基于邻居选取策略的人群定向算法

Crowd Orientation Algorithm Based on the Neighbor Selection Strategy

刘旺勤
Wangqin Liu
南昌师范学院,中国·江西 南昌 330000
Nanchang Normal University, Nanchang, Jiangxi, 330000, China

【摘　要】人群定向是广告商投入广告常使用的用来寻找目标客户的一种手段,如百度、淘宝就用此方法来推荐广告,是广告推荐系统的重要技术手段,通过对人群的行为数据的分析,为广告商找到合适的人群,但是中国目前的人群定向算法还是以协同过滤法为主。但是该算法有其自身的缺陷,如精准程度不高、抗攻击力不强等问题,为了使这一问题得到完美的解决,基于邻居选取策略的人群定向算法就由此应运而生。论文首先介绍了什么是人群定向,接下来介绍用户相似度分析,最后阐述了基于邻居选取策略的人群定向算法分析。

【Abstract】Crowd orientation is often used by advertisers to find targeted customers, such as Baidu, Taobao uses this method to recommend advertising. It is an important technical means of advertising recommendation system. Through analyzing the behavior data of the crowd, it can find suitable people for advertisers, but the current crowd orientation algorithm in China is mainly collaborative filtering. However, the algorithm has its own shortcomings, such as low precision, weak anti-attack and so on. In order to solve this problem perfectly, the crowd orientation algorithm based on neighbor selection strategy emerges as the times require. This paper firstly introduces what is crowd orientation, then introduces the analysis of user similarity, and finally expounds the crowd orientation algorithm based on the neighbor selection strategy.

【关键词】人群定向;用户相似度分析;基于邻居选取策略的人群定向算法

【Keywords】crowd orientation; analysis of user similarity; crowd orientation algorithm based on the neighbor selection strategy

【DOI】http://dx.doi.org/10.26549/gcjsygl.v2i3.729

1 引言

随着互联网的快速发展,中国互联网用户数量剧增,中国成为互联网使用人数最多的国家，而当今的广告企业也搭乘了互联网的高度列车,利用互联网的优势来进行广告投放。为了保证广告可以准确地投入到合适的人群中，人群的寻找定位就是重中之重,人群定向技术可以准确地对客户进行定位,这满足了广告主的投放需要。传统的人群定向算法只是依据对商品评分来寻找到客户人群,而这往往没有将客户的年龄、职业、性别等考虑在内,比较笼统,所以基于邻居选取策略的人群定向算法就解决了这一难题,受到广告主的青睐。

2 人群定向的相关概述

人群定向，是基于广告主投放广告的需要而形成的一种寻找客户人群的一种方法,在对目标进行定向选择时,使用的技术就是人群定向技术。人群定向技术是根据用户行为来进行数据分析,由此找到具有相同行为的客户群,广告主们选择合适的媒体来将广告投放给他们,可以减少广告的投放成本,并且达到好的广告效果。人群定向是数字信号处理(DSP)的核心特征之一，增强了互联网广告市场的透明度，使其更加高效,提升其可控制性,是未来网络广告要走的一条发展之路,人群定向分析直接决定 DSP 的竞价策略，分析得出的结果对广告主的投放策略以及广告的投放效果起到决定性的作用。

3 用户相似度分析

用户通过对各个网站进行点击访问,如淘宝网、优酷网、百度等，那么用户的信息就在每一次点击访问中留存在网站之中,网站就可以直接获得这些信息,帮助广告主投放广告,这就是用户行为相似度;每个用户都是单独的个体,存在着个体差异，而用户的有些信息是无法通过访问网站来直接得知的,这就是用户性别、年龄、学历、职业等人口属性;娱乐、游戏、健康等兴趣爱好,这就是用户特征相似度。而这两者就构成了传统意义上的用户相似度，接下来就重点介绍用户的相似度[1]。

3.1 人口属性的相似度

人口属性有数值型和名称型之分，数值型属性，顾名思义,就是属性是以数值形式表现出来的,如年龄和收入,这些就是数值型人口属性；名称型人口属性与数值型人口属性相对,不可通过数值表现出来,也就是非数值型人口属性,例如性别、学历等,无法将这些属性以数值形式表现出来。因为二者的表现形式不同，所以在计算相似度时不能使用一种算法来对二者加以计算,要有所区分,保证计算出的相似度是科学合理的。

数值型属性的相似度计算：第一步就是将 2 个用户属性

的数值之差的绝对值计算出来，并将最大的差的绝对值和最小的差的绝对值划分区间$[V_{min},V_{max}]$，然后将区间划分为 N 个等距的子区间，即$[V_{min},V_1]$,$[V_1,V_2]$.....$[V_{n-1},V_{max}]$，并且每个子区间的距离为 1,2,3.....n。若是 2 个客户的绝对值差在第 i 个子区间内，那么该属性的距离就是 d=i(i≤n)。以下面的例子来方便人们理解。

