0.0	1.0
1	0	1	1	0.0	1.0
1	1	0	0	0.0	1.0
1	1	0	1	0.0	1.0
1	1	1	0	0.0	1.0
1	1	1	1	0.0	1.0

(b) 结点 A 的 CPT

(b) CPT of node A

图 1　一个简单的 UBN

Fig. 1　A simple UBN

对于包括 n 个结点的 DAG，式(1)的执行次数为 $O(n^2)$；若假设每个用户搜索的关键词不超过 m 个，则式(1)的计算时间为 $O(m)$；由于广告搜索记录的稀疏性，确定边的方向时，式(2)的执行次数将远远小于 $O(n^2)$. 因此，以上方法可在 $O(mn^2)$ 时间内构建 DAG. 然而，实际的广告搜索记录中，m 和 n 往往都具有较高的数量级. 基于本文的研究和初步探索，针对大规模的广告及搜索记录构建模型，是我们今后要开展的工作.

2　基于 UBN 的近似推理及 CTR 的预测

2.1　UBN 的近似推理算法

基于第 1.2 节中的用户间相似度度量函数 $\mathrm{sim}(U_i,U_j)$，可以得到具有直接相似关系的用户. 而用户之间往往也存在着大量的间接相似关系，因此我们考虑利用 UBN 的概率推理机制，以一种通用的方式来获取直接或间接的用户相似关系，为 CTR 预测奠定基础.

基于 BN 进行边缘概率、条件概率或后验概率的计算，是 BN 推理的基本任务. 基于 BN 的概率推理，可针对给定的证据值计算目标变量可能取值的不确定性. 为了基于 UBN 定量地推断与特定用户存在广告搜索行为方面具有相似性的用户，我们利用 BN 的概率推理机制计算给定结点为证据(当前用户)情形下其他结点的概率值，作为相似用户推断的依据.

然而，BN 的精确推理具有指数计算时间[10-11]，并不能有效地支持相似用户的发现. 马

尔可夫链蒙特卡罗(Markov Chain Monte Carlo, MCMC)方法是近年来被广泛应用的一种贝叶斯计算方法;Gibbs 采样[16,17]又是 MCMC 方法中最简单、使用最广泛的一种,它从满条件分布中迭代地进行抽样,当迭代次数足够大时,就可以得到来自联合后验分布的样本,也可以得到来自边缘分布的样本. Gibbs 采样能够支持高效的条件概率和后验概率计算,为 UBN 的高效概率推理提供了理论基础,为基于 UBN 中推断相似用户提供了有效的计算手段. 因此,我们基于 Gibbs 采样进行 UBN 的近似推理,为了简化计算而又不失一般性,对于当前的采样结点,仅考虑其马尔可夫覆盖①中的结点对它的影响,以上思想由算法 1 描述.

算法 1 基于 Gibbs 采样的 UBN 近似推理

输入:G:UBN $G=(U, E)$,$U=\{U_1, U_2, \cdots, U_n\}$,$U_i$ 为 1 或 0 分别用 u_1 和 u_0 表示

Φ:证据变量(当前用户)

e:Φ 的取值(是否搜索广告)

Z:非证据变量

q:目标变量(拟推断的用户)

s:采样总次数

输出:$p(q=1|e)$

变量:z:Z 的取值

$U_{(-i)}=\{U_1, \cdots, U_{i-1}, U_{i+1}, \cdots, U_n\}$

$v=e\cup z$:G 的当前状态

$N(q_1)$:q 值为 1 的样本数量

1. 初始化
   - z(随机地为 Z 中的每个 U_i 赋值,$v^{(0)}\leftarrow e\cup z$,$c^{(0)}\leftarrow v^{(0)}$
   - $N(q_1)\leftarrow 0$
2. 产生样本序列

   For $k=1$ to s Do

   1) 基于当前状态 $c^{(k-1)}$ 计算被选变量的概率
      - 随机地从 Z 中选择一个非证据变量 U_i
      - $B(p(U_i=0|c_{MB(U_i)})+p(U_i=1|c_{MB(U_i)})$

      //$c_{MB(U_i)}$ 为 U_i 的马尔可夫覆盖中各变量在当前状态 c 下的值

   2) 通过 $p(U_i|c_{MB(U_i)})$ 对 U_i 采样
      - 随机产生 $r_k\in[0, B]$,并利用(公式 3)确定 U_i 的值

$$Ui=ui=\begin{cases}0 & rk\leqslant p(Ui=0\mid c_{MB(Ui)})\\ 1 & p(Ui=0\mid c_{MB(Ui)})<r_k\leqslant p(Ui=0\mid c_{MB(Ui)})+p(Ui=1\mid c_{MB(Ui)})\end{cases}\tag{3}$$

