北京大學

硕士研究生学位论文

题目：**搜索开放平台广告的点击率影响因素分析**

姓　　名：	吴恩宇
学　　号：	1301210997
院　　系：	软件与微电子学院
专　　业：	软件工程
研究方向：	电子商务与物流
导师姓名：	李杰　　教授

二〇一五年七月

摘要

搜索开放平台广告是近五年迅速发展起来的搜索引擎广告。搜索开放平台广告使得暗网数据被搜索引擎捕获，也使得广告主的数据前置，搜索用户获取信息的路径变得更短，极大地方便了使用者。但作为搜索引擎公司商业产品的一种，搜索开放平台广告还存在着上线前无法确定商业价值的弊端，公司的成本与收益具有极大的不可控性。导流能力往往是衡量搜索开放平台广告商业价值的重要标准，因此需要对搜索开放平台广告的点击率进行预估。

围绕着搜索引擎公司在上线前需要尽早确认上线广告效果的需求，本文的主要研究内容包括：1)对搜索开放平台广告进行介绍，分析流程中存在的弊端，提出问题 2)通过阅读其他学者研究文献及访问搜索引擎公司领域专家，确定搜索开放平台广告点击率与展现样式（图片、按钮、输入项)、关键词、展现量、广告位排序以及网站知名度有关，通过一定的方式将这些因素量化，作为分析点击率影响因素的指标 3）在分析搜索引擎广告与预测方法的前提下，提出了贝叶斯网络算法预测的思路，基于前面收集处理的因素数据对搜索开放平台广告的点击率预测进行了贝叶斯网络的搭建，而后在该模型的基础上引入完全主观变量用户体验（UE）搭建了新的模型，通过实证数据对两个模型进行了检验，数据表明后者较前者的准确率有极大的提升 4)利用研究发现的一些结论对现在的搜索开放平台广告上线流程提出优化建议。通过预测模型不单单可以使搜索开放平台广告的商业价值变得更加容易估计，也使得搜索开放平台广告在上线前就可以进行优化，从而实现“上线即最优”的结果。

关键词：搜索开放平台广告，影响因素，点击率，预测，优化

Analysis on Influence Factors of Search Open Platform Advertisement Click Through Rate

Wu Enyu (Software Engineering)

Li Jie

ABSTRACT

Search Open Platform Advertisement is one kind of advanced Search Engine Advertisement in the recent five years. Search Open Platform not only make search engine get the data of hidden web, but also shorten the way in wich search engine user get the knowledge they need,which facilitate people a lot.However,as one kind of business product for search engine company, Search Open Platform Advertisement still has a weakness that its click through rate can't be predicted before it is launched on the web.It means risk for company.So it is a must to predict the click through rate of Search Open Platform.

In this paper, I've mainly studied the following: (a). Introducing Search Open Platform Advertisement and analysize the weakness before the advertisement is launched (b).By communicating with some experts on search engine company and reading documents,I ensure that the click through rate of search open platform advertisemet has relationship with its' appearance(Picture,Button,Input), keywords, page view,rank,web fame.I collect some data about these factors to mearture them. (c).On the basis of studying search engine advertisement and predicton methods, I come up with a idea to predict the click through rate with Bayesian Network on the basis of some objective factors.When the pure subjective factor User Experience is brought into the model,the new one has a better performance over the old one(d).Make some optimization suggestion to the preparation procedure of search open platform advertisement.

KEY WORDS： Search open platform advertisement,Influence factor,Click through rate;prediction; optimization

目 录

第一章　绪论

1.1　研究背景及问题提出

1.1.1　研究背景

最早的互联网始于 1969 年的美国，是美军在阿帕网制定的协定下，首先用于军事连接，后将美国西南部的加利福尼亚大学洛杉矶分校、斯坦福大学研究学院、加利福尼亚大学和犹他州大学的四台主要的计算机连接起来。互联网的出现固然是人类通讯技术的一次革命，然而，如果仅从技术的角度来理解互联网的意义显然远远不够。互联网的发展早已超越了当初 ARPANET（阿帕网，美国国防部研究计划署）的军事和技术目的，几乎从一开始就是为了人类的交流服务的。随着互联网的逐渐发展普及，越来越多的服务器与计算机被连接进入了互联网络之中，互联网的民用价值以及商用价值也开始逐渐体现。

在日益繁多的互联网数据面前，人类的懒惰性与智慧促使了网络搜索的诞生。1991 年，第一个连接互联网的友好接口在 Minnesota 大学被开发出来，当时学校只是想开发一个简单的菜单系统可以通过局域网访问学校校园网上的文件和信息，用户需精确地输入FTP文件名进行匹配,而后由服务器对用户反馈用户所找文件的FTP具体地址[1]，用户自行前往下载。就在同一年，前面提到的创意启发了 Minnesota 大学的一位学生，他尝试在文件中进行纯文本的搜索，于是基于用户关键词进行搜索的 Gopher 就这样诞生了。爬虫程序的诞生促成我们现代搜索引擎的出现，MIT（麻省理工大学）的马修开发出来的用于信息检索的爬虫程序，最初是用于监测互联网的相关信息数量以进行网络规模统计，而后逐渐演变成抓取 URL。在 1994 年，将爬虫程序引入到搜索程序当中 Lycos 出现了，它可以当之无愧地被称为现代搜索引擎始祖。如今鼎鼎大名的搜索引擎 Google（谷歌）起源于斯坦福大学的一个名为 BackRub 的小项目，后由拉里佩奇创建了谷歌公司。支持精准搜索和泛搜索的 Google 在众多功能领域中集成搜索，使搜索引擎进入了一个崭新时代。发展至今，已经不单单只支持文字搜索，对图片搜索、声音搜索的功能也进行了拓展，可以说谷歌在搜索引擎领域已经走在了世界的最前列。在中国，由于文化及政策的种种原因，谷歌没能成功占领中国市场，而百度则占据了中国最大的搜索引擎份额。

从互联网的诞生到搜索引擎的诞生，从一家搜索引擎公司建立到多家搜索引擎公司并存，随之而来的是网络用户的便利以及公司的商业价值。搜索引擎带给用户最大的便利就是使人们可以快速地从海量互联网数据中找到自身所需，在海量数据中，用户的细化需求被搜索引擎匹配，用户在少量且排列有序的有效数据中必然可以更轻松

地满足自身需求。而搜索引擎公司正是在满足这种诉求的同时，发明了一系列通过搜索引擎盈利的方式。而搜索引擎营销恰好是搜索引擎公司创造出来的首要盈利模式，简单地说它是基于搜索引擎平台帮助广大广告主进行广告的投放，辅助广告主在互联网平台上进行推广从而达到广告主售卖服务或者售卖产品的目的。在诸多的推广渠道中，互联网已经成为不可忽视一条渠道。有调查报告显示 2014 年中国网民人均上网时间已经达到 26.1 个小时，网民搜索引擎使用率达到 80.3%[4]，网络及搜索引擎使用的广泛性使得广告主将营销内容通过搜索引擎传递给潜在客户日益成为可能[2]。从搜索结果的呈现的类别来看，笔者认为搜索引擎营销结果可以分为以下四类，分布位置示意图如下：

图 1.1 常见搜索结果示意图

第一类,品牌直达广告（又名品牌专区）。固定位于搜索结果左侧第一位，见图 1.1 中的 A 位置。展现形式多以图片混合文字输出，占据搜索结果首屏二分之一左右的面积，形式新颖且突出显眼，对互联网用户具有较强的吸引力。该结果的触发词要求为广告主品牌词及广告主产品词，搜索引擎公司会对该类结果上线前对广告主进行资质审核，审核通过也就意味着搜索引擎公司对广告主的资质背书。因为品牌直达广告的位置显眼突出且网站可信度具有较高保障，所以极受品牌类广告主的青睐。

第二类，品牌地标。固定位于搜索结果右侧第一位，见图 1.1 中的 B 位置。美国著名网站设计工程师杰柯伯·尼尔森发现网络用户大多不由自主地以“F”字母形状的模式来阅读网页，见图 1.2，图中颜色越深说明浏览用户目光集中度越高。

图 1.2 常见网页热感仪分析图

研究发现，网民已经形成了一种普遍的阅读习惯，即无论是电商网站还是搜索结果页，浏览者的视线将不自觉地以 F 型游走从而达到迅速捕获重要信息的目的[3]。通过对比图 1.1 和图 1.2，我们发现搜索引擎营销广告的布局与网民浏览习惯的相当吻合，在网民的视线集中区域均存在商业广告布局，有效地贴合了搜索用户的使用习惯。搜索引擎公司恰恰是利用上述规律，在搜索结果的右侧第一位（F 的上边一横的右侧）发明了品牌地标广告。该广告对搜索触发词的审核并不像品牌专区那般严格,对某些通用词同样可以触发，尽管因为处于右侧的原因点击率并不高，但是却因为具有更强的曝光而受到一些追求曝光的广告主偏爱。

第三类，竞价广告（又名关键字广告）。通常位于品牌专区下侧，见图 1.1 中的 C 区域。关键字广告可以说是搜索引擎营销收入的来源主体。顾名思义，广告主针对一些词或者字打广告，这些词或者字被称为关键字，当用户搜索关键字时，关键字广告被触发，展现在搜索结果页当中，用户每通过该广告点击跳转到广告主链接处一次，搜索引擎向广告主收费一次，这种收费形式被称为按点击收费。而竞价广告之所以被称为竞价广告的原因就是搜索引擎为了不使商业结果过多而破坏了正常的搜索结果，对竞价广告条目进行了限制，当多位广告主竞争该广告时，搜索引擎将按广告主对每一次点击的支付意愿价格进行排序，展现排名前几位的广告。因为竞价广告不生效不收费的特性以及灵活的收费方式，竞价广告是广告主最愿意使用的一种广告手段。

第四类，搜索开放平台广告结果。搜索开放平台广告结果是搜索引擎对正常搜索展现结果的一种扩展，通常位于正常搜索结果之中，排位并不固定，但由于其展现形式特殊，较普通搜索结果有更强的吸引力，所以还是会在排位的加权算法中逐渐排升到靠前的位置（图 1.1 示意图中的 D 位置在其他搜索结果之上）。

无论是互联网搜索引擎的普及性还是搜索结果的多样性，都在反复地说明着一个不争的事实——搜索引擎营销是广告主推广自身产品或者服务，提升其知名度的一种有效途径。截至 2014 年六月，我国搜索引擎用户规模达到 5.07 亿，占网络用户的比例为 80.3%[4]。广泛的互联网覆盖，较高的搜索引擎使用率使广告主们看到了隐藏在搜索引擎之后的大量潜在客户，而各种各样的广告形式则使广告主可以实现预期需求。但是在搜索引擎上打广告和传统的电视广告、广播广告还是有很大差异的。电视广告或者是广播广告都是灌输式的传播，潜在消费者属于被动接收信息，而且也不存在消费者与广告之间的互动，衡量广告好坏的其实只是创意是否有吸引力或者能让消费者记忆深刻，广告主不需要去关注一些其他因素。而互联网的搜索引擎广告则不然，企业的广告效果受到广告形式、排名、展现形式等多方面因素的影响，针对搜索开放平台广告的点击率研究就是在这种情况下提出来的。

1.1.2 问题的提出

搜索开放平台是搜索引擎公司对数据拥有者开放的一个数据共享平台，其最大的功能是帮助互联网公司捕获自然抓取不到的“暗网”数据。之所以称之为开放平台，一方面是搜索开放平台的获取数据全部都开放给搜索用户，另一方面是因为搜索引擎方会在后台提供一个用户提交结构化数据的平台，开放给广告主或开发者。数据提交者通过提交数据，可以使自己的数据在搜索结果页以某种形式直接展现给搜索引擎用户，缩短了搜索用户信息获取路径，提升了搜索引擎使用体验。而搜索开放平台广告正是基于搜索开放平台的商业化提出来的一种广告形式，通过对广告主售卖独特的搜索开放平台结果的展现样式的使用权限，从而达成商业化的目的。

搜索开放平台广告的触发规则和之前提到的品牌专区、竞价广告是相同的，都是通过搜索关键字进行触发。但是展现形式却更加的灵活，因为该产品出现的目的就是为了优化用户体验，所以针对不同的数据形式会出现不同的展现结果。例如在图 1.3 中左侧为搜索开放平台对彩票的展现样式，右侧为菜谱的展现样式。对天气、机票、视频等各种搜索词，搜索开放平台会以用户体验第一为根本，设计出符合数据特性的展现形式，因此对于一些不单单满足于单纯的品牌曝光的广告主来说，展现形式更具体更细致的搜索开放平台广告无疑是一种更好的广告形式。搜索开放平台广告使广告性质变得更隐蔽，更不容易使用户感觉突兀，也更容易搜素引擎用户在未点击进来之前就对广告主的某些产品产生认识了解，进一步地也使得搜索用户在广告点击之后形

成购买亦或深入搜索广告主的信息的意愿变得更强，基于以上这些优点，搜索开放平台广告业广受广告主的偏爱。

图 1.3 搜索开放平台广告样式（彩票、菜谱）

作为一种可商业化的数据展现形式，搜索开放平台广告的收费方式有两种：一种是像竞价广告一样，按点击收费，但是因为对搜索开放平台广告来说，每一个结果一般都只展现一家客户的数据，不存在多家竞争关系所以点击价格只能由搜索引擎公司来制定，价格的高低、价格维持的时间长短等问题就会接踵而至。此外，和竞价广告相比搜索开放平台的展现形式更加完善细致，如果按点击收费则使广告主只关注点击效果反而忽略了曝光等品牌露出效果。综上所述，搜索开放平台广告一般不会采取按点击收费的策略，至多把这种收费方式放在广告上线初期对广告主进行短期测试，而后还是会转成另一种收费方式。另一种收费方式则是按 CPT（Cost Per Time）收费，即按上线时长进行收费。搜索引擎公司综合考虑搜索开放平台结果展现的次数、可能的点击量以及对其他搜索广告收入（主要是竞价广告）的影响，制定一个该搜索结果在线上维持一定长度时间（以月为单位）的打包价格售卖给广告主。和按点击价格收费相比，搜索引擎公司固然还需要考虑价格制定等因素，但是该项工作只需在上线初期制定，而后只需维持搜索结果的正常展现即可，因此避免了许多需要人力耗时的工作，节省了大量人力。目前大量在线的搜索开放平台广告都是采用这种按时长收费的方式存在的。

