搜索引擎用户查询的广告点击意图分析

靳岩钦，张 敏，刘奕群，马少平

（清华大学 智能技术与系统国家重点实验室 100084 北京）

摘 要：搜索引擎广告点击率的多少直接影响搜索引擎的收入，而深入分析用户查询的广告点击意图则是提高广告点击率的基础性工作. 针对与此，基于商用搜索引擎的用户查询点击日志，统计分析了搜索引擎用户查询的广告点击率，提出基于查询词内容匹配和基于贝叶斯分类的两种方法预测搜索引擎用户查询的广告点击意图. 在大规模的真实用户查询点击日志上的实验结果表明，所提出的方法能够预测查询的广告点击意图，将广告投放的精度从3.0%提高到36.8%，广告投放的平均F-measure值从0.060提升到0.408. 通过广告点击意图预测，有效缩小了广告投放范围，并适用于在线广告意图的实时预测.

关键词：搜索引擎广告；用户行为分析；查询意图；广告点击预测

中图分类号：TP391 **文献标志码**：A **文章编号**：0367－6234(2013)01－0124－05

Ads-clicking intent analysis of search engine queries

JIN Yanqin, ZHANG Min, LIU Yiqun, MA Shaoping

(State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China)

Abstract: Click through rate (CTR) on sponsored search ads determines the search engine's revenue, thus analysis on users' ads-clicking intent is one of the fundamental work to improve CTR. Based on the search logs provided by a Chinese search engine, this paper presents statistical analysis of ads clicks, and further proposes two methods to predict ads-clicking intent of query, namely query content match based prediction and Bayesian classification, respectively. Experimental results on large scale real data show the improvements from 3.0% to 36.8% in precision and from 0.060 to 0.408 in F-measure on sponsored search ads delivery. The proposed methods are capable of predicting the intent of user queries and enhancing the effect of search engine advertising, and are also applicable for online prediction of advertising click intent of user queries.

Key words: sponsored search; user behavior analysis; query intent; ads click prediction

近年来，搜索引擎的商业模式已经逐渐转化为搜索广告，即搜索引擎针对用户提交的查询内容，将相应的广告显示在搜索结果页面中，根据用户对这些广告的访问和点击情况，向广告商收取相应的费用[1]. 因而用户对广告的实际点击情况，直接影响搜索引擎的收入. 于是搜索引擎投放广告的点击率分析已经成为近年来研究的热点.

现有的搜索广告投放过程中并不区分用户的意图，只依靠关键词匹配进行“普遍撒网”，实际点击率很低，产生了很多无效广告投放. 而单纯增加投放广告的数量会严重影响普通用户的使用体验，甚至使用户放弃用该搜索引擎. 因此准确预测用户的查询是否具有广告点击意图，从而有针对性地选择投放内容相关的广告，具有很好的研究意义与应用价值.

本文探讨了从用户查询的广告点击意图预测方法，首先介绍相关研究工作；然后分别提出两种用户查询的广告点击意图的预测模型，并基于大规模真实用户查询点击日志给出相应实验结果；接下来通过广告点击曲线拟合模型分析方法的可行性；最后给出结论与未来工作.

收稿日期：2010－11－04.

基金项目：国家自然科学基金资助项目(60903107,61073071)；国家高技术研究发展计划资助项目(2011AA01A205).

作者简介：靳岩钦(1988—)，男，本科生；
马少平(1961—)，男，教授，博士生导师.

通信作者：张 敏，z-m@tsinghua.edu.cn.

1　相关工作

当前提高广告点击率的工作主要面向上下文广告和搜索广告这两种类型.

在上下文广告方面，一些大公司在相关研究上起到主导作用. Yahoo！的研发人员提出了一系列提高上下文广告投放相关性的方法，如 B. Ribeiro-Neto 等[2]比较了若干种基于关键词的页面与广告的匹配算法；A. Broder 等[3]利用海量数据和热门广告词，建立了针对网页以及广告的分类体系，并利用相关类别改进广告分类效果；随后在文献[4]中探讨了根据用户点击反馈提高上下文广告投放效果. A. Anagnostopoulos 等[5]研究了上下文广告投放过程中的实时性和效率问题. 但是这些工作都没有涉及查询的广告点击意图预测.