例一：假设 2 个客户的年龄最小值和最大值的区间为$[V_{min},V_{max}]$，并且将其划分为 3 个子区间，即$[V_{min},V_1]$,$[V_1,V_2]$,$[V_{n-1},V_{max}]$，那么用户 A 和 B 二者的年龄差就是$|l|=|a-b|$，其中 a 表示的是用户 A 的年龄，b 显示的就是用户 B 的年龄，并且$|l|\in[V_1,V_2]$，那么就可以清晰地得知，二者的距离为 2。

针对所有的数值型属性$a_1,a_2,a_3...a_n$，用户 A 和 B 的距离为：

$$D=1,\forall_{dj}=0,$$

$$D=\frac{1}{n}\sum_{j=1}^{n}dj,\exists_{dj}>0,$$

其中d_j表示 2 个用户在属性a_j上的距离，若所有的d_i都为 0，则 D 的默认值就是 1。

例二：假设名称型属性的取值数目为 N，那么对所有取值进行人工评级，即为$r_1,r_2,r_3...r_n$，若两个用户在该属性上的评级为r_i和r_j，那么 2 个用户的距离为$|ri-rj|$。

下面以学历来作为例子，加深理解，将学历分为专科、本科、硕士、博士，对应的人工评级分别为 1,2,3,4，若两个用户 A 学历是本科，B 的学历是博士，那么二者学历距离是 3。

针对所有名称型属性$b_1,b_2,b_3...b_n$，A 和 B 的距离为：

$$D1=1,\forall_{dj}=0$$

$$D1=\frac{1}{n}\sum_{j=1}^{n}dj,\exists dj>0,$$

其中d_j表示 2 个用户在属性b_j上的距离，若所有的d_j都为 0，则 D 的默认值就是 1。

如果两个用户的距离越小，那么二者的相似度就越大，那么人口属性的相似度即：

$$\mathrm{sim(A,B)}=\frac{D+D1}{2\times D\times D1}$$

3.2 兴趣相似度

输入：用户兴趣概率矩阵$P=(P_{ij})$，P 是 m×n 矩阵、用户U_x的兴趣概率向量$P_x=(P_{x1},P_{x2}...P_{xn})$，n 个兴趣；

输出：用户的兴趣指纹 F。

$H\leftarrow\varnothing,F'\leftarrow\varnothing;$

for $i\leftarrow1$ *to* n *do*/* 本文中 K 的取值为所有兴趣的计数/

对于兴趣 i，生成一个 k 为 hash h_i;

$H\leftarrow H\cup H_i;$

end for

for $i\leftarrow1$ *to* K *do*

$sum\leftarrow0;$

for $j\leftarrow1$ *to* n *do*

$q\leftarrow h_j$的第 i 位;

if $q=1$ *do*

$sum=sum+p_{xj}$

else if $q=0$ *do*

$sum=sum-p_{xj};$

end if

end for

if $sum>0$ *do*

$f_i=1;$

else $f_i=0;$

end if

$F'\leftarrow F'\cup F_i;$

end for

将F'中的元素进行链接，生成$f_1\ f_2\ f_3..f_k;$

$F\leftarrow f_1\ f_2\ f_3..f_k;$

return F.

则二者的兴趣相似度为：

$$Sim(U_a,U_b)=\frac{K-\sum_{i=1}^{K}f_{Ai}\oplus f_{Bi}}{K}$$

3.3 用户特征相似度

假设有 3 个用户，A、B、C，其 A 和 B 人口属性和兴趣相似度分别为 0.8 和 0.2，而 A 和 C 的人口属性和兴趣相似度为 0.6 和 0.4。若是采用算数平均值方法来计算相似度，A 和 B 的相似度为 0.32，A 和 C 的相似度为 0.48，A 和 C 的相似度高于 A 和 B[2]。

3.4 用户行为相似度

例一：假设有 m 个用户组成的集合$U=\{u_1,u_2,...u_m\}$，和 n 个媒体组成的集合$M=\{M_1,M_2,...M_n\}$，那么用户-媒体访问矩阵可以用 m×n 矩阵 C 来表示，$C_{ij}(1\le i\le m,1\le j\le n)$表示用户$U_i$访问媒体$M_j$的计数，并且用户 A 和 B 的共同访问媒体集为$T$，那么 2 个用户的行为相似度为：

$$\mathrm{sim(A,B)}=\frac{\sum_{M_j\in T}\left|(C_{A,j}-\bar{C}_A)(C_{B,j}-\bar{C}_B)\right|}{\sqrt{\sum_{M_j\in T}(C_{A,j}-\bar{C}_A)^2}\sqrt{\sum_{M_j\in T}(C_{B,j}-\bar{C}_B)^2}},$$