      - $v^{(k)}(\leftarrow v_{(-i)}^{(k-1)}, u_i)$;$c^{(k)}\leftarrow v^{(k)}$

   End For k

3. 计算 $N(q_1)$和 $p(q=1|e)$

   For $k=1$ to s Do

   If $c_q^{(k)}=1$ Then //在第 k 个样本 q 是否为 1

   $N(q_1)(N(q_1)+1$

① X 的马尔可夫覆盖(Markov Blanket, MB)[10]是包括 X 的直接孩子结点、X 的直接父亲结点、以及 X 的直接孩子的其他父亲结点的结点集合,记为 MB (X).

End if

End For

Return $p(q=1|e)(N(q_1)/s$　// 估计 $p(q=1|e)$的值

利用算法 1，可以得到与某一用户具有直接或间接相似关系的用户. 通过设定相似度阈值 λ，若 $p(U_j=1|U_x=1)\geqslant\lambda$（为了表达的方便，记为 $p(U_j|U_x)\geqslant\lambda$），则认为 U_j 与 U_i 相似. 文献[16-17]已经给出结论，Gibbs 采样算法只要采样次数足够多，它总能收敛到一个静态分布(Stationary distribution)、且为正确值，这保证了算法 1 的收敛性和有效性. 一般地，蒙特卡罗概率算法能以较高执行效率得到预期的答案，这从理论上保证了算法 1 能快速地收敛到正确的值. 第 3 节中我们将通过实验测试算法 1 的收敛速度.

2.2　基于 UBN 的 CTR 预测

利用 UBN 及其上的概率推理机制，可获得存在广告点击记录的用户与不存在点击记录的用户之间在广告搜索行为方面的相似性，因此，我们通过存在点击记录的用户对广告的 CTR，来预测将该广告推荐给未知用户时可能的 CTR.

假设 U_x 为不存在广告点击记录的用户，根据算法 1 可以得到其相似用户的集合 $US=\{U_j|\ p(U_j|U_x)\geqslant\lambda, U_j(U,\ j\neq x\}$，设 $A=\{A_1,A_2,\cdots,A_l\}$ 为 U_s 中的用户点击过且已知 CTR 的广告的集合. 我们基于 U_x 与 US 中用户在广告搜索行为方面的相似性，来预测 U_x 可能点击 A 中的广告的 CTR，对与 A 中的 $A_i(i=1,\ 2,\ \cdots,\ l)$，用 $y_{Ai}=1$ 和 $y_{Ai}=0$ 分别表示广告 A_i 是否被点击，则 U_x 对是否点击广告 A_i 可由(公式 4)度量，其中 $p(y_{Ai}\mid A_i,U_j)$ 表示 U_j 对 A_i 的 CTR：

$$\sum_{U_j} p(y_{Ai}\mid A_i,U_j)p(U_x\mid U_j). \tag{4}$$

对 A 中所有 A_i，我们进一步对(公式 4)归一化，得到 U_x 对 A_i 的 CTR，如下：

$$p(y_{Ai}\mid A_i,U_x)=\frac{\sum_{U_j} p(y_{Ai}\mid A_i,U_j)p(U_x\mid U_j)}{\sum_{A_i\in A}\sum_{U_j} p(y_{A_k}\mid A_k,U_j)p(U_x\mid U_j)}. \tag{5}$$

例 3　对于图 2 所示的相似用户和已知的 CTR，利用上述方法，通过与用户 U_1 相似的用户 U_2、U_3 和 U_4 来预测 U_1 对广告 A_1、A_2、A_3、A_4 和 A_5 的 CTR. 以 A_4 为例，根据式(4)可以得到 $p(y_{A4}=1|A_4,U_2)p(U_1|U_2)+p(y_{A4}=1|A_4,U_3)p(U_1|U_3)+p(y_{A4}=1|A_4,U_4)p(U_1|U_4)=0.26\times0.48+0.32\times0.62+0.13\times0.59=0.399\ 9$，归一化后可得 $p(y_{A4}=1|A_4,U_1)=0.446$. 类似可计算 U_1 对其他广告的 CTR. 最终可以得到向用户 U_1 推荐广告的顺序为 A_4，A_2，A_3，A_5，A_1.