在搜索广告方面，K. Debmbsczynski 等[6]利用引发投放某个广告的所有查询的内容来构建广告的标题和主题，并根据搜索结果页面的特征（广告的排名和搜索结果页的编号）以及广告 URL，建立模型，预测新广告的点击率. M. Regelsonl 等[7]发现不同的词项（term）在引发广告点击的可能性上具有特定的差别. 例如“数码相机”和“人脑结构”的广告意图是不同的，前者引发广告点击的可能性高于后者. 因此，在论文中提出通过词项的点击率来反映这种固有的差别. A. Ashkan 等[7]意识到更好地理解用户查询的意图有助于提供个性化的搜索结果并且提高用户满意度，他们利用历史点击信息、查询自身的特征以及搜索结果页的内容分析预测用户查询的商业意图，研究表明将这 3 组特征结合起来可以有效检测出用户的查询意图. 在国内，陈磊等[9]统计了各大商用搜索引擎搜索广告方面的统计数据，研究了大量真实用户和搜索广告的实际交互行为. 王家卓等[10]研究了在搜索结果页面放置广告对用户体验的影响，广告链接的实际收效，以及不同关键词或位置的广告对用户的吸引力等问题.

实时性也是广告点击意图预测中一个必须考虑的因素. 为了达到提高广告点击率的目的，必须在用户提交查询时对查询意图做出预测，这也是当前广告投放遇到的挑战. 如上所述，基于查询词项的广告点击预测是目前最主要的预测方法.

本文采用国内著名商用搜索引擎 1 个月（2009 年 11 月）的用户查询点击日志进行分析，包括超过 200 000 000 条用户查询及相应的点击信息. 在处理搜索日志的时候，只考虑发生了点击（包括点击广告或点击返回的网页结果）的查询，而不考虑无点击的查询.

2　基于查询词内容匹配的点击预测模型

2.1　基本思想

搜索引擎根据用户查询词与广告的关键词匹配的程度及相应竞价来决定以什么顺序展示哪些广告. 因此对于如何预测哪些关键词引发广告点击的可能性比较大，一个直观的思路是：从搜索日志中挑出所有引发了广告点击的用户查询，统计每个词项在这些查询中的频度，按照频度的降序对词项进行排序，词项的位置代表了它引发广告点击的可能性.

然而，这种方法忽略了一个很重要的问题：某些词项不仅在引发了广告点击的查询中出现频度很高，在没有引发广告点击的查询中也会大量出现. 可见一个词项引发广告点击的可能性是个相对量. 因此本文将查询分为两类：引发过广告点击的和从未发生过广告点击的. 对所有词项，根据它们在两类查询中的频度进行排名，以及根据在两个排名位置的比较，来判断词项的广告点击意图.

更进一步地，本文的目的是预测用户提交给搜索引擎的完整查询的广告意图. 因而需要根据每个 term 的广告意图，通过一定的映射关系，计算出完整查询的广告意图.

2.2　模型描述

设 S 为所有查询构成的集合，对其中的查询进行中文分词，得到所有出现在 S 中的词项的集合 T. 将全体查询分为两个部分 S_1 和 S_2，其中：S_1 为引发了广告点击的查询集合，S_2 为未引发广告点击的查询集合.

对 S_i，统计 T 中的每个词项在其中出现的频度，并且按照频度的高低排名，形成词表 L_i. $|L_1| = |L_2| = |T|$. 对于 T 中的每个词项 t，获取它在 L_1 和 L_2 中的排名 $\mathrm{rank}(t, L_1)$ 和 $\mathrm{rank}(t, L_2)$，计算两个排名的比值 $v(t) = \mathrm{rank}(t, L_1)/\mathrm{rank}(t, L_2)$，并且根据这个比值的大小，对 T 中的所有词项按降序排列，得到词表 L_3. 其格式如表 1 所示.

表 1　查询词信息包含内容格式

term t	$v(t)$	t 在非广告查询中出现的频率	t 在广告点击中出现的频率

在基于查询词内容匹配的预测算法中，本文只需要前两项的信息，即词项和排名比值.

给定任意用户查询 Q，进行中文分词，得到一个词项集合 s，并定义一个映射 g 即

$$s = \mathrm{segment}(Q) = \{t_1, t_2, \cdots, t_n\},$$

$g(Q) = g(t_1, t_2, \cdots, t_n)$.

这样,就获得由若干词项组成的完整查询 Q 的广告点击倾向性的量化度量. 如果 $g(Q)$ 大于某阈值,则判定 Q 具有引发广告点击的倾向,反之则没有. 在实验分析中可以看到,映射 g 对预测算法的性能有一定影响.