对搜索引擎公司来说，前面提到的搜索开放平台广告收费只是广告上线较后期的工作，而搜索开放平台广告在收费前的工作才是重中之重。2009 年，国内的百度公司最先创建了搜索开放平台并将之命名为“阿拉丁”，至今为止，国内几大搜索引擎公司先后推出了相同产品，360 搜索将之命名为“one box”,搜狗搜索则称之为“vertical result”，但是无论是哪个公司的哪个称呼，搜索开放平台广告的本质是不变的,各网站站长或者广告主这些独特信息数据的拥有者需要提交数据给搜索引擎，从而解决现

有搜索引擎无法抓取和检索的暗网信息，同时也缩短了网络用户的搜索路径，使搜索用户实现真正的“即搜即得”。为了实现这一目的，无论是哪个搜索引擎公司，都需要经历一些不可避免的工作，概括的讲就是作为搜索引擎方，需要审核网站站长或者开发者的数据，并根据数据的不同确定不同的展现样式，尽可能以最优化搜索用户体验的方式向用户呈现数据。将上线前的流程具体细化，则在搜索开放平台上线过程中的大致工作可以用下面的流程图来描述，其中或因不同的搜索引擎公司有所差异，但以下基本环节都是必不可少的：

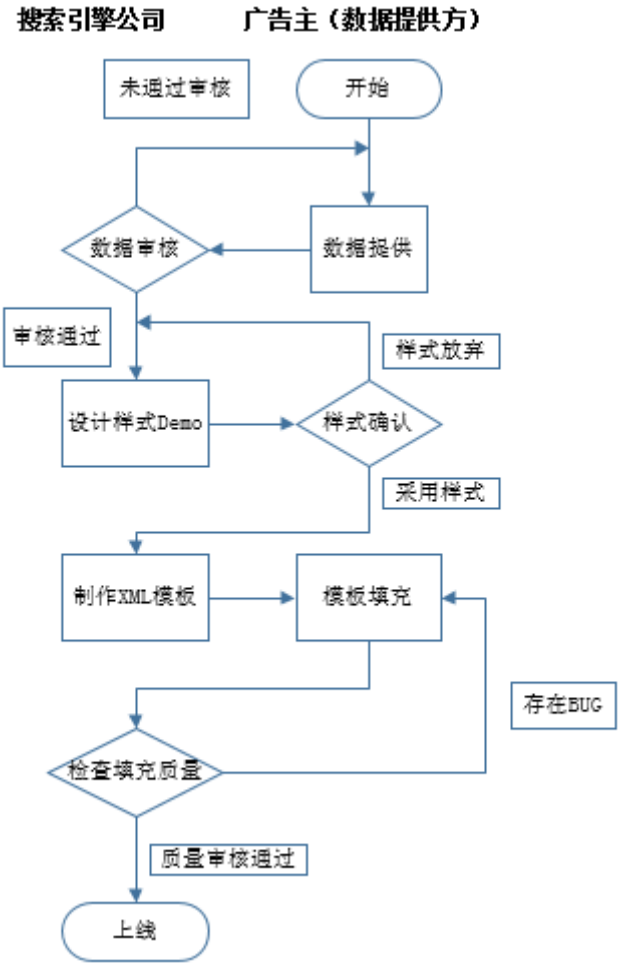

图 1.4 搜索开放平台广告上线流程图

通过上面的流程图我们可以看到，在收费上线前的步骤相当多，且可能存在多次搜索引擎公司与数据提供方的反复交互。根据实际经验，从数据提供方提交数据开始，到搜索开放平台广告正式在搜索用户页面呈现，整个流程往往需要一个月甚至更久。而在搜索结果未正式上线前，无论是数据提供方还是搜索引擎公司都无法对搜索开放平台广告的最终的展现效果——点击量亦或点击率进行准确预估，这种现象在飞速发展的互联网行业无疑是不能容忍的。和传统的电视广告或者广播广告相比，搜索引擎

营销广告最大的优势就是拉近与目标消费者的距离。广告已经远远不是让消费者看到或者听到那么简单，广告主更希望能让这些看到广告并对广告情境感兴趣的目标人群更深入的了解信息，最终产生消费效果。

综上所述，搜索开放平台广告效果是每一个利益相关方都关注的问题。衡量搜索广告效果好坏的一个重点指标就是点击量，对广告主或数据提供方来说点击就意味着流量、潜在消费者甚至是公司的前景，对搜索引擎公司来说点击则意味着商业价值。搜索引擎公司希望可以提前预测点击量来确认商业化价值，数据提供方希望对点击量进行预测以提前对广告效果进行预期确认广告投放是否有价值，同时也可以有条不紊地应对流量上涨，因此尽可能准确地对搜索开放平台广告点击量进行预测是十分重要的。

基于上述的真正需求以及作者自身的研究兴趣，本文将对搜索开放平台广告的点击量影响因素进行分析。为了使点击量这种绝对数值变得更具比较意义，对点击量的研究则被转换为对点击率的研究，点击率是点击量与展现量之比，数值在零到一之间，是一个相对数值，不会受到不同搜索开放平台的展现量具备不同大小的影响，因此更具有普适意义。

1.2　研究目的与意义

从 2009 年百度最先提出搜索开放平台广告的概念至今已经五年有余，越来越多的数据拥有者都加入到了搜索开放平台广告的行列当中。毫不夸张的说，在未来任何搜索结果都可能会存在搜索开放平台的广告。日益增多的搜索开放平台广告势必要求搜索引擎公司对搜索开放平台广告上线流程标准化、模块化。众所周知，流水线的生产方式使福特汽车公司上个世纪声名大振，流水线之所以可以实施就是老福田将汽车生产的环节进行了标准化模块化处理。同样的，对于搜索引擎公司来讲，如果想将搜索开放平台这种广告商业化做的更成功，必须要对流程进行标准化并规避掉不可控因素。

如果搜索开放平台广告直到上线时搜索引擎公司才知道实际的广告效果的话，往往会面临两个潜在问题：一是根据实际效果报价后发现价格过低，根本无法弥补之前投入的人力成本，但是工作已经付出成为沉没成本，唯有获取收益才是最优结果，所以即使价格不足以弥补前期人力投入也只能上线。二是搜索引擎公司开出价格后广告主并不愿意支付，那么问题就可能因广告主砍价将问题转变为上一个问题或者因为双方无法接受协调价格最终放弃上线，如果是后者的话那么前期投入的所有成本的收益都将是零。

为了防止上述情况的发生，提前对搜索开放平台广告的点击率进行预测具有非常重要的意义。拥有了点击率，搜索引擎公司可以提前预测投入的人力成本是否可以取

得令人满意的收益，也可以提前就价格与广告主进行沟通，如双方无法对价格达成一致也就无需进行无收益的合作。这使得搜索引擎公司可以更有效的使用人力，同时也会更具效率地对广告上线流程制定相应的标准，方便标准化操控。

1.3 国内外研究现状

目前国内不同搜索引擎公司对搜索开放平台广告的称呼皆有不同，分别以易记的代号称呼，如百度称其为阿拉丁，360 搜索称其为 one box,搜狗搜索称其为 VR，但是无论针对哪个公司的搜索开放平台广告，都没有相关的点击预测研究。但是，搜索开放平台广告的本质仍是搜索广告的一种，从相似性上来看，竞价搜索广告与搜索开放平台广告最相似。因此，本文将从竞价搜索广告来对国内外研究现状进行分析。

竞价搜索广告，又名关键词广告，如前文提到的那样广告主通过对不同的关键词进行竞价，实现在搜索引擎竞价关键词被搜索时的广告位排序争夺，用户根据自己的偏好和需求点击相应的广告，广告链接就会将用户链接到相应产品或服务的网站页面上，由此引导用户做出消费选择。这种竞价机制秉承的一个基本理念就是排位靠前的广告将获得更高的点击量。和竞价搜索广告相比，搜索开放平台广告除了不需要提供竞争价格以外，同样需要提供关键词、展现内容等其他信息，它们的目的都是希望通过吸引点击最终引导用户消费或者注册等行为，所以用竞价搜索广告的研究现状来分析搜索开放平台广告的研究现状是非常恰当的。当前，对竞价搜索广告的研究根据人群被分为了三个方向：从搜索引擎用户角度研究、从广告主角度研究以及从搜索引擎公司角度研究。

1.3.1 从搜索引擎用户角度研究

这一研究方向是站在搜索引擎用户的角度，分析搜索用户与搜索广告的之间的相互影响。在搜索引擎的搜索结果影响着搜索引擎用户的同时，搜索引擎的使用者们也在逐渐改变着搜索引擎。例如当尼尔森发现的互联网用户的 F 型网页浏览习惯[3]并发表了该研究之后，各种网站的页面设计师都开始将网页的关键目录以竖排目录的方式在网页左侧展现，并极尽可能地利用网页的顶端位置。国内的童强通过数据仿真以及问卷调查的方法获取了相当量的数据，通过数理统计的方法对数据进行了分析，研究了网民对搜索广告的信任度、态度与广告点击行为之间的相关性。他的研究最终表明，广告的排序位置、推广内容、网民的社会属性以及广告的展现形式等因素都会影响搜索竞价广告的广告点击行为[5]。雅虎实验室的 Erick Cantu-Paz 和 Haibin Cheng 基于大量搜索用户点击竞价广告的行为日志样本，抽取了除人口统计学特征之外的一些可以反映搜索用户点击可能性的特征，把这些特征融入非个性化点击预测模型。通过该

模型，雅虎把模型判定搜索用户点击可能性最大的的广告推送到竞价广告的最高位，成功地提升了竞价广告的精准度，使搜索量得到了更大化的变现[6]。Bernard 和 Simone 基于美国一个大型零售商的搜索引擎营销数据进行数据分析，引入评估购买漏斗模型，研究归属于处于购买漏斗模型不同阶段的关键词对消费者的影响。研究结果表明，分类处于漏斗模型第一阶段意识（具体阶段分为意识、研究、决策、购买）的词在搜索广告活动中更有效果[7]。

1.3.2 从广告主角度研究

对搜索引擎广告的另一个研究方向是以广告主的角度来研究。国内的许建盈对竞价广告进行了研究，他基于广告主在多家搜索引擎公司的的竞价广告投放历史数据建立了贝叶斯网络模型，针对不同家搜索引擎公司数据（谷歌、雅虎、百度）分别进行了参数学习，对广告投放收益进行了有效预估[8]。引擎搜索排名的机器学习过程一般需要人工编辑对数据进行标签分类处理，耗时且繁琐。事实上，点击日志同样可以被当做是引擎搜索排名的一种隐性反馈，可以在无需花费过多的前提下近似替代人工编辑的标签。ChaPellehe 等提出了一个动态贝叶斯网络模型从而可以根据点击日志提供无偏的相关性预测，即利用链条式网络来分析用户行为的广义级联模型，在他们的研究中是以用户查看下一项结果的概率与广告位置及广告页面相关为假设前提，研究的结果证明该点击模型在点击率预测和相关性上大大超出了其他已知模型[9]。

1.3.3 从搜索引擎公司角度研究

对搜索引擎公司来说，搜索广告的相关数据的获取更加便利，视角也更加全面，可以覆盖到各个角度，因此可研究的方向也更加的丰富，这也是本课题所要研究的主体方向。斯坦福大学的 Jason 和 Mukund 利用了 Yahoo webscope 的数据对广告主的投放目标进行了研究。根据关键字出价成本以及广告主的收益数据，他们证明了经济学中做出的理性人假设—收益最大化策略确实存在于竞价行为当中[10]。哈佛大学国外的 Benjamin Edelman 和斯坦福大学的 Michael Ostrovsky 通过对 Overture 和 Google 的竞价搜索中竞价者的策略数据研究发现，如果可以提供另一种可以减少广告主竞价策略行为的机制，那么搜索引擎方将获得更大的收益，而整个市场也将更有效率。同时他们还指出，广告主的竞价策略行为并不会随着时间消失而会存在于搜索引擎中，最终形成一种动态的平衡[11]。在竞争激烈的互联网广告环境，搜索引擎提供商们对广告者提供专业的咨询建议服务将成为各自的优势，而对广告投放的预测也成为了各自提供的建议中比较重要的一个部分。Gluhovsky 对建立预测模型进行了研究，基于传统的单调回归模型开始对广告位置以及广告点击效果分别进行分析，最终对用极大似然估计的方法对后者建立了一个更新颖的模型。该模型涉及到单个广告主对期

望的点击率和点击量的竞价问题，根据该模型制定投放策略的效果显著超出传统单调回归模型[12]。

本文就是站在搜索引擎公司的角度，以预测搜索开放平台广告的点击率为目标，运用贝叶斯网络预测模型作为研究方法的搜索开放平台广告影响因素的分析研究。本文在对搜索开放平台广告上线前的准备流程进行分析后，确定了在搜索开放平台广告上线前可以提前获取的变量以及可以人工操作的变量，构建出基于贝叶斯网络的预测模型。基于该模型可以在搜索开放平台广告上线前有的放矢地进行分析，便于搜索引擎公司对产品上线的商业价值把控。同时，在研究搜索开放平台广告的点击率影响因素，也可以对可人工操纵的某些变量进行提前优化，使产品上线初期即达到较优的效果。

1.4 研究内容与思路

1.4.1 研究内容

本文是站在搜索引擎公司的角度对搜索开放平台广告的点击影响因素研究分析，在总结其他学者的研究成果的前提下，基于自身在搜索引擎公司掌握的一些数据信息，确定影响搜索开放平台点击数据的因素以及这些因素之间的相互关系，并将这些因素作为相关变量构建贝叶斯网络预测模型。构建好预测模型后，利用一定量的数据进行验证，然后对模型进行参数学习。根据变量的实际值与预测值的对比来判定预测模型的有效性。最后根据研究出的影响因素，更具普遍性地分析搜索开放平台上线前需要侧重的东西。

研究的重点在于建模预测部分，建模部分中有以下最重要的三个环节:

第一环节是模型的因素选择。因素选择的主要目的在于确定可以影响点击的因素。方法是根据前人的研究基础，基于以往文献研究以及搜索引擎公司专业人员多年实际经验总结来确定模型涉及到的相关变量。通过对搜索广告的研究分析，可能影响点击的因素有展现效果、搜索相关性、广告位置、网页质量等[13]。通过对上述一些因素的细化或者替代，本文中贝叶斯网络模型涉及到的相关变量有填充关键词量、搜索触发词上线前展现量、数据填充方的网站知名度、搜索开放平台广告的搜索展现排序位置等。

第二环节是贝叶斯网络的构建。完整学习的构建方法是无论是变量节点、贝叶斯网络结构还是分布参数均由人工给出，完全受人工经验指导。部分学习的方法是结构和分布参数由计算机根据历史数据给出，其他部分由专家指定。当然，在上述两种方法之间还存在着一种“中间态”：仅通过机器学习的方式获取贝叶斯网络的参数，其余未知信息人为给出[14]。

相比较来看，方法一过于受限于专家知识，由于人类可获取知识的可能性和有限性，完全由专家指导构建的模型往往与实践中积累下的数据存在很大偏差。方法二完全是一种数据驱动的方法，具有很强的适应性，但是得出的模型有可能不能被人所理解。方法三则是前两种方法的折衷方法，当变量之间的关系明显的情况下，这种方法能大大提高学习的效率。

因此，在本文中的贝叶斯网络模型构建将用专家法确定网络结构，再使用参数学习的方法构建参数模型。

第三环节是贝叶斯网络推理。贝叶斯网络推理，即利用推理算法对测试数据进行推理验证网络模型的预测效果。和最基本的贝叶斯网络推理算法——变量消元法相比，团树传播法的主要优点是它使得两次不同推理的中间结果可以共享，因此，当需要做多次推理时，团树传播法比变量消元法更为合适。本文中贝叶斯网络推理将使用团树传播法来进行。

1.4.2　研究方法

本文的研究方法主要运用了以下三种方法：一是文献分析方法。通过对国内外相关文献的收集研究得出现有的搜索广告的理论基础及发展现状，并对贝叶斯网络模型方法的文献资料进行收集、整理、归纳、分析、提炼总结，深刻领悟该方法在搜索开放平台广告点击预测方面的应用可行性及意义，基于贝叶斯网络模型及搜索广告相关理论完成对搜索开放平台广告点击预测的模型框架。二是统计分析方法。在对所获取到的数据进行粗洗，剔除无效数据处理后，根据需要把一些连续型数据分割为离散型数据。然后，运用统计软件对收集到的搜索开放平台广告相关数据进行整理分析，对不同的变量之间的关系进行统计检验，得出支持本文模型的结论。三是模型分析方法。基于构建出的贝叶斯网络预测模型，提出搜索开放平台广告上线的运营建议，譬如如何调整一些因素尽可能地提升搜索开放平台广告的点击数量，提升搜索引擎广告展现效果，实现搜索引擎方与数据提供方的共同效益最大化。

1.4.3　技术路线

本文的技术路线图如下图 1.5 所示。

在以下技术路线图中，分析问题确定研究思路及建模方法这一环节主要是在明确预测点击是解决问题的关键点的前提下，确定预测的方法。在这里，预测的方法不单单指算法的选择，也包括了工具的选择。本文从诸多预测方法中，选择了贝叶斯网络的方法，同时因为 matlab 中具有贝叶斯网络工具包，所以借用了 matlab 的环境来实现预测。

因素的选择与专家知识确定结构这两个步骤其实存在着工作交叉，本文的因素选

择主要是依据论文文献的参考以及搜索领域专家的建议，而在专家建议因素时也给出了因素之间可能的因果关系，因此结构也就有了一个雏形。

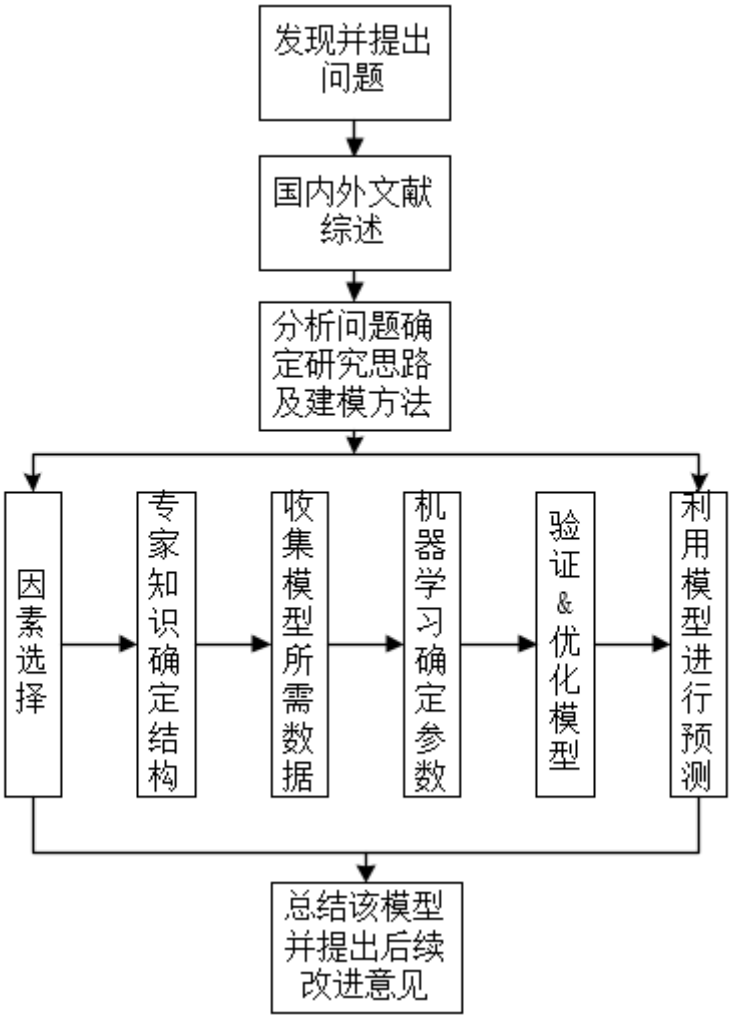

图 1.5 技术路线图

根据专家法确定的结构模型，通过对收集到的历史数据用最大似然估计的方法对参数进行学习，最终获得了模型。在通过建立好的贝叶斯网络模型对数据进行验证后，查找可以进行优化的点进行优化，这也就是验证与优化模型这一步骤。

最后将所得模型运用到实际数据中，也就是利用模型进行预测。

第二章　理论基础

2.1　搜索引擎广告理论基础

2.1.1　传统搜索引擎广告

从竞价广告诞生起，它就被当做了搜索引擎公司的主要盈利来源。即使是商业广告形式逐渐繁多的今天，竞价广告仍然是搜索引擎公司收入的主要来源。广告主通过对某个关键字出价使得该广告主投放的广告内容在搜索结果页呈现给搜索该关键字的用户，之所以称为竞价是因为对同一个关键字可能由多家广告主购买，当出现这种情况时，不同广告主的广告将按出价排序，出价最高者排序在第一位。超出广告展现条数的竞价排序位置不予展示。下图 2.1 是一个简单的竞价广告示意图，如果三个广告主分别对“鲜花”，当搜索用户搜索“鲜花”时，搜索引擎公司瞬间排序对“鲜花”出价的广告主，图中依次排序为 A、B、C,如果竞价广告最多展示两条，则只展现广告主 A 和 B 的广告，如果用户点击了 B 的广告，则搜索引擎公司对 B 收费 1.4 元。

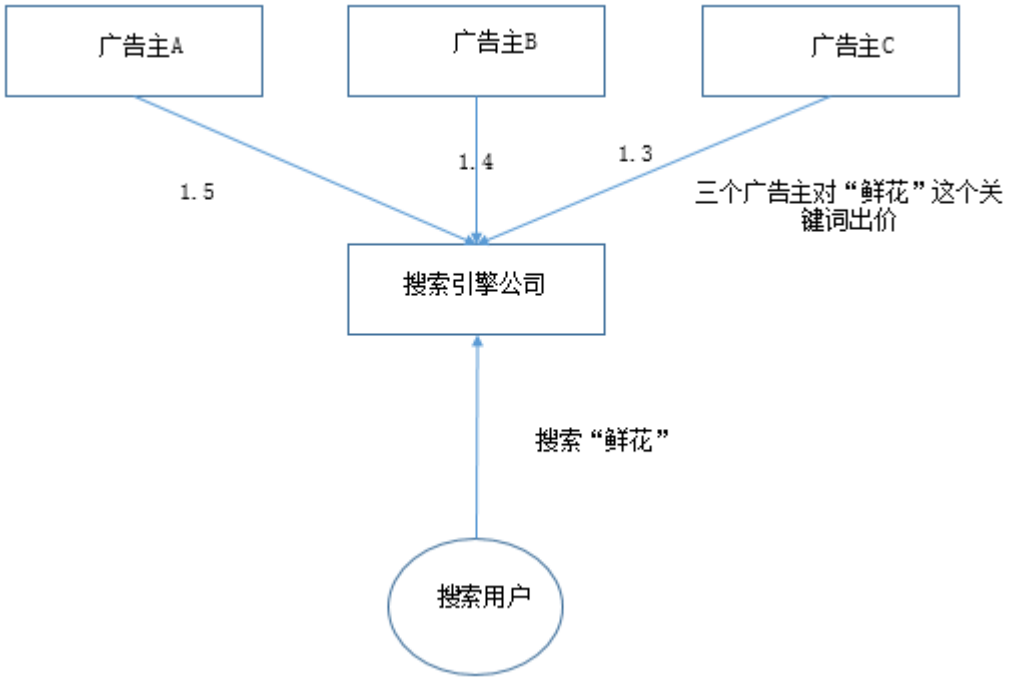

图 2.1 竞价广告简单示意图

搜索竞价广告具有以下四点优势：

1.精准投放。竞价广告只会呈现在搜索了相关关键词的搜索结果页，因此广告投

放完全是精确匹配，直接针对有需求的客户。

2.可跟踪可衡量的广告效果。搜索引擎可以提供广告的数据资料，由此生成完整的报告，方便掌握广告投放效果，及时调整相应的营销战略。例如在某一段时间在上海“可口可乐”一词的搜索量猛增，那么可口可乐的供应商就会根据数据相应地在上海提前备货。

3.广泛的受众。前文已经提到，中国的搜索引擎用户量已经达到5.07亿，覆盖网民人口的80%，覆盖中国人口的30%以上，在如此庞大的受众基数面前，任何品牌的推广都可以达到广告主心中的预期。

4.更有效的收费方式。传统的电视广告中，广告主只要投放就需要付费。而竞价广告则不同，竞价广告只展现不被点击是不会计费的。在关键字被搜索到被点击的过程中，竞价广告进一步地筛选掉非潜在客户，只对点击收费成功地使广告主实现了“好钢用在刀刃上”，避免了无效广告投放。

西北大学的营销大师舒尔茨认为传统的4P营销理论应该被新的SIVA理念代替，即“解决方案(solutions)、信息(information)、价值(value)和途径(access)”，营销人员不再主导一切，权力移交转移到消费者手上，客户或潜在客户成了发送信息的人，而不是索取信息的人，组织变成了接收者与呼应者[15]。SIVA模型的重点在于以消费者为核心，以搜索引擎的广泛使用为驱动力，品牌扮演的角色是为消费者找到答案。当SIVA理论与搜索平台结合，便能为消费者提供实时的解决方案。信息在不断更新，以消费者希望的方式出现，同时消费者还可以参与进来，去评估、修改问题，以至重新搜索。简言之，SIVA理论可以在搜索平台上得到完整体现。通过消费者自主的寻找发现，搜索引擎广告悄无声息地完成了真正最大规模的植入广告，当互联网用户搜索关键词时，搜索引擎给用户呈现的恰恰是与关键词相关的品牌信息，呈现品牌与关键词高度契合，所以通过该关键词进行搜索引擎广告的品牌自然会对搜索用户产生正向激励，使该用户在SIVA模型中的价值和途径环境更倾向于该品牌。

2.1.2 搜索开放平台广告

搜索开放平台广告是搜索引擎广告中的一种形式，相比于传统的竞价广告，搜索开放平台广告则是一种优化与完善。前文已经提到，竞价广告在某种程度上可以被称为植入广告，但是竞价广告因为表现形式单一，只能以文字条目较高排名的方式呈现，所以并不能很好的体现植入广告的精髓。但搜索开放平台广告则不同，搜索开放平台广告具备了更智能式的表现，搜索开放平台广告的展现结果被自然排列在正常搜索结果之中，以文字、图片、flash、视频等形式呈现，更有吸引力而商业广告气息则不是那么浓厚。受众可以直接阅读广告背后内容，还可以点击该广告进入包含广告内容的链接网站。人性化设计更方便，受众更直接，更有价值。

和传统的搜索引擎广告相比，搜索开放平台广告将数据信息前置，成功地使大量的“暗网”即搜索引擎检索不到的信息被发现，使搜索用户获取某些信息的路径变短，方式更简单。基于上述特点，搜索开放平台广告结果已经逐渐受到各大站长、数据拥有者的青睐，笔者大胆推测在未来搜索引擎搜索结果首页将全部由搜索开放平台呈现结果替代，搜索开放平台展现方式将引领搜索引擎的未来。

2.2　预测算法理论基础

预测就是指在掌握现有信息的基础上，依照一定的方法和规律对未来的事情进行测算，以预先了解事情发展的过程与结果。如果根据预测的性质进行分类，则预测可以被分为定性预测、定量预测、综合预测等。本文的主题对点击率进行预测则属于定量预测范畴。在进行定量预测时，需在数据齐备准确无误的前提下，运用传统的统计方法或者其他数学模型，根据已知推测未知的量。

2.2.1　传统统计学预测方法

传统的统计学预测方法可以被粗略分为回归分析和时间序列分析两类：

1.回归。回归模型是揭示变量之间的关系一种最常见的工具。回归的本质是假定因变量与自变量之间存在含有待定系数的特定函数关系，当数据量足够多时，可以根据已知数据对未知的待定系数进行求解，在求解得出因变量与自变量函数关系的后，可以将已知的自变量代入到求解得出的函数式，求出需要预测的因变量。根据函数表达式的线性与否，回归被分为线性回归、非线性回归。由于函数关系的简单直接，可解释性强，因此成为传统统计学预测方法中最普遍的预测方法。