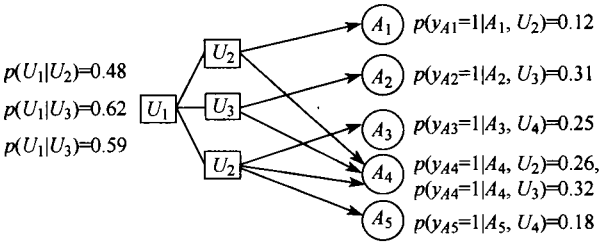


图 2　相似用户和已知点击率

Fig. 2　Similar users and the given CTRs

3 实验结果

为了测试本文方法的有效性,我们实现了 UBN 的构建,近似推理及基于 UBN 的 CTR 预测方法,测试了 UBN 构建和近似推理的效率、基于 UBN 的 CTR 预测效率和准确性. 实验建立在如下配置的 IBM 服务器上: Intel Xeon Processor E5405 (12M Cache, 2.00 GHz, 1 333 MHz FSB)处理器,2 GB 内存,Ubuntu 11.10 操作系统,使用 Python 语言编写程序,MongoDB 存储数据. 实验采用 KDD Cup 2012-Track 2[18] 中的测试数据.

考虑到在实际的广告推荐中,用户首先输入关键词,然后广告媒介根据该关键词来为用户推荐广告,也鉴于本文是对基于 BN 的 CTR 预测进行的初步试探性研究,以及现有实验设备的计算能力,对于包括 474 455 个用户、每个用户可能搜索 118 868 个关键词的情形,我们从测试数据中对点击过同一个关键词(即广告)的用户进行分组,每组包括 10～100 个用户,基于此完成实验测试.

3.1 构建 UBN 的效率

我们从源数据中选取包括 10,20,…,100 个用户的数据片段,测试了构建 UBN 的时间开销,对于每个不同的用户数,记录了 10 次测试的平均时间开销. 图 3 给出了包括与不包括数据库 I/O 开销的 UBN 构建时间. 可以看出,构建 UBN 的时间开销随结点数的增加基本呈线性增长趋势,数据库的 I/O 开销大约占总时间的 25%,但基本不随用户数增加而提高.

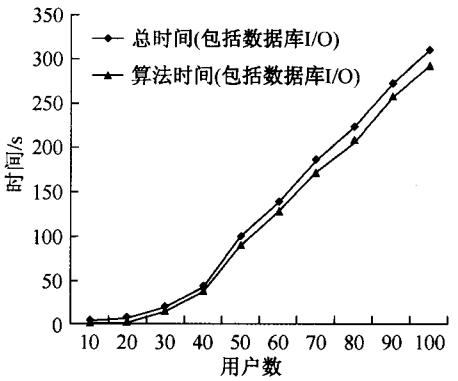


图 3 构建 UBN 的效率

Fig. 3 Efficiency of UBN construction

3.2 UBN 推理效率

我们首先测试了不同用户数的情形下,随着采样次数增加算法 1 返回结果的收敛性和执行的时间开销,分别如图 4 和图 5 所示. 从图 4 可以看出,随着采样次数的增加,不同用户数情形下的 UBN 推理结果均能较快地收敛到一个稳定的值. 这说明算法 1 能较快地收敛到静态分布,即算法 1 能高效地进行 UBN 的概率推理、发现相似用户;同时,用户数越多,与特定用户相似的用户具有越高的相似度. 从图 5 可以看出,不同用户数的情形下,算法 1 的执行时间都随着采样次数呈线性趋势增加. 这说明 UBN 近似概率推理算法的高效性.