2.3 实验结果分析

令 S 为 2009 年 11 月所有查询构成的集合,T 为在 S 中出现过的所有词项构成的集合,为了去除噪声以及过于稀疏数据的影响,除去那些稀有的词项或者发生广告点击次数过少的词项(实验中设为点击次数 < 10). 以后讨论中所用到的 L_3 均如此.

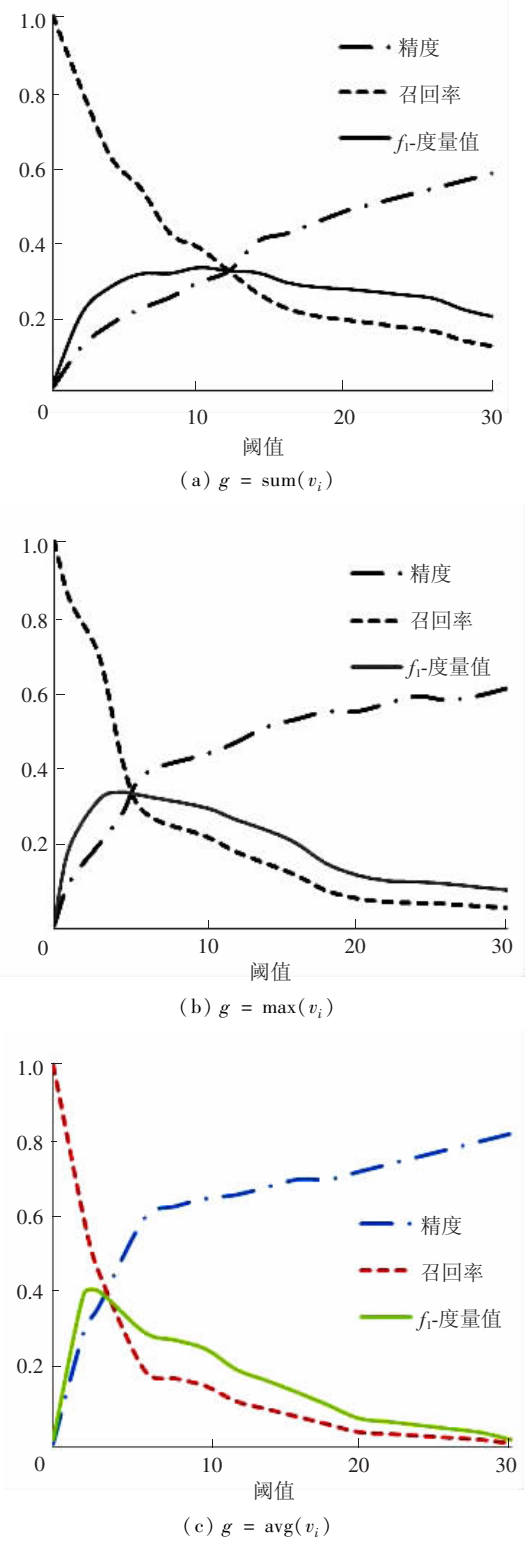
本文对真实搜索引擎任意一天(为保证开放测试,选取非 2009 年 11 月的日期,例如 2009 年 12 月15 日)的查询进行验证,预测这些查询是否可能引发广告点击,并根据实际发生的广告点击情况进行预测的精度与召回率等性能评价. 测试集含有约 7 000 000 条不同的查询,忽略当天点击次数 < 10 的查询,一共包括约 35 000 条引发了广告点击的查询.

为了保证判断的准确性,要求 t_1, t_2, …, t_n 中必须有 2/3 的词项出现在 L_3 中,即某些包含过多稀有查询词项的查询会直接被忽略. 由于广告商购买的都是一些比较常见的词项,因此这样做并不会影响对于用户查询意图的判断. 在模型描述中已经提到,g 为用来将词项的广告点击意图映射到完整查询的广告意图的函数,不同的 g 函数会影响预测算法的效果. 图 1 给出使用不同映射函数得到的预测效果对比, 表 2 给出使用不同映射方法得到的最优预测精度. 其中映射 g 分别为

$$g(\text{sum}) = \sum_{i=1}^{n} v_i,\ g(\text{max}) = \max(v_i),$$

$$g(\text{avg}) = \sum_{i=1}^{n} v_i/n.$$

如果不加预测而对于每个用户提交的查询都投放广告,那么精度 = 1.000 0,召回率 = 0.005 0,f_1- 度量值 = 0.010 0,结果劣于采用预测算法的情况. 基于查询内容匹配的模型对于非稀有查询的意图预测还是比较有效的. 采用映射函数 $g(\text{avg})$,能够使得广告投放的精度从 3.0% 提高到 36.8%,$f1$- 度量值从 0.060 提升到 0.408. 本文在更多日期上的预测实验表明各映射函数对应的阈值基本稳定. 其中 $g(\text{avg})$ 函数一般取阈值为 3.