2.时间序列分析。时间序列是按时间顺序的一组数字序列。时间序列分析就是利用这组数列，应用数理统计方法加以处理，以预测未知的未来状态。基于事物发展具有一定规律性的假定，时序分析将这种规律性分为长期上升、下降或维持水平的趋势因素，以相同长度的时间序列反复出现的周期因素，趋势和周期间隔出现从而导致的季节因素以及完全不可掌握的不规则波动。对时间序列分析存在多种方法，针对时间序列中可能存在的因素，学者们针对不同的因素组合创造了不同的方法来有效对时间序列进行分析，如指数平滑法、Box-Jenkins 法以及对时间序列适用性最强的 ARIMA 方法等。

2.2.2　几种常见的数学模型预测方法

数学模型的数值预测方法较传统统计方法更加复杂，但却更具有适用性。常见的几种方法数学模型预测方法有：神经网络、随机森林以及本文中将要使用的贝叶斯网

络。

1.神经网络模型。该模型是对生物神经系统的模拟，因生物的信息处理功能极其强大，所以神经网络模型模拟了神经系统中的神经元的激活特性、神经元的传输方式、突触联系的强度来构造网络单元的输入输出特性、网络的拓扑结构、连接权的大小以及神经元的阈值。通过确定神经网络类型、神经网络的结构，神经网络可以较为智能对数据进行预测。在众多预测系统中，神经网络模型得到了广泛应用[16]。

2.随机森林模型。随机森林的名称很形象，随机的构建许多没有关联的决策树，“树木”积少成多形成了“随机森林”。因为决策树之间不相关，每一个决策树得到的结果都独立于其他决策树，因此随机森林整体的判断结果取决于多个“决策树专家”的投票结果。因为决策树的数量较大，最终的结果有效地避免了单独某个决策树的“偏见”对最终结果的影响。

2.3 贝叶斯网络理论基础

2.3.1 贝叶斯网络的基本概念

贝叶斯网络（Bayesian network），又称信念网络（belief network）或是有向无环图模型（directed acyclic graphical model），是一种基于贝叶斯公式的概率图形模型[17-22]。贝叶斯网络是对朴素贝叶斯的一种升华，因为朴素贝叶斯有一个限制条件，就是特征属性必须有条件独立或者基本独立，当这个条件成立时，朴素贝叶斯方法的准确率极高。但事实上，现实应用中几乎不可能做到完全独立，哲学上讲物质都是相互联系的，而现实中各个特征属性之间往往会出现较强的相关性，这也就限制了朴素贝叶斯的能力。而贝叶斯网络则放宽了变量无关的假设，将贝叶斯原理和图论结合，建立起一种基于概率推理的数学模型，对于解决复杂的不确定性和关联性问题有很强的优势。

前面提到贝叶斯网络的另一个称呼为有向无环图模型，因此贝叶斯网络模型必然要包含一个有向无环图，其中每一个节点代表一个随机变量，而有向边（在图论中也被称为弧）则表示两个随机变量之间的联系，表示指向结点影响被指向结点[23-24]。不过如果仅有一个图的话，只能定性给出随机变量间的关系，如果要定量，还需要一定的数据，即每个节点对其直接前驱节点的条件概率，而没有前驱节点的节点则使用先验概率表示。因此，贝叶斯网络模型需要由有向无环图和条件概率表两部分组成。下图2.2是一个简易的贝叶斯网络模型图，每一个节点就是一个圆，代表了一个事件或者状态，而弧则代表了它们的条件概率。

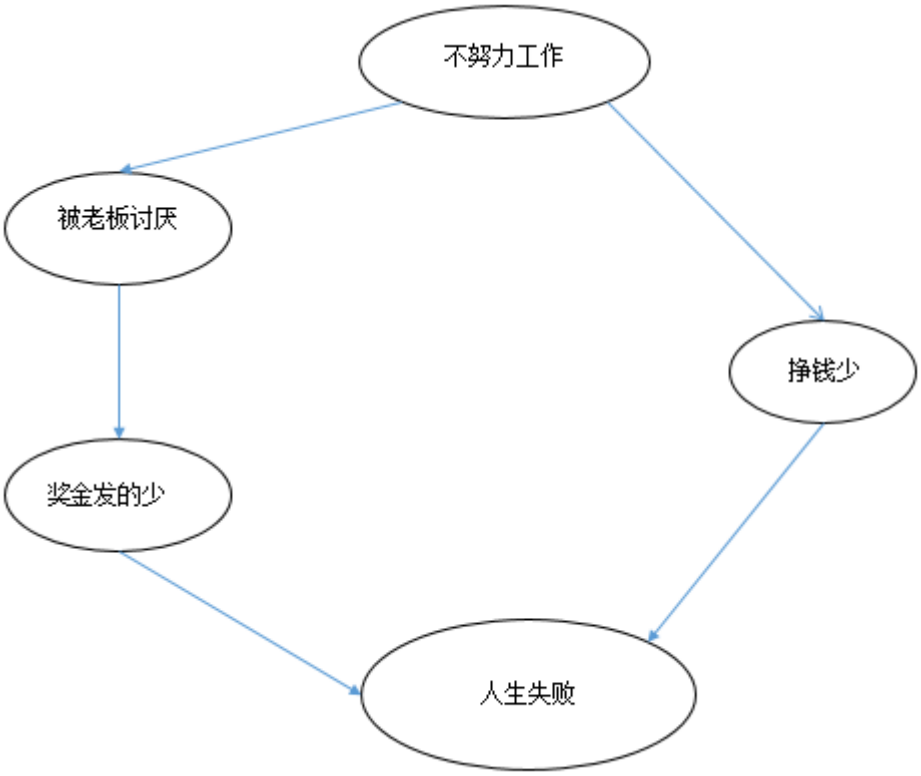

图 2.2 贝叶斯网络示意图

2.3.2　贝叶斯网络的特征与性质

贝叶斯网络定义由有向无环图（DAG）表示的结构模型以及对应的条件概率表组成，其中有向无环图中的不同节点表示模型中的不同变量，节点相连的弧表示变量之间的因果关系。而条件概率表则记录了针对不同父节点的状态每个节点的条件概率[25]。如果用 D 代表一个由 N 个节点组成的有向无环图，这 N 个节点可以分别与集合 W={w1，w2,…，wn}中的元素一一对应，η代表节点之间的条件概率分布，则由结构模型 D 和参数模型η构成的贝叶斯网络模型 B，可以表示为 B=<D, η>。在该模型的结构模型中只需满足条件独立性的假设，则网络中所有的节点概率值都可以通过相关节点的值和先验数据计算出来，即任意变量的联合概率都是可以被计算出的[26-27]。其中，条件独立性的假设表述为网络中每个节点都条件独立于非该节点后代节点构成的任意节点子集，如用 R（wi）代表 wi 的直接双亲节点，N（wi）代表非 wi 后代的节点子集，则条件独立性的公式表述如下：

$$P(w_i \mid R(w_i), N(w_i)) = P(w_i \mid N(w_i)) \tag{2.1}$$

以上述集合 W 为例，W 上的联合概率分布构成了贝叶斯网络。根据条件概率的链式法则，则有变量 vi(i=1,2,…,n)的联合概率分布必定满足以下公式：

$$P(w_1, w_2, \cdots, w_m) = \prod_{i=1}^{n} p(w_i \mid N(w_i)) \tag{2.2}$$

综上，一个完整的贝叶斯网络具有以下几点特征：

1.贝叶斯网络必然有一个结构模型。首先，需要有一个网络节点集合，每个节点都代表了状态、过程、进展等某个随机变量。其次，需要有一个弧的集合，弧具备了方向，方向代表了被指向节点是指向节点的子节点。

2.父节点对子节点的影响均可通过条件概率分布 P(wi|N（wi）)进行量化，其中 wi 是任意节点，N（wi）是 wi 的非 wi 后代的节点集合。

3.贝叶斯网络又名有向无环图模型，所以不能存在环结构。

除了上述特征，贝叶斯网络有一条极为重要的性质：每一个节点仅与父节点相关，跟所有不直接关联的前驱节点条件独立。

这个性质很类似 Markov 过程，因此也被称为贝叶斯网络的马尔科夫毯（Markov blanket)。在某种程度上，对 Markov 链做非线性扩展就可以得到贝叶斯网络。这条性质极大化地便利了贝叶斯网络中求取联合概率分布的方法。通常，我们通过以下公式计算联合条件概率分布：

$$P(x_1, x_2, ..., x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)...P(x_n|x_1, x_2, ..., x_{n-1})$$

由于马尔科夫毯性质，上述公式被化简成：

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i|Parents(x_i)) \tag{2.4}$$

其中 Parents（xi）是 xi 父节点的集合。

2.3.3 贝叶斯网络的优点

和关联规则、随机森林、神经网络相比，贝叶斯网络参考了图论中的基础，使其无论在计算还是在概念理解上都较其他方法有着一定的优势：

1.解释简单。这一优点与其模型的可视化有着极大的关联，神经网络、模拟退火等算法或多或少参考了某一少见的知识领域的知识来解决问题，这也就使得常人在理解这些模型时存在疑问。而贝叶斯网络不同，它的解释更加直接明了，变量之间的相关性只需检查两节点是否通过弧相连即可。

2.计算时间与空间的节省。前文提到，贝叶斯网络可以看做是马尔科夫链的非线性扩展。因为某些节点只与父节点相关，因此与常规的排列组合方式多个节点可能性相乘相比，贝叶斯网络节点变为可能性相加，极大地减少了计算的时间与空间。

3.判断迅速。对人类而言，它更能轻易地得知各变量间是否条件独立或相依与其局部分配（local distribution）的型态来求得所有随机变量之联合分配。

4.适用性强。贝叶斯网络是模拟人的认知思维推理模式，用一组条件概率函数以及有向无环图对不确定性的因果推理关系建模，而基础概率知识的可以解释大多问题，因此贝叶斯网络对不确定性问题具有很强的适用性。

5.先验与后验的有效结合。和大多数机器学习的方法相比，贝叶斯网络有个极其明显的优势（或许也有可能是劣势），它不会完全受到数据的影响。在由专家先确定结构的前提下进行参数学习使得模型不会因极端数据而产生完全无法经验解释的模型（劣势的原因就是数据挖掘往往需要挖掘出人类未知的一些知识，一旦模型由专家参与构建，那么发现未知自然变得不可能）

2.3.4　贝叶斯网络的构造与学习

贝叶斯网络作为一种不确定性的因果推理模型，其应用范围非常广，在医疗诊断、信息检索、电子技术与工业工程等诸多方面发挥重要作用，而与其相关的一些问题也是近来的热点研究课题。例如，Google 就在诸多服务中使用了贝叶斯网络。

具体地讲，在确立模型参数后建立贝叶斯网络分为以下两步：

1.明确模型参数之间的因果关系，使自变量变量通过弧指向因变量，形成 DAG。通常这一步可以通过算法求解（如贪婪搜索算法、爬山算法等）或者由领域专家指定，如本文中所用的结构都来自于多位领域多年从业人员的经验综合构成。当然，一次构造完 DAG 模型的搭建并不一定会完结，通常想要建立一个好的拓扑结构，需要不断地构建模型验证然后迭代改进。

2.训练贝叶斯网络，即参数学习，这一步也就是要完成条件概率表的构造。从常识中我们知道，对于某种因变量，人类可以获取到的自变量往往并不完全，或者说也不可能实现真正的完全，因此我们把数据集分为两类：完备数据集和残缺数据集。对于一些残缺数据集，即在贝叶斯网络节点中存在的某些隐藏变量节点，训练方法就是比较复杂，例如使用梯度下降法。对于完备数据集，贝叶斯估计和最大似然估计都是比较好用的参数学习方法。两者之间的共同点都在于寻找最大值，不同的是贝叶斯估计法以先验概率为基础，以网络结构和已知数据集求解未知后验概率最大值，而最大似然估计法则以似然函数为基础，计算不同节点的概率来找出使似然函数取最大值时的函数参数值[34-45]。

以下为贝叶斯网络应用的一个示例，如果 SNS 社区中需要进行不真实账号检测，我们的模型中存在四个随机变量：账号真实性 R，头像真实性 H，日志密度 L，好友密度 F。其中 H，L，F 是可以观察到的值，而我们最关心的 R 是无法直接观察的。结构模型图如下图 2.2 所示。

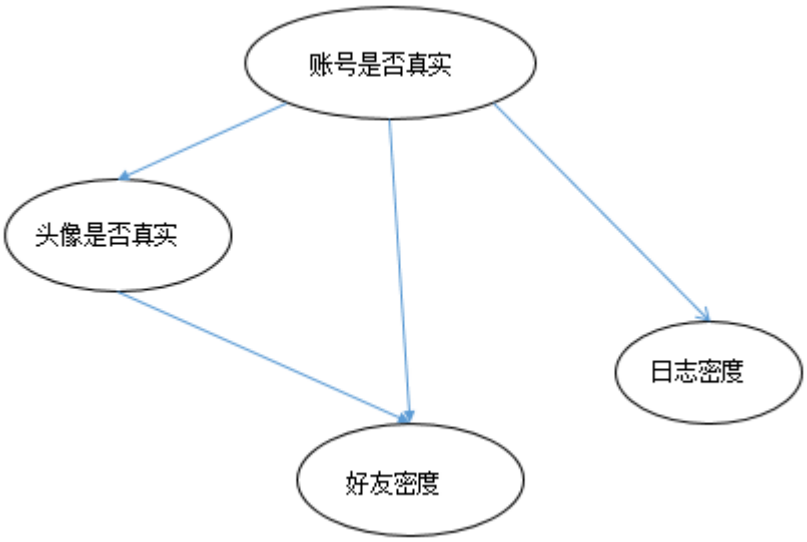

图 2.2 社交网络真实性判断结构模型

于是，这个问题就划归为通过 H，L，F 的观察值对 R 进行概率推理。推理过程可以如下表示：

1. 使用观察值实例化 H,L 和 F，把随机值赋给 R。

2. 计算 P(R|H,L,F)=P(H|R)P(L|R)P(F|R,H)。其中相应概率值可以查条件概率表。

2.4 本章总结

为了支持第三章节中的理论研究，本章先对搜索引擎广告，特别是本文的研究重点搜索开放平台广告进行了介绍；其次简单介绍了常见的预测算法；最后着重介绍了贝叶斯网络，对贝叶斯网络的基础性质、优点以及构造方法进行了细致描述与分析。本章一方面是为下一章的理论应用进行铺垫，另一方面也阐述了贝叶斯网络在本文中的应用选择的合理性与可行性。