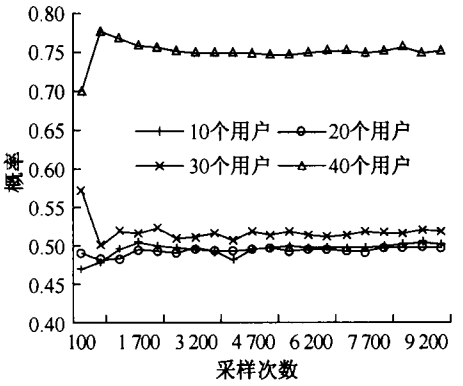


图 4　UBN 近似推理结果的收敛性

Fig. 4　Convergence of UBN's approximate inference results

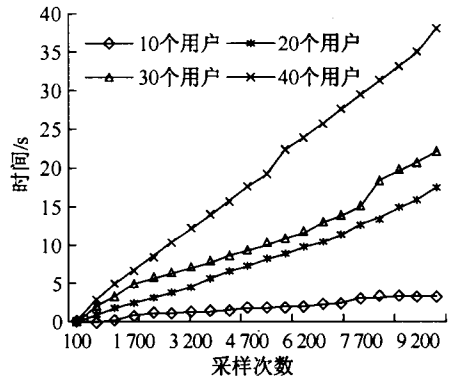


图 5　UBN 近似推理算法的效率

Fig. 5　Efficiency of UBN's approximate inference

3.3　CTR 预测的准确性

我们从测试数据中随机抽取了 12 592 个用户和广告(每个广告用一个关键词描述)，将基于本文第 2 节中的 CTR 预测结果与真实的 CTR 进行了比较. 为了展示 CTR 预测结果的有效性，在这 12 592 次测试结果中随机选择 10 次. 表 1 给出了用户 ID、关键词、真实 CTR、预测得到的 CTR(预测 CTR)及两者之间的误差(真实 CTR 与预测 CTR 之差的绝对值). 进而可以得到，CTR 预测的最大、最小和平均误差为 0.666 667、0 和 0.187. 同时，对所有 CTR 预测结果的误差进行了分段统计，基于本文方法的 CTR 预测结果，误差低于 0.1、0.1—0.2、0.2—0.3、0.3—0.4、0.4—0.5，以及高于 0.5 的广告数如图 6 所示，所占总广告数的比例分别为 32%、20%、13%、9%、15%、11%. 由此可以得出结论，本文所提出的 CTR 预测方法能得到较准确的结果. 然而，实际中广告的历史点击记录往往较稀疏，并且点击率的值往往也较小. 因此，进一步降低 CTR 预测结果的误差，是本文所提出的方法能有效用于实际 CTR 预测的前提，这也是我们将要进一步研究测试的方面.

表 1 部分 CTR 预测误差

Tab. 1 Some errors of CTR prediction

用户 ID	关键词	真实 CTR	预测 CTR	误差
12 203 747	42 453	0.5	0.5	0
18 618 596	8 382	1	0.333 333	0.666 667
21 554 134	52 033	1	0.5	0.5
17 647 407	53 344	0.333 333	0.5	0.166 667
13 124 169	5 297	0.5	0.5	0
10 750 244	19 168	0.5	0.333 333	0.166 667
9 025 547	20 047	1	1	0
11 955 508	7 170	0.5	0.361 111	0.138 889
15 253 596	15 623	0.5	0.166 667	0.333 333
19 958 274	11 677	1	1	0

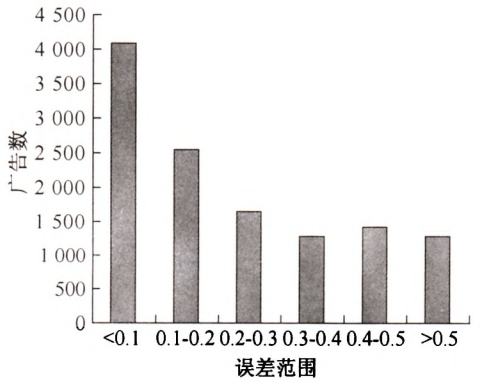


图 6 CTR 预测误差分布

Fig. 6 Distributions of CTR prediction errors

4 总结与展望

互联网广告的 CTR 预测,是计算广告领域研究的热点和关键;考虑用户偏好、预测不存在点击记录的用户对广告的 CTR,又是目前 CTR 预测亟待解决的问题之一. 本文以贝叶斯网这一重要的概率图模型作为广告检索、点击和推荐中不确定性表示和推理的基本框架,利用 BN 构建用户的相似模型,基于 BN 的概率推理机制定量地度量用户之间的行为相似性和相似的不确定性,继而利用用户相似关系和已知点击率来预测可能推荐给用户的广告的 CTR. 作为基于概率图模型进行广告 CTR 预测的初步试探性研究,本文主要针对这一思路本身的阐述和讨论,旨在探索这一思路的有效性和适用性,为后续工作奠定基础.

本文通过小规模真实数据集上的实验,测试了所提出思路和方法的正确性,实验结果从一定程度上验证了这一思路的可用性、继续深入研究的必要性;而从实际中 CTR 预测问题的特点和需求来看,如何处理海量的广告及用户数据,进一步探索基于 BN 预测广告 CTR 这一思路的实用性,满足海量数据处理需求和针对实际情形 CTR 预测的可行性,是我们以本文中初步探索性研究结果为基础、将来要开展的工作. 例如:基于在线学习机制的模型构建和增量更新;将本文的方法与 LDA 主题模型相结合,在统计计算的基础上进行语义分

析；考虑用户及广告的分类间的相互关系，建立层次模型以支持高效的 CTR 预测.