(a) $g = \text{sum}(v_i)$

(b) $g = \max(v_i)$

(c) $g = \text{avg}(v_i)$

图 1 基于词表内容匹配的广告点击预测效果

表 2 基于词表内容匹配的广告点击最优预测结果对比

最佳结果	精度	召回率	f_1- 度量值
Baseline(不预测)	0.030 0	1.000 0	0.060 0
$g = \text{sum}$	0.290 5	0.381 1	0.329 7
$g = \max$	0.296 5	0.435 7	0.352 9
$g = \text{avg}$	0.368 5	0.456 5	0.407 8

3　基于朴素贝叶斯分类的预测模型

3.1　模型设计

对用户所提交查询的意图进行预测实际上也可以看做一个分类问题. 由此本文提出基于词项历史点击信息对用户查询意图进行分类的朴素贝叶斯预测模型. 所有查询被分为两类：不含有广告点击意图和含有广告点击意图，分别表示为 C_1 和 C_2. 分别计算每类查询出现的先验概率 $P(C_i)$，通过分别统计在所有查询中具有和不具有广告点击意图的查询的比例来估计；每个词项的条件概率 $P(t \mid C_i)$：可以通过统计 t 出现在 C_1 和 C_2 类查询中的频度估计. 将查询描述为其对应的词项的集合. 假设各词项独立，计算 $P(C_1 \mid q)$ 和 $P(C_2 \mid q)$，并考虑到 $P(q)$ 是一个常数，因此有

$$p(c_i \mid q) = \frac{p(c_i)p(q \mid c_i)}{p(q)} = \frac{p(c_i)\prod_{j=1}^{n} p(t_j \mid C_i)}{p(q)},$$

$$p_1 = P(C_1)\prod_{j=1}^{n} p(t_j \mid C_1),$$

$$p_2 = P(C_2)\prod_{j=1}^{n} p(t_j \mid C_2).$$

如果 $p_1 > p_2$，则 q 属于 C_1，不含有广告点击意图，这时应减少投放广告的数量甚至不投放广告；如果 $p_1 < p_2$，则 q 属于 C_2，q 更可能含有广告点击意图，应投放相关的广告.

3.2　实验结果分析

采用与上述同样的数据集，对朴素贝叶斯预测模型效果进行验证，如表 3 所示.

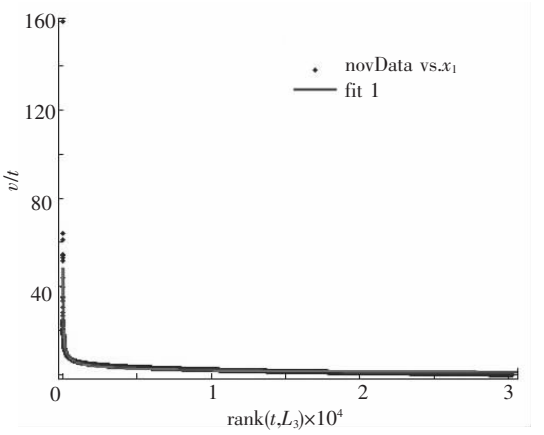
表 3　基于朴素贝叶斯的广告点击预测效果

日期	精度	召回率	f_1-度量值
2009 年 12 月 8 日	0.21	0.29	0.23
2009 年 12 月 15 日	0.20	0.29	0.23
2009 年 12 月随机采样（10% 比例）	0.17	0.20	0.18
Baseline（不预测）	0.03	1.00	0.06

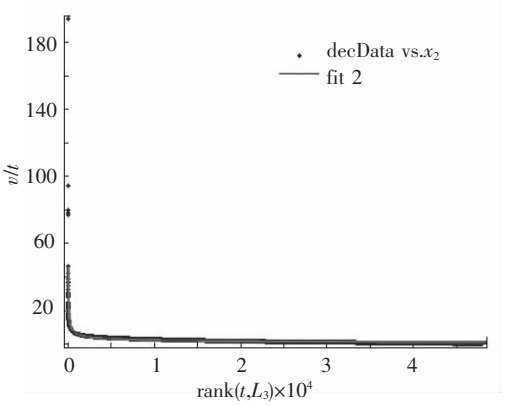
表 3 中列出了应用 11 月的数据训练而成的贝叶斯分类器对于 12 月部分查询数据进行测试的结果. 其中“2009 年 12 月数据随机采样（10% 比例）”为从 2009 年 12 月的所有查询中按照 10% 的概率随机抽取一部分查询作为测试集. 预测的综合效果（f_1-measure）虽然比不预测有很大提高，但是并没有达到与基于词表匹配的模型预测性能.