第三章　基于贝叶斯网络的预测模型

本章将根据上一章讨论的贝叶斯网络理论对搜索开放平台广告建立预测模型。首先，通过笔者的实际经验选择一定的客观变量，通过传统统计学方法验证这些变量与预测目标点击率之间的相关性，使这些相关变量组成贝叶斯网络的有向无环图。其次，通过收集到的搜索开放平台广告上线前一天的相关数据以及上线一天的点击率进行参数学习，确定模型中的参数。完成模型后，通过模型对搜索开放平台广告点击率的预测验证模型的有效性。最后，在前面研究的模型基础上引入主观评价（打分）参数，验证搜索用户对不同的搜索开放平台结果的打分评价确实会对点击率产生影响，且引入该参数可以使模型的点击率预测变得更加准确。

3.1　模型数据说明

前文中已经讨论过，搜索开放平台广告是搜索引擎广告中的一种优化，因此对搜索开放平台广告关注的效果有一部分与竞价广告相同，如搜索展现量、点击量、展现排序位置。不同的是，根据网站站长或者广告主提交数据的不同，搜索开放平台广告的触发词（关键词）是不同的，因搜索开放平台广告内可能含有成百上千条数据，所以触发词也可能是成百上千项，针对不同的触发词展示该广告内不同的数据，只是展现形式都会相同。此外，搜索开放平台广告较竞价广告最大的优势就是展现效果更好，即有可能存在图片或者互动输入项。此外，以笔者实际操作经验，即使是相同的搜索展现结果，当展现结果的标题使用不同的同类网站名称时，点击率会出现明显变化，这种现象可以解释为搜索用户对不同的网站信誉度有差别，因此笔者引入数据来源网站排名（webrank）在模型中对这一现象进行解释。除了上述这些客观存在的参数外，还有一个主观参数——美观性，亦或者是用户体验。用户体验，这个词是互联网行业或者所有创造行业的一个非常重要的衡量，它是用户在使用产品或者体验某种服务时纯主观建立起来的感受，用户体验的好坏决定了产品或者服务的成功与失败[28]。以下为在模型中出现的数据的具体说明：

1.Rank(搜索开放平台广告排序位置)。搜索开放平台广告区域在竞价广告位置之下，具体位置参见第一章图 1.1 的 D 区域。尽管目前为止，搜索开放平台广告的 rank 都是按照搜索引擎的内部逻辑变换位置，但 rank 仍是一个可控变量，即搜索引擎公司可以决定一个搜索开放平台广告在上线第一天的排序位置。而且笔者认为 rank 位如其在竞价广告中的地位一样，对点击率仍然有着不可忽视的影响。在本文中使用的 rank 源数据为搜索开放平台广告上线第一天中按其在 24 小时内占据不同位置的时间长短取

的均值。

2. Pic(图片量)。搜索开放平台广告的展现形式之所以称得上丰富就是因为搜索开放平台广告可以增加多图展示，甚至可以增加动态图片或者 FLASH 展现。图片一方面占据了更广的位置使其在显示屏幕中更加显著，另一方面也使得广告更加简单易懂。在不同的搜索开放平台广告中，使用的图片数量与图片质量都是不同的，因图片质量不便客观衡量，因此不作为本模型的参考依据，仅使用 Pic 来对搜索开放平台中的图片数量进行描述。通过确认图片项对搜索开放平台广告点击率的影响，搜索开放平台广告展现形式在设计之初就可以把握一些规律，预先把设计做到最优水平。

3. Input(输入项数量)。搜索开放平台广告将暗网数据前置，使搜索用户浏览路径变短。在某些搜索中，搜索开放平台广告会增加输入项窗口，使得用户可以直接调取到广告主数据或者向广告主提交数据。笔者认为随着输入项数量的增多，使用者的懒惰将无意愿对广告进行点击，因此会使点击率下降，因此将输入项数量引入模型。下图 3.1 为带有输入项的搜索开放平台广告示例。

图 3.1 带输入项的搜索开放平台广告示例

4. Keywords(关键词数量)。关键词也被称为触发词，即搜索用户在搜索这些词时广告就会在搜索结果中展现。和竞价广告的一个广告只可以对应一个触发词不同，搜索开放平台广告往往对应着多个触发词。搜索开放平台广告的触发词往往是以组合词的方式形成的，以申办信用卡的搜索开放平台广告为例，信用卡网站可以提供的是不同银行的信用卡，如银行名称为 A，“信用卡”三个字为 B，则 AB 组合或者 BA 组合都代表了搜索用户对某行信用卡感兴趣，即可触发该广告，而广告主的数据覆盖了多家银行，在触发词上搜索开放平台广告只需更新词表 A 即可。此外，如“申办”、“申请”这些动词被称为可有可无词，无论有或者没有其实都代表了搜索用户的信用卡的意愿，所以被称为可有可无，它可以被放在 AB 之前，亦可在之间，亦可在之后。所以，对搜

索开放平台广告的触发词数量往往是一个排列组合，其覆盖范围之广，数量之大，较搜索竞价广告有着明显的区别。

5.PV(Page View 的缩写，也被称为 Impression，即展现量)。广告的每一次触发都会被搜索用户看到一次，毫无疑问，PV 增加点击才会增加，因此 PV 对点击有着极大的影响。在绝对数值上，关注点击量就必须要关注 PV 数值。本文的目的是在搜索开放平台广告上线前预测点击率，在触发词量已知的情况下，触发词量与展现量存在着较强的因果联系。举极端假设，如果触发词无穷大，覆盖了所有用户搜索词，即搜索用户无论搜索什么词都会触发广告，那么展现量可能将以百亿为单位。展现量往往是点击数量的上限，一般对一个搜索结果，搜索用户只可能点击一次，当然也不排除某些结果含有较丰富的信息使某些人点击多次的情况，但是对于搜索用户整体，对一个搜索结果的日均点击量普遍地会小于等于展现量。

6.Click（点击量)。点击量，顾名思义，广告被点击的次数。点击量直接影响搜索开放平台广告的效果，对广告主来说每一次点击就意味着一个流量来源，也就是一个潜在价值客户。

7.CTR（Click Through Rate，点击率)。本文的研究目标，它是一个相对值，CTR=Click/PV。本文的研究就是希望通过预测 CTR 这个相对概念来确定在 PV 已知的情况下点击量的多少，进而也就确定了项目的商业价值。若 CTR 很高，而 PV 很低，那么增加触发词将是点击量提升的有效手段，如果无法增加触发词，那么项目的价值不够很有可能就会被放弃。如 CTR 较低，那么对于从项目设计之初搜索引擎公司的设计人员就应在设计阶段想方设法提高 CTR，避免后期上线商业价值低而带来的困境。

7.Button（按钮数量)。点击的源自搜索用户对按钮的点击，因此按钮数量越多，意味着每一次广告展现就越有可能被用户多次点击，这也是某些搜索引擎广告点击率可能超过 100%的原因之一。

8.Webrank(数据来源网站排名)。相同的展现结果带有不同网站的标题对点击率会有极大的影响，这就是数据来源网站的名气对搜索用户的信任度影响。可以肯定的是，越是知名的网站，搜索用户信任度越高，其对点击率的影响也是正向的。在没有其他更有效的衡量方式前提下，笔者使用了 Alexa 排名来衡量，Alexa 排名记录了全球几十亿个网站的访问量和浏览量，在覆盖广度上具有很大的优势。通过 Alexa 对每个网站的访问量排名来间接衡量搜索用户对该网站的信任度，笔者认为是可行的。

9.UE（User Experience,用户体验)。前文提到的八个变量都是客观变量，不以某个人的意志为转移。但是 UE 却是一个完全主观且对搜索开放平台广告的点击率可能影响比较重要的一个变量。简单地讲，用户体验就是一个用户在使用了一个产品或者服务后的感受，用户体验好不单意味着该用户可能再次或者多次使用该产品，留住顾客，更意味着该产品或服务可能通过口碑营销的方式得到广泛传播。更进一步地，用户体

验甚至都不需要用户完全经历产品或者服务的整个流程，如饭店做的菜很难看顾客很可能就不会去吃而是直接退单。同样的，本文中对搜索开放平台广告的用户体验指标也不需要体会完整的服务流程，笔者请人对诸多搜索开放平台广告展现结果样式进行了打分，最低分 1 分，最高分 5 分。之所以只对展现结果样式打分，而不考虑其他一些因素,主要是因为搜索开放平台广告在上线前我们最有可能获取的就是它的 demo 图，而其他诸如进入网站的速度、跳转页面的布局这些我们是无法提前获取的。

3.2 模型数据统计分析

3.2.1 数据来源及数据处理

本文的目的是尽可能早地在搜索开放平台广告上线前实现对其点击率的预测，因此搜集的数据除了最终的预测目标点击率 CTR 是在搜索开放平台广告上线后 24 小时的统计数据，rank、pic、input、webrank、keywords 均为上线前可已知数据。其中 rank 数据虽然只有在上线后方可统计，但是因为可人工控制所以仍作为上线前已知数据，rank 数据也是搜索开放平台广告上线 24 小时的一个均值；webrank 作为 Alexa 流量排名，代表了网民的信任度，但是某一天的流量排名并不会直接对搜索用户产生影响，因此使用的排名为 Alexa 流量排名在三个月的均值；除 pic、input 外其它数据都采用的是广告上线前一周的日均值。

本文在实证研究过程中，收集了某搜索引擎公司从 2014 年 1 月 1 日起至 2014 年 12 月 31 日截止共 9921 条有效的搜索开放平台广告的历史数据。事实上，在该时间段内上线的广告数量远不止于此，但受限于笔者收集数据的时间与精力有限，且有些数据无法统计齐全，所以在数据集数量上只能止步于此。

数据处理是指对收集到的数据进行加工整理，形成适合数据分析的样式，它是数据分析前必不可少的阶段。数据处理的基本目的是从大量的、杂乱无章、难以理解的数据中，抽取并推导出对解决问题有价值、有意义的数据[29]。从数据挖掘的角度看，数据的处理占据了数据挖掘人员一半以上的工作量。本文中的数据很大程度上来自于笔者人工收集，在收集的过程中已经剔除了无效数据、空缺数据，也去掉了一些异常数据项，如点击率极高或者搜索量极大的数据，这些数据占总数据比例较小，但因为数值极其特殊会对模型产生很大的偏差影响，所以均被作为异常数据删除了。

在数据处理过程中，笔者认为数以万计的网站排名 webrank 单独的几位波动并不会对搜索用户信任度产生较大影响，因此将网站排名分为了 11 个等级，其中数值越小代表了网站排名越靠前，具体参见下表 3.1。

表 3.1 Webrank 排名分类等级处理

Alexa 排名	Webrank 等级
1-10	1
11-50	2
51-100	3
101-500	4
501-1000	5
1001-2000	6
2001-4000	7
4001-8000	8
8001-20000	9
20001-50000	10
50000 以上	11

基于同样的原因，笔者对关键词数量与展现量做了分级处理，毕竟在数以千记甚至万计的关键词数量或者展现量面前，考虑到系统误差的存在，其实精准到个位的预测并不会对我们估计产品的商业价值有太大的影响。此外，将数据处理为离散变量存在几个明显好处：（1）在贝叶斯网中对参数可以设定为离散值或者连续值，如需要设定为连续值，则参数需要符合高斯分布的假设，设定为离散值可以避免约束性过强的情况。（2）设定为离散值后，如参数设定为连续值，模型对计算机的内存要求（数据以矩阵方式存在，占用内存以平方形式增长）会变得更高，且运算速度也无法得到保证。坦率地讲，对参数离散化也是对准确度与机器学习时间成本的一个平衡。综上，对关键词数量与展现数量以 10 倍关系进行分级处理，见下表 3.2。

表 3.2 Keywods 与 PV 分级

实际数值	分级等级
1-10	1
11-100	2
101-1000	3
1001-10000	4
10001-100000	5
100001-1000000	6
100000+	7

对参数 Rank，出于搜索引擎公司对 rank 可控的原因，笔者将其引入模型作为自变量。尽管源数据是对 rank 在一天中的时间权重均值，可以为小数，但是现实生活中却并不存在第几点几位的描述方法，因此对 Rank 进行了向下取整的处理。

对于预测目标 CTR，笔者以 1%为精度进行分级。相较于搜索引擎公司动辄 10%的描述估计方法，1%的偏差已经非常精准。对目标参数分级的主要目的同样是为了减少运

算时间，如有更精准的需求，只需对目标参数进行更细致的划分即可。

3.2.2 模型变量的描述统计与定性分析

本文模型中涉及到的变量共 9 个，下表 3.3 通过最大值、最小值、中间值、均值、标准差五个基本统计量对已经处理后的数据进行了统计描述。

表 3.3 模型变量的总体描述统计

变量名称	变量描述	最大值	最小值	中间值	均值	标准差
rank	广告搜索结果排名	12	1	1	2.38	2.40
pic	广告图片数量	26	1	2	4.40	5.10
input	广告输入项数量	5	1	1	1.69	1.09
pv	广告展现量	7	1	4	4.33	0.80
keywords	关键词数量	6	1	4	4.31	0.74
button	广告按钮数量	63	1	9	10.68	7.79
webrank	数据来源网站流量排名	11	1	8	6.93	2.50
UE	用户体验	5	1	2	2.87	1.45
CTR	点击率	82	1	8	9.97	6.66

在上表中对 CTR 的描述统计表示，CTR 的最小等级为 1，即小于 2%，该值与 CTR 最大等级对应值 82%相差百分比绝对值 80%，相差之大，效果差异之明显可想而知。若对搜索开放平台广告的效果没有一个准确预估，放任点击率出现在最大与最小值之间的任何一个区间，那么对企业的商业生产的破坏是剧大的。正是出于解决上述问题的目的，搜索引擎公司迫切地需要对点击率进行预估。下面开始对各变量进行定性分析。