其中，$C_{A,j}$和$C_{B,j}$分别表示用户 A 和 B 的访问媒体M_j的计数，sim(A,B)表示用户 A 和 B 的行为相似度，$\bar{C}_A$和$\bar{C}_B$分别表示用户 A 和 B 的所有媒体的平均计数。

4 基于邻居选取策略的人群定向算法

人群定向是广告主进行广告投放而采取的一种手段，因为过去的传统算法存在不足之处，无法满足广告主投放广告的需求，所以为了弥补传统算法的缺陷，基于邻居选取策略的人群定向算法受到广告主的青睐，其可以进行准确的人群定向计算，帮助广告主更好地投入广告。首先，找出和种子人群行为相近的人群，这是由人们对网站的访问频率来决定的；其次，相关人员在进行一系列相似度计算之后，找出每个种子用户的邻居，并且将他们的所有邻居都作为候选人群；第三，通过基于邻居选取策略的人群定向算法的计算，找出潜在目标。

4.1 基于用户行为和用户特征的候选人群选取算法

输入：种子用户集合 $S=\{S_1,S_2,...S_r\}$，人群集合 $U=\{U_1,U_2,...U_m\}$；种子用户–媒体访问矩阵 $C=(C_{ij})$；C 是 r×n 矩阵；人群–媒体访问矩阵 $C'=(C_{ij})$；C' 是 m×n 矩阵；种子用户–兴趣概率矩阵和人群–兴趣概率矩阵，种子用户及人群的人口属性，参数 k；

输出：候选人群 P

$S\leftarrow\varnothing, u_c\leftarrow u, M_c\leftarrow 0, P\leftarrow\varnothing$；/* u 为中心用户，$M_c$ 为中心用户 u 的访问的媒体量 */

for $i\leftarrow 1$ *to* n *do*

$num\leftarrow 0, sum\leftarrow 0$;

for $j\leftarrow 1$ *to* r *do*

if $C_{ij}\neq 0$ *do*

$sum\leftarrow sum+C_{ij}$;

$num\leftarrow num+1$

end if

end for/* 给 M_c 的第 i 个元素赋值 */

$M_c(i)\leftarrow sum/num$

end for

$sumSim\leftarrow sumSim+sim_B(u_i,u_c)$;

end for

$T_{sim}\leftarrow sumSim/|U|$;

for each $u_i\in U$ *do*

if $sim_B(u_i,u_c)\geq T_{sim}$ *then*

$S\leftarrow S\cup\{u_i\}$;

end if

end for

for each $u_i\in S$ *do*

for each $u_j\notin S$ *do*

计算 $sim(u_i,u_j)$;

end for

对行为相似人群中的用户按相似度进行降序排列，并选择前 k 个用户放入 P，并将这 k 个用户从 S 中删除；

end for

return P

4.2 基于邻居选取策略的人群定向算法

输入：种子用户集合 S、候选人群 P_{cand}、种子用户–媒体访问矩阵、种子用户–兴趣概率矩阵、候选人群–媒体访问矩阵、候选人群–兴趣概率矩阵、种子用户及候选人群的人口属性；

输出：目标人群集合 A_1；

$A_1\leftarrow\varnothing, sum\leftarrow 0, Q\leftarrow\varnothing$;

for each $u_i\in P_{cand}$ *do*

计算 $sim(u_i,u_j)$;

$sumSim\leftarrow sumSim+sim(u_i,u_j)$;

end for

$sim_i\leftarrow sumSim/|S|$;/* sim_i 为用户 u_j 与种子人群的相似度 */

$sum\leftarrow sum+sim_i$;

$Q\leftarrow Q\cup\{(sim_i,u_i)\}$;

end for

$T_{sim}\leftarrow sum/|Q|$;

for each$<sim_i,u_i>\in Q$ *do*

if $sim_i\geq T_{sim}$ *then*

$A_1\leftarrow A_1\cup\{u_i\}$;

end if

end for

return A_1.

5 结语

综上所述，论文首先对人群定向以及人群定向技术做了简要的概述，帮助人们了解什么是人群定向，紧接着对用户相似度进行了分析，最后举出了基于邻居选取策略的人群定向算法，克服传统算法的不准确、易受攻击的缺点，虽然该算法有这些优点，但是也存在不足之处，那就是没有将系统的时间属性考虑在内，人们在以后的实际工作中，要进一步完善人群定向算法，更好地为广告主服务，长路漫漫，还需要人们进行不断地求索。

参考文献：

[1]荣辉桂,火生旭,胡春华,等.基于用户相似度的协同过滤推荐算法[J].通信学报,2014,35(2):16-24.

[2]周孟,朱福喜.基于邻居选取策略的人群定向算法[J].计算机研究与发展,2017,54(7):1466-1475.

基金项目：

本论文是江西省教育厅科技项目《DVE 对等网络传输——基于兴趣行为的邻居发现策略研究》(编号:171125)。