4　广告点击曲线拟合模型

为了验证使用基于历史数据是否能够对新的广告点击意图进行有效预测，本文进行了广告点击的曲线拟合，即考察词表与相对排序关系的稳定性，如图 2 所示.



（a）2009 年 11 月数据拟合结果



（a）2009 年 12 月数据拟合结果

图 2　2009 年 11 月和 12 月的查询词项广告点击数据拟合曲线

图 2（a）是 11 月数据的拟合结果. 曲线解析表达式为 $f(x) = ax^b$，$a = 114.5$，$b = -0.4399$，拟合误差 SSE $= 1.4\mathrm{e}+004$，$R^2 = 0.9106$. 其中 SSE 为误差平方和，值越小说明拟合的误差越小；R^2 为决定系数，常用来衡量曲线对真实数据点的近似程度，$R^2 = 1$ 则为精确匹配实际情况。图 2（b）为 12 月数据的拟合结果，拟合误差 SSE $= 2.047\mathrm{e}+004$，$R^2 = 0.902$，拟合系数 $a = 129.7$，$b = -0.4432$. 拟合结果说明对于不同的两个月的真实用户查询点击日志，拟合所得到的参数具有一致性，分布平稳，且 rank(t, L_1)/rank(t, L_2) 随排名的降低而迅速减少，因此验证了所提出方法的可行性. 首先，过滤低频不会对结果造成很大影响；其次，数据稳定，所以可以用以前的数据预测进行分析从而预测未来的点击情况.

5 结 论

1) 提出了基于查询词内容匹配的点击和基于朴素贝叶斯分类的两种预测模型;

2)通过实验验证,表明两种模型均能改进广告投放效果,其中第 1 种模型效果更佳;

3)通过对不同月份的查询词项的分布进行了曲线拟合,验证了所提出预测方法的可行性. 该方法可适用于用户查询广告点击意图的在线预测.

参考文献

[1] FAIN D C, PEDERSEN J O. Sponsored search: a brief history [J]. Bulletin of the American Society for Information Science and Technology, 2006, 32(2): 12 – 13.

[2] RIBEIRO-NETO B, CRISTO M, GOLGHER P B, *et al.* Impedance coupling in content-targeted advertising [C]//Proceedings of the 28th Annual International ACM SIGIR conference on Research and Development in Information Retrieval. New York, NY: ACM, 2005: 496 – 503.

[3] BRODER A, FONTOURA M, JOSIFOVSKI V, *et al.* A semantic approach to contextual advertising [C]// Proceedings of the 30th Annual International ACM SIGIR conference on Research and Development in Information Retrieval. New York, NY: ACM, 2007: 559 – 566.

[4] 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. New York, NY: ACM, 2008: 417 – 426.

[5] ANAGNOSTOPOULOS A, BRODER A, GABRILOVICH E, *et al.* Just-in-time contextual advertising [C]// Proceedings of the 16th ACM conference on Conference on Information and Knowledge Management. New York, NY: ACM, 2007, 331 – 340.

[6] DEBMBSCZYNSKI K, KOTLOWSKI W, WEISS D. Predicting ads click-through rate with decision rules [C]//Proceedings of the Workshop on Target and Ranking for Online Advertising. New York, NY: ACM, 2008: 578 – 586.

[7] ASHKAN A, CLARKE C L A, AGICHTEIN E, *et al.* Characterizing query intent from sponsored search clickthrough data[C]//Proceedings of the Workshop on Information Retrieval and Advertising. Singapore: SIGIR – IR, 2008:15 – 22.

[8] REGELSON M, FAIN D C. Predicting click-through rate using keyword clusters[C]//Proceedings of Second Workshop on Sponsored Search Auctions. New York, NY: ACM, 2006: 1 – 7.

[9] 陈磊,刘奕群,茹立云,等. 基于用户日志挖掘的搜索引擎广告效果分析[J]. 中文信息学报,2008, 22(6),92 – 97.

[10] 王家卓,刘奕群,马少平,等. 基于用户行为分析的竞价广告效果分析[J]. 计算机研究与发展,2011, 48(1): 133 – 138.

(**编辑** 张 红)