1.搜索开放平台广告排序位置(Rank)与点击率之间的散点图如图 3.2 所示。Rank 最大值为 12，而自然搜索结果首页展现结果最多在 10-12 条不等（其上的竞价广告条数波动变化），可以说 Rank 最大值时已经处于首页最底部。而在散点图中，我们可以明确的看到 CTR 与 Rank 存在负相关，Rank 越大，搜索开放平台广告位置越靠下，CTR 也就越低。

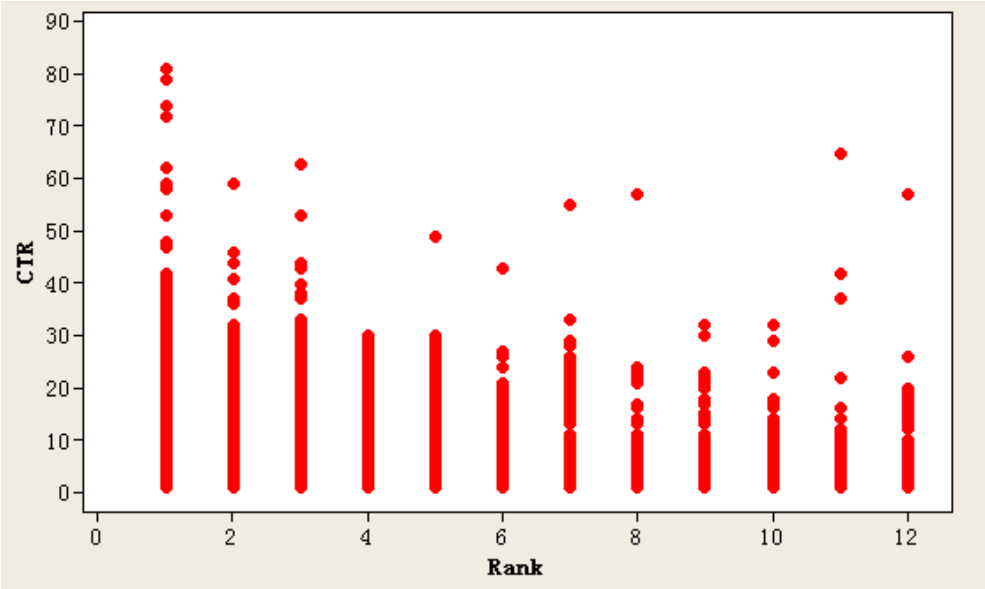

图 3.2 CTR-rank 散点图

除了从上图我们可以看到 CTR 与 Rank 之间的关系，我们还可以看出在整体的样本数据中，大部分数据密集地集中在 Rank 为 1 之上,统计 Rank 为 1 的样本量，发现其占总样本数量的 53%。如果上图表明了 Rank 位与 CTR 之间的负相关关系，那么大量样本 Rank 值为 1 也就证明了搜索开放平台广告的点击率确实比正常的搜索结果更优秀。

2．搜索开放平台广告展现结果的图片量（Pic）与点击率之间的散点图如图 3.3 所示。

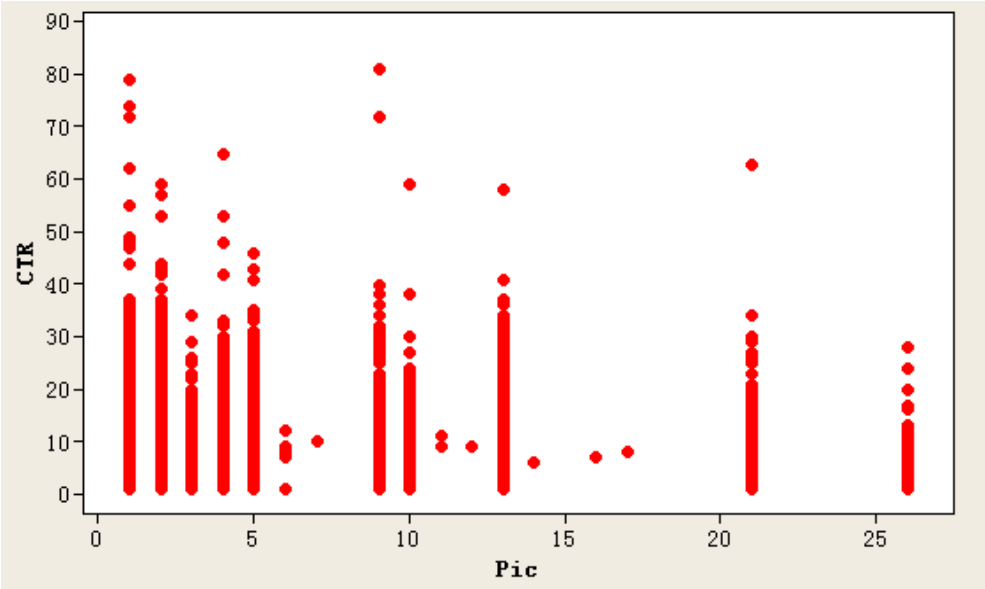

图 3.3 CTR-pic 散点图

尽管从逻辑上判断认为图片的存在会增加点击率，但是从散点图上分析，搜索开

放平台广告展现图片数量与点击率之间并无明显线性相关性。如果把偏离主体的某些异常点考虑进去，散点图整体有下行趋势，即图片数量增多会使点击率下降。为了验证 CTR 与 pic 之间的相关关系，本文对这两个变量做了皮尔斯相关系数检验，其结果如表 3.4 所示。根据下表显示，显著性=0.000,小于 0.01，拒绝原假设 pic 与 CTR 不相关。且因为皮尔森相关系数为负，搜索开放平台广告点击率与图片数量呈现负相关，验证了从散点图中观测到的图片越多点击率越低的结论。这一现象可能与搜索引擎用户对广告的选择性忽略有关，当图片过多时，搜索开放平台广告的展现结果太过商业化，从而使得某些希望通过搜索实现知识发现的用户忽视该结果。

表 3.4 CTR-pic 皮尔森相关系数检验

		pic	CTR
pic	皮尔森 (Pearson) 相关	1	-.178**
	显著性 （双尾）		.000
	N	7461	7461
CTR	皮尔森 (Pearson) 相关	-.178**	1
	显著性 （双尾）	.000	
	N	9921	9921

**. 相关性在 0.01 层上显著（双尾）。

3.搜索开放平台广告输入项数量(input)与点击率之间的散点图如图 3.4 所示。

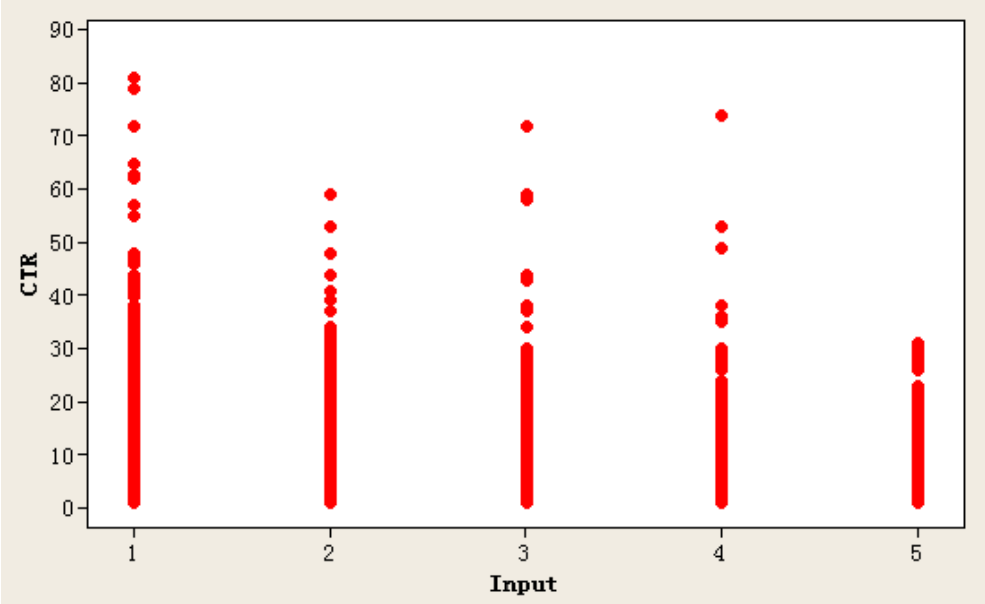

图 3.4 CTR-input 散点图

搜索开放平台广告将暗网数据前置，使搜索用户浏览路径变短。在某些搜索中，搜索开放平台广告会增加输入项窗口，使得用户可以直接调取到广告主数据或者向广告主提交数据。尽管带有输入项的搜索开放平台广告数量不多，且散点因为输入项数

目并不连续，从散点中仍可看出输入项的增多会使点击率下降，即因为在点击前搜索用户需要进行输入，使点击的时间成本增高，减少了某些因好奇或者被广告吸引的使用者对搜索开放平台广告的点击。因此，输入项数量也将被引入预测模型中。

4.搜索开放平台广告展现量（PV）与点击率之间的散点图如图 3.5 所示。

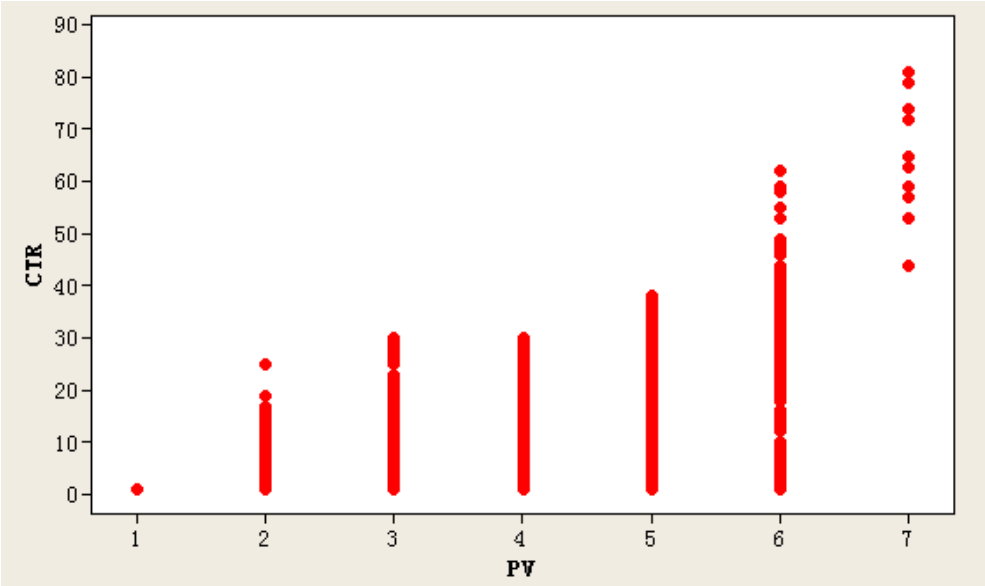

图 3.5 CTR-pv 散点图

从上图我们可以明显看到点击率与展现量存在正相关关系。PV 展现量越大，点击率出现较高值的可能也就越大。从逻辑上讲，某一相对数值并不应因绝对数值的变化而变化，但是从散点图上却明确呈现了该趋势。下表 3.5 进一步验证了散点图中的展现结果。表中数据表明，显著度=0.000，小于 0.001，拒绝 PV 与 CTR 不相关的假设。此外，从皮尔森相关系数=0.723>0 可以得出 PV 和 CTR 呈现正相关趋势，即展现量的增多会增加 CTR。一方面有可能是大展现量的搜索开放平台广告涵盖了大量的数据，更容易满足搜索用户需求，另一方面也可能是该类搜索属于社会密集搜索，即大比例的搜索用户都会关注此类查询，因此也会提升点击率，具体例子如天气查询、汽车摇号查询等，大量的用户都会在网络搜索天气状态、汽车摇号中标情况，因此点击率会有显著的提升。

表 3.5 CTR-pv 皮尔森相关系数检验

		pv	CTR
pv	皮尔森 (Pearson) 相关	1	.723**
	显著性 （双尾）		.000
	N	9921	9921
CTR	皮尔森 (Pearson) 相关	.723**	1
	显著性 （双尾）	.000	
	N	9921	9921

**. 相关性在 0.01 层上显著（双尾）。

为了确认搜索量的增加与CTR呈正相关是因为搜索热度还是因为覆盖数据的广泛，本文进一步对关键词数量与展现量进行了分析，两者的散点图如 3.6 所示。因为 PV 和 Keywords 都是离散值，所以在散点图中看到的只是几个离散稀疏的点，但是从图中仍可以看到，随着关键词的增加展现量也在增多。因此，我们推测 CTR 的增多与主要是数据覆盖更加的原因，即关键词增多使搜索用户的一些长尾需求被覆盖。具体讨论可见关键词数量与点击率之间的关系。

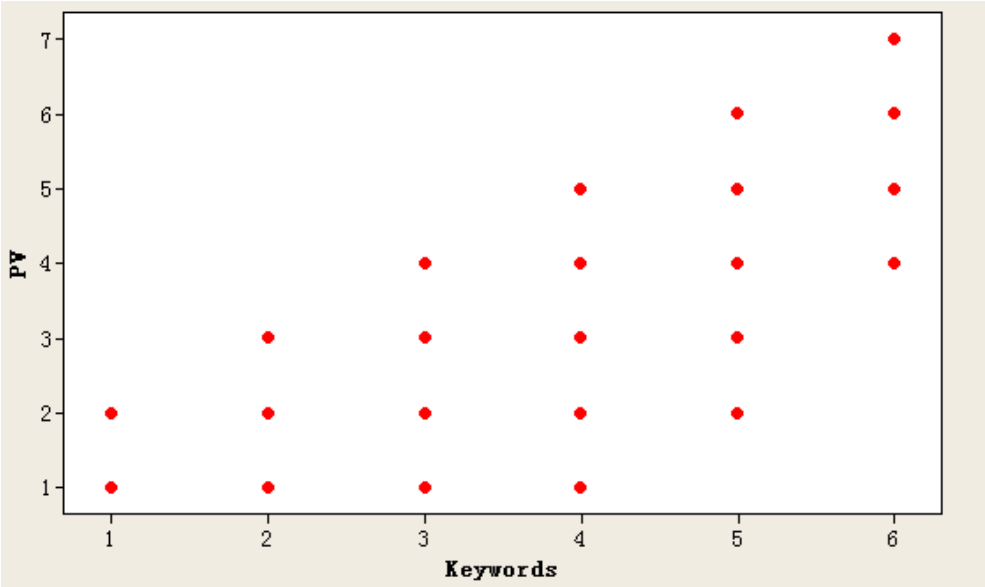

图 3.6 PV-keywords 散点图

5.关键词数量（keywords）与点击率之间的散点图如图 3.7 所示。从图 3.7 我们可以看出，点击率随着关键词的增加而增加。因此我们断定，随着关键词的增多，点击率也会增多。产生这种现象的原因极可能是关键词因为排列组合覆盖到长尾词，搜索开放平台的广告更加容易满足长尾需求。尽管某些词的日均搜索量较小，但在该类词被搜索时触发广告的情景也更加精准，因此更符合搜索用户的搜索需求，所以点击率会增加。该散点图的分布与展现量与点击率的散点图极其相似，但是展现量作为一