致谢　感谢云南大学高性能计算中心为本文研究提供了良好的实验环境和软硬件设备的支持.

[参　考　文　献]

[1] 周傲英，周敏奇，宫学庆. 计算广告：以数据为核心的 Web 综合应用 [J]. 计算机学报，2011，34(10)：1805-1819.

[2] REGELSON M，FAIN D. Predicting click-through rate using keyword clusters[C]//Proceedings of the Second Workshop on Sponsored Search Auctions，EC 2006. Michigan：ACM，2006.

[3] AGARWAL D，BRODER A，CHAKRABARTI D，et al. Estimating rates of rare events at multiple resolutions. Proceedings of the ACM SIGMOD International Conference on Management of Data. Beijing：ACM，2007：16-25.

[4] RICHARDSON M，DOMINIWSKA E，RAGNO R. Predicting Clicks：Estimating the Click-Through Rate for New Ads[C]//Proceedings of the 16th International Conference on World Wide Web，WWW 2007. Banff：ACM，2007：521-530.

[5] CHAKRABARTI D，AGARWAL D，JOSIFOVSKI V. Contextual Advertising by Combining Relevance with Click Feedback[C]//Proceedings of the 17th International Conference on World Wide Web，WWW 2008. Beijing：ACM，2008：417-426.

[6] GOLLAPUDI S，PANIGRAHY R，GOLDSZMIDT M. Inferring Clickthrough Rates on Ads from Click Behavior on Search Results[C]//Proceedings of the Workshop on User Modeling for Web Applications，Fourth International Conference on Web Search and Web Data Mining，WSDM 2011. Hong Kong：ACM，2011.

[7] YAN J，LIU N，WANG G，et al. How much can Behavioral Targeting Help Online Advertising? [C]//Proceedings of the 18th International Conference on World Wide Web，WWW 2009. Madrid：ACM，2009：261-270.

[8] AHMED A，LOW Y，ALY M，et al. Scalable distributed inference of dynamic user interests for behavioral targeting[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego，CA：ACM，2011：114-122.

[9] WANG X，LI W，CUI Y，et al. Click Through Rate Estimation for Rare Events in Online Advertising. Online Multimedia Advertising：Techniques and Technologies，Chapter1 [M/OL]. 2011[2012-06-15]. http://labs.yahoo.com/node/434.

[10] PEARL J. Probabilistic Reasoning in Intelligent Systems：Networks of Plausible Inference [M]. San Mateo，CA：Morgan Kaufmann Publishers，1988.

[11] RUSSEL S，NORVIG P. Artificial Intelligence—A Modern Approach [M]. Boston：Pearson Education，Publishing as Prentice-Hall，2002.

[12] DEMBCZYNSKI K，KOTLOWSKI W，WEISS D. Predicting Ads' Click Through Rate with Decision Rules. [EB/OL]. 2008-03-31[2012-06-15]. Yahoo Research，http://research.yahoo.com/workshops/troa-2008/papers/submission_12.pdf.

[13] GRAEPEL T，BORCHERT T，HERBRICH R，et al. Probabilistic Machine Learning in Computational Advertising Microsoft Research [EB/OL]. 2010-12-10 [2012-06-15]. http://research.microsoft.com/en-us/um/ beijing/events/mload-2010/.

[14] CHAPELLE O，ZHANG Y. A dynamic Bayesian network click model for web search ranking[C]//Proceedings of the 18th International Conference on World Wide Web，WWW 2009. Madrid：ACM，2009：1-10.

[15] 张少中，高飞. 一种基于小世界网络和贝叶斯网的混合推荐模型 [J]. 小型微型计算机系统，2010，31(10)：1974-1978.

[16] HRYCEJ T. Gibbs sampling in Bayesian networks [J]. Artificial Intelligence，1990，46：351-363.

[17] PEARL J. Evidential reasoning using stochastic simulation of causal models [J]. Artificial Intelligence，1987，32：245-257.

[18] KDD CUP 2012 Track 2：Predict the click-through rate of ads given the query and user information [EB/OL]. 2012-02-20[2012-06-15]. http://www.kddcup2012.org/c/kddcup2012-track2.