个后验参数，只能在搜索开放平台广告上线后才能得出。从这个角度上看，通过研究关键词与展现量的关系以及展现量与点击率之间的关系，间接得出关键词与点击率的关系的方法才是有效地预测方法。

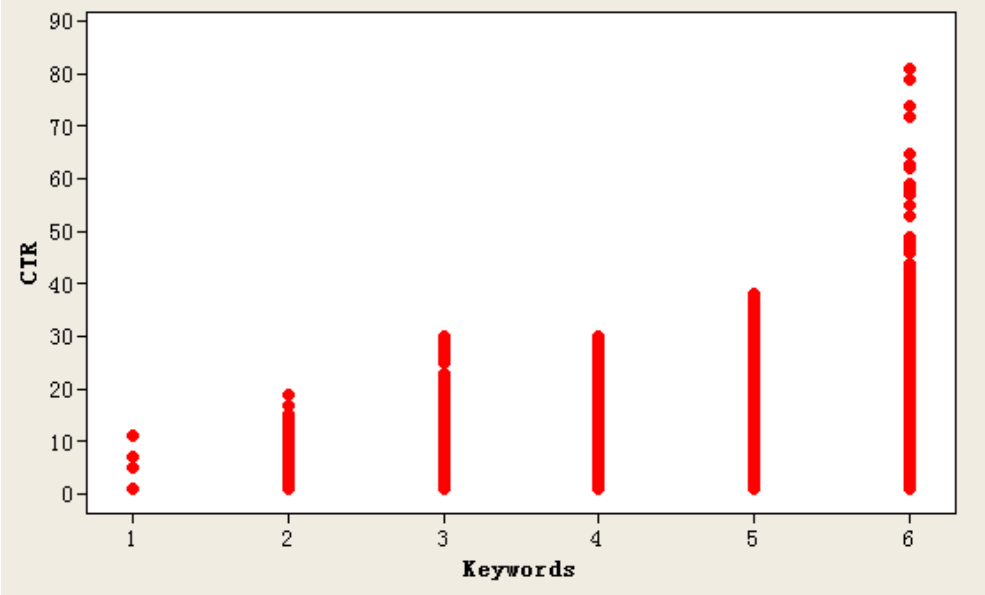

图 3.7 CTR-keywords 散点图

7. 按钮数量（button）与点击率之间的散点图如图 3.8 所示。

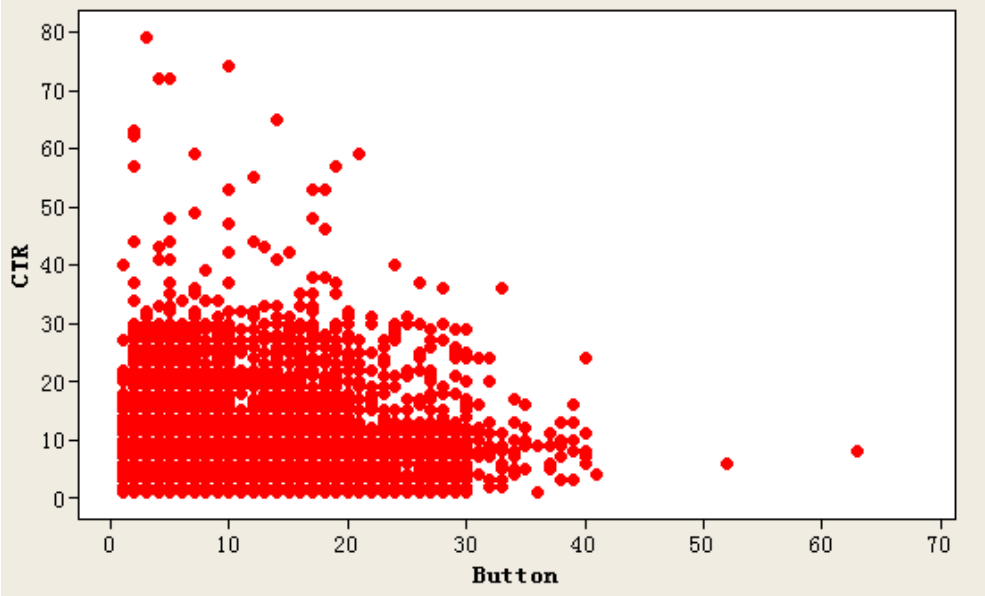

图 3.8 CTR-button 散点图

从图 3.8 散点图上观察，按钮数量与点击率呈现负相关，而这一观察结论其实与经验推测不太吻合。因此，对按钮数量与点击率进行皮尔森相关系数检验，结果如下表。

表 3.6 CTR-button 皮尔森相关系数检验

		Button	CTR
Button	皮尔森 (Pearson) 相关	1	-.702**
	显著性 （双尾）		.000
	N	9221	9221
CTR	皮尔森 (Pearson) 相关	-.702**	1
	显著性 （双尾）	.000	
	N	9221	9221

**. 相关性在 0.01 层上显著（双尾）。

上表数据表明，显著性=0.000，小于 0.001，可以拒绝 CTR 与 button 不相关的原假设。此外，两者间的皮尔森相关系数=-0.702，小于 0，因此随着按钮数量的减少点击率也会增加。这一结果的得出确实出乎笔者的意料，经验推断上看，按钮越多，搜索引擎用户点击广告的“通道”也就越多，点击率应该越大。但数据证明，按钮越多点击率越低，这一现象产生的原因笔者推测有两方面：一方面按钮越多，广告商业性越明显，用户也越容易选择性忽略。另一方面，在有限的展现空间下，放置很多按钮，很有可能影响整体结果的美观，从而使得对搜索用户的吸引力下降。

8. 网站排名（webrank）与点击率之间的散点图如图 3.9 所示。前文提到，为了使网站排名数据更有意义，网站排名是根据 Alexa 排名分级后得出的数据。

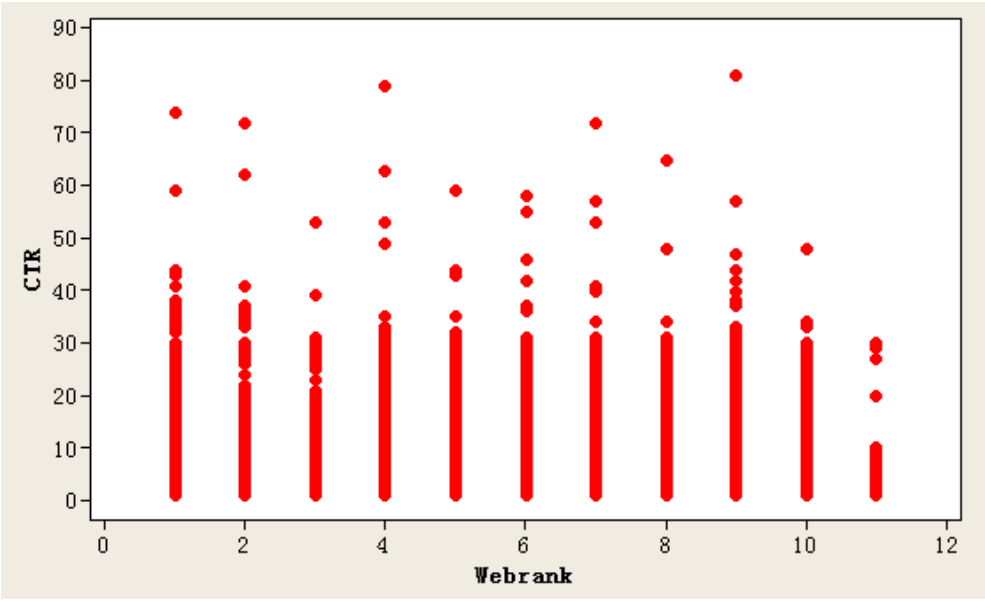

图 3.9 CTR-webrank 散点图

从上图中可以看到，随着网站排名增大点击率分布数值及数量都变得更小，也就是点击率随着网站流量的减少而降低。为了进一步验证这一观测结论，做相关性检验表如下。下表的显著性拒绝 CTR 与网站排名不相关假设，并表明两者呈负相关关系。

笔者推测，如对每个搜索开放平台广告按行业分类，如 IT 行业、汽车行业等，再根据 Alexa 流量排名对网站在该行业进行重新排序，那么排名对点击率的影响可能会更明显。但受限于时间原因，本推测可以在新的研究中再次验证。

表 3.7 CTR-webrank 皮尔森相关系数检验

		webrank	CTR
webrank	皮尔森 (Pearson) 相关	1	-.121**
	显著性 （双尾）		.000
	N	9221	9221
CTR	皮尔森 (Pearson) 相关	-.121**	1
	显著性 （双尾）	.000	
	N	9221	9221

**. 相关性在 0.01 层上显著（双尾）。

9. 用户体验（UE）与点击率之间的散点图如图 3.10 所示。前文提到，该参数完全主观，因此笔者使用了调查问卷的方式请被调查者对搜索开放平台广告 Demo 图进行人工打分（1—5）。因此，数据量较前面的客观变量少了很多，笔者共获取了 372 条有效打分。

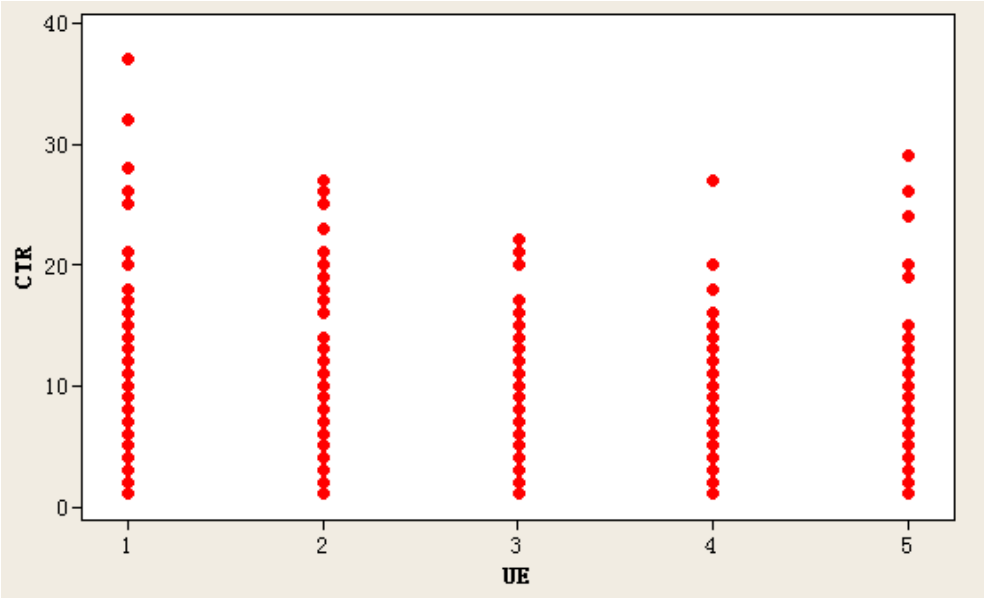

图 3.10 CTR-UE 散点图

从散点图上，并不能看到明显的相关关系，因此做相关性检验，得到下表 3.8，通过表中数据我们看到点击率与用户体验呈正相关，用户体验分值越高，点击率越高。

表 3.8 CTR-UE 皮尔森相关系数检验

		UE	CTR
UE	皮尔森 (Pearson) 相关	1	.164**
	显著性 （双尾）		.000
	N	372	372
CTR	皮尔森 (Pearson) 相关	.164**	1
	显著性 （双尾）	.000	
	N	372	372

**. 相关性在 0.01 层上显著（双尾）。

3.3 基于客观参数的静态贝叶斯网络预测模型

在第一章中讨论了构造贝叶斯网络的流程：1 定义变量；2 结构学习；3 参数学习。对于这三个流程的工作划分将构造贝叶斯网络分为了三种不同的方式：1.完整学习。三个流程完全由人定义，即领域专家确定贝叶斯网中的变量，通过专家的知识来确定贝叶斯网络的结构，并指定它的分布参数。这种方式太过主观，且受人类认知维度及知识的有限性影响极大。2.部分学习。由人主观定义贝叶斯网络中的结点变量，然后通过大量的训练数据来学习贝叶斯网的结构和参数。这种方式完全是一种数据驱动的方法，具有很强的适应性。在 matlab 的工具包中，有 K2 算法、贪婪搜索算法以及爬山算法都可以用来进行结构学习。3.完整与部分学习的结合，即由领域专家确定贝叶斯网络中的结点变量，通过专家的知识来指定网络的结构，再通过机器学习的方法从数据中学习网络的参数。本文也将使用第三种方法，通过向搜索引擎公司业内多年从业人士请教，收集到一些可能影响搜索开放平台点击率的因素，在排除了那些因权限或者其它原因不可收集到的因素数据后，笔者人工收集了相关因素数据.在结构确定的过程中，一方面是依据模型参数中存在明确的前后因果关系，另一方面也参考了多位业内从业人士的建议，最终建立了本文的提到的结构模型[30-31]。

根据上一节对各变量与 CTR 之间的相关性分析，我们首先可以认为 button、input 与 pic 这三个在搜索开放平台广告设计过程的可控变量与 CTR 存在因果关系。

尽管 rank 数据的获取在搜索开放平台广告上线之后，但该变量可控且与 CTR 存在明显的相关性，所以将作为贝叶斯网络中的一个节点。Webrank 作为一个客观不可操控变量，因在模型中代表了搜索用户对广告数据背后的广告主的信任度，与 CTR 存在明显因果关系。

搜索引擎用户在搜索某关键词后触发广告，进而导致用户对广告的点击。根据这一明显的逻辑推断，我们认为在关键词数量、展现量、点击率之间存在一条明确的因

果关系线路。

综上，我们得到搜索开放平台广告点击率的静态贝叶斯预测模型有向无环图如下图 3.11 所示。

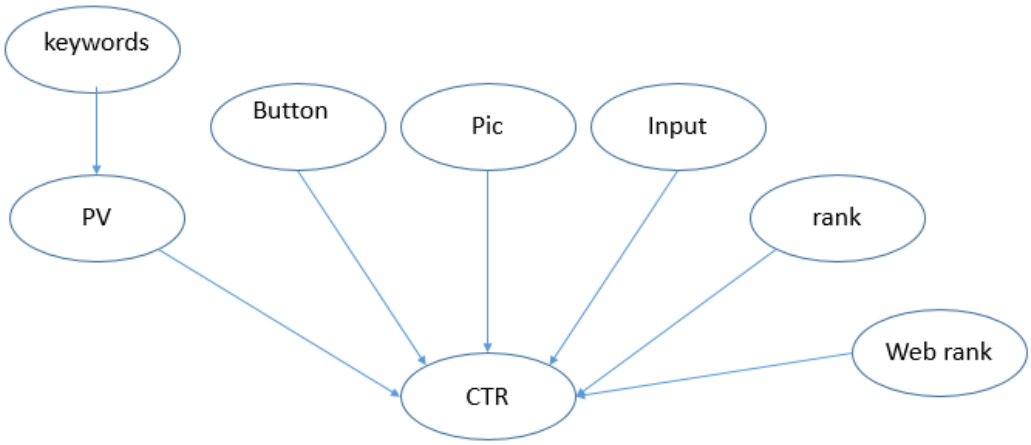

图 3.11 静态贝叶斯结构模型图

上图中的各变量描述请参见表 3.9。

表 3.9 模型参数解释说明

变量名称	变量描述	变量类型	取值范围
Keywords	广告关键词数量	离散	1,2,3,4,5,6(分级处理，方法参见表 3.2)
PV	广告展现量	连续	1,2,3,4,5,6,7(分级处理，方法同 keywords)
Button	广告按钮数量	离散	1 至 63 之间的整数
Pic	广告图片数量	离散	1-26 之间的整数
Input	广告输入项数量	离散	1,2,3,4,5
Rank	广告搜索结果排名	离散	1-12 之间的整数
Webrank	数据来源网站流量排名	离散	1-11 之间的整数（分级处理，分级方法参见表 3.1）
CTR	预测目标点击率	连续	1-80 之间的整数，单位 1 代表 1%

进一步，为了防止模型中的自变量参数存在相关关系，对于 CTR 直接相连的六个变量与 CTR 做统一相关性分析。因为 Keywords 与 PV 之间存在较强的相关性，且有先

验后验的区别，所以不需要验证 Keywords 的相关性。统一的相关性检验表如下，经过统一的相关性验证，几个自变量间不能拒绝无关假设，而自变量与因变量均存在显著相关性。因此，基于客观参数的静态贝叶斯网络预测模型将基于图 3.10 中的结构模型构造。

表 3.10 模型参数统一相关系数检验

		PV	Rank	Webrank	Button	Input	Pic	CTR
PV	皮尔森(Pearson)相关	1	.000	.012	.012	-.014	.013	.247**
	显著性(双尾)		.966	.226	.240	.153	.205	.000
	N	9921	9921	9921	9921	9921	9921	9921
Rank	皮尔森(Pearson)相关	.000	1	.007	-.014	-.003	.003	-.051**
	显著性(双尾)	.966		.494	.166	.758	.796	.000
	N	9921	9921	9921	9921	9921	9921	9921
Webrank	皮尔森(Pearson)相关	.012	.007	1	-.005	-.002	-.025*	-.280**
	显著性(双尾)	.226	.494		.587	.808	.014	.000
	N	9921	9921	9921	9921	9921	9921	9921
Button	皮尔森(Pearson)相关	.012	-.014	-.005	1	.003	-.013	.029**
	显著性(双尾)	.240	.166	.587		.788	.201	.003
	N	9921	9921	9921	9921	9921	9921	9921
Input	皮尔森(Pearson)相关	-.014	-.003	-.002	.003	1	.004	-.024*
	显著性(双尾)	.153	.758	.808	.788		.668	.018
	N	9921	9921	9921	9921	9921	9921	9921

续表 3.10 模型参数统一相关系数检验

Pic	皮尔森(Pearson)相关	.013	.003	-.025*	-.013	.004	1	-.027*
	显著性(双尾)	.205	.796	.014	.201	.668		.024
	N	9921	9921	9921	9921	9921	9921	9921
CTR	皮尔森(Pearson)相关	.247**	-.051**	-.280**	.029**	-.024*	-.027*	1
	显著性(双尾)	.000	.000	.000	.003	.018	.024	
	N	9921	9921	9921	9921	9921	9921	9921

**. 相关性在 0.01 层上显著（双尾）。

*. 相关性在 0.05 层上显著（双尾）。

3.3.1　对静态贝叶斯网络模型的假设

根据上一小节的数据分析，本文所建立的静态贝叶斯网络模型是建立在以下假设上的：

假设 1：搜索开放平台广告位所占位置越高，点击率越高。根据一般用户的浏览习惯我们知道，占据位置越靠前，越容易在搜索用户搜索结果首页被优先点击，点击的用户也就越多。排名靠后的广告，由于用户不愿意浪费更多的时间成本而往往被忽略。

假设 2：网站排名越低，点击率越低。本文中使用 alexa 排名作为网民对网站信任度的指标，网站排名越低意味着网民对网站的关注度低，名气不足以为该网站提供的数据作背书，这也就导致搜索用户对该网站点击率降低。越是排名高的网站，越容易被搜索用户信任，也更容易被搜索用户接纳。

假设 3：输入项越多，点击率越低。每一个输入项的存在对搜索用户都意味时间成本，搜索用户需要对输入项进行输入才能获取自己需要得到的数据，而人类的懒惰性使搜索用户中因好奇探索的用户不再愿意去点击这个广告。而没有输入项的广告对搜索用户来说，即使点击也并不需要付出其它成本，因此点击率自然会高一些。

3.3.2　贝叶斯网络模型各节点的概率描述

图 3.11 的贝叶斯网络模型中，各节点的定性关系表示为：

PV=PV（Keywords）　　(3.1)

CTR=CTR(PV,Button,Pic,Input,Rank,Webrank)　　(3.2)

整个模型用条件概率公式表示为：

P（Keywords，PV，Button,Pic,Input,Rank,Webrank,CTR）=P(Keyword)*P(PV|Keyword)*P(Button)* P(Pic)* P(Input)* P(Rank)* P(Webrank)* P(Rank)*P(CTR|PV，Button,Pic,Input,Rank,Webrank) (3.3)

根据全概率公式，这几个变量的量和概率表示为：

P（Keywords，PV，Button,Pic,Input,Rank,Webrank,CTR）=P(Keyword)*P(PV|Keyword)*P(Button)* P(Pic)* P(Input)* P(Rank)* P(Webrank)* P(Rank)*P(CTR|Keywords,PV,Button,Pic,Input,Rank,Webrank) (3.4)

由此，我们得出

P(CTR|PV，Button,Pic,Input,Rank,Webrank)=

P(CTR|Keywords,PV,Button,Pic,Input,Rank,Webrank) (3.5)

通过上式，我们可以得出在该静态贝叶斯网络中，在给定PV、Button、Pic、Input、Rank、Webrank下，点击率独立与关键词数量（keywords）。

3.3.3 基于客观参数的参数学习

通过与多位搜索引擎领域从业多年人士的交流，确认了贝叶斯网的结构模型之后又对各节点之间的关系进行量化。

参数学习的过程主要是通过Matlab中有一个名为FULLBNT的工具包来实现的。通过该工具包构建模型过程中所用到的变量为：关键词数量（Keywords）、搜索开放平台广告排序位置（Rank）、广告展示量（PV）、图片数量（Pic）、输入项数量（Input）、按钮数量（Button）、网站排名（Webrank），在前文中的数据处理以及提到，为了平衡计算时间与精准度，对所有的参数进行了分级处理，将所有非目标变量设定为离散型（tabular），而目标参数因为可取数值较多，即使进行离散处理，仍可以当做连续函数，因此设定自变量为连续型（gaussian）。运用到的函数包括[32]：

mk_bnet()：静态贝叶斯网络模型构造函数。通过输入由矩阵表示的结构图，以及各节点可能取值的数量(离散节点数量皆标为可取离散值数量，连续节点可取值数量为1)，即可生成符合结构图的静态贝叶斯模型对象。该函数中有许多可以指定的参数，如通过对‘discrete’参数指定离散节点、对‘observed’指定所有已知数据节点等，这些设置可以有效的帮助贝叶斯网络更多地掌握节点信息，从而节省参数学习时间。

tabular_CPD ()：构建离散型节点函数。在生成贝叶斯模型对象后可用该函数及gaussian_CPD()函数对每个节点进行设定离散或者连续的设定，在本文中使用的函数仅涉及简单的离散或连续节点，因此涉及的节点设定函数仅有这两个。如果对节点的父节点或子节点有额外要求，还有softmax_CPD()或者binary_CPD()等函数可用。

gaussian_CPD()：构建连续型节点函数。

learn_params()：参数学习函数。通过输入已有的贝叶斯网络及历史数据，可以生成新的进行参数学习过的贝叶斯网络对象。

jtree_inf_engine()：Matlab 中对联合树推理引擎的调取函数，每个生成的贝叶斯网络对象都有一个对应的推理引擎，通过该函数可以实现团树传播法的推理过程，通过参数学习后贝叶斯网络模型需要使用该函数测试数据进行预测验证。

enter_evidence()：对已有的贝叶斯网络模型输入证据的函数。通过该函数把证据加入参数学习后的引擎中，从而进行未知节点数据的推断。

在参数学习的过程中，需要重点关注的 matlab 函数就是 learn_params（）函数。贝叶斯网络的参数学习大致可以分为两种方法：贝叶斯估计和最大似然估计。而本文中所使用的参数学习函数 learn_params（）背后调用的方法就是最大似然估计的方法。其背后的逻辑解释如下：假设一个贝叶斯网络 BNET 由 n 个变量构成，它们分别用 X_1、X_2…X_n 表示。其中节点 X_i 有 x_i 个取值 1,2，…，x_i，其中父节点 Parents(X_i)取值共有 l_i 个组合 1,2, …，l_i，若 X_i 无父节点，则 l_i=1,则贝叶斯网络参数为：

$$\theta_{ijk} = P(X_i = k \mid Parents(X_i) = j) \tag{3.6}$$

用 θ 记所有 θ_{ijk} 所组成的向量，根据数据样本 D=（D_1，D_2，…，D_n）构造参数的对数似然函数如下所示:

$$L(\theta \mid D) = \log \prod_{i=1}^{n} P(D_i \mid \theta) = \sum_{i=1}^{n} \log(P(D_i \mid \theta)) \tag{3.7}$$

通过已有数据对对数似然函数求解最大值，所得解 θ 即为贝叶斯网络的参数。

3.3.4　基于客观参数的预测分析

通过上一小节所提的方法，对已得的结构模型用最大似然估计的方法进行了参数学习后，就可以通过该模型进行我们的点击率的预测。贝叶斯网络主要有预测推理、诊断推理以及原因关联推理等推理功能。而本文运用的则是它的预测推理功能，即根据已知原因数据预测推理结果。将已有数据拆分为学习集和验证集，90%（8921 条）用于参数学习，10%用于验证（1000 条）。在给定搜索关键词数量(Keywords)的条件下，利用模型对展现量数值（PV）进行预测推理。图 3.12 中给出了是展现量的实际值、预测值及两者之间的差值。

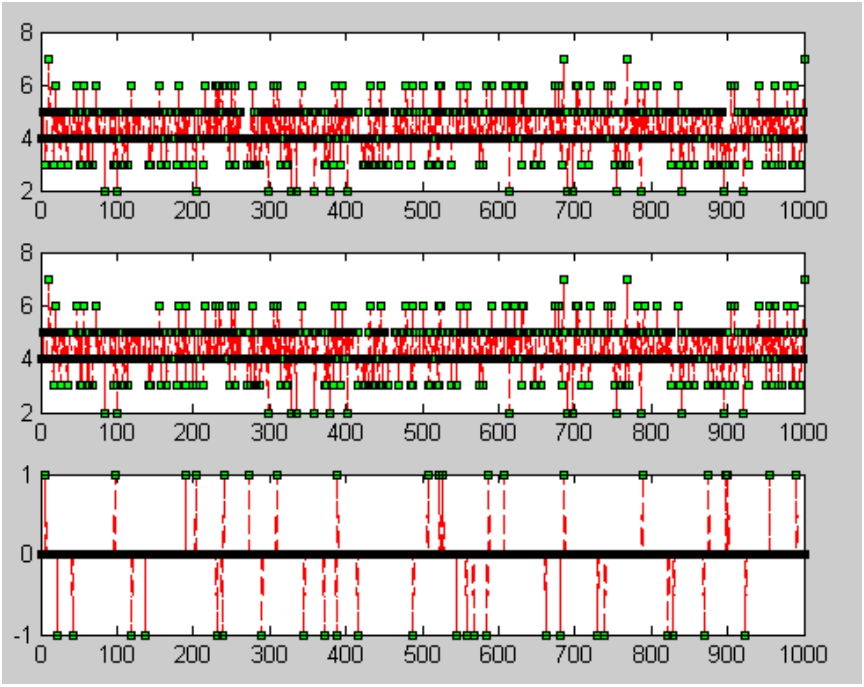

图 3.12 PV 的实际值、预测值与差值图

在给定 keywords、rank、webrank、button、pic、input 的前提下，利用模型对广告点击率（CTR）值进行预测估计。图 3-13 由上至下给出了 CTR 的实际值、预测值及两者之间的差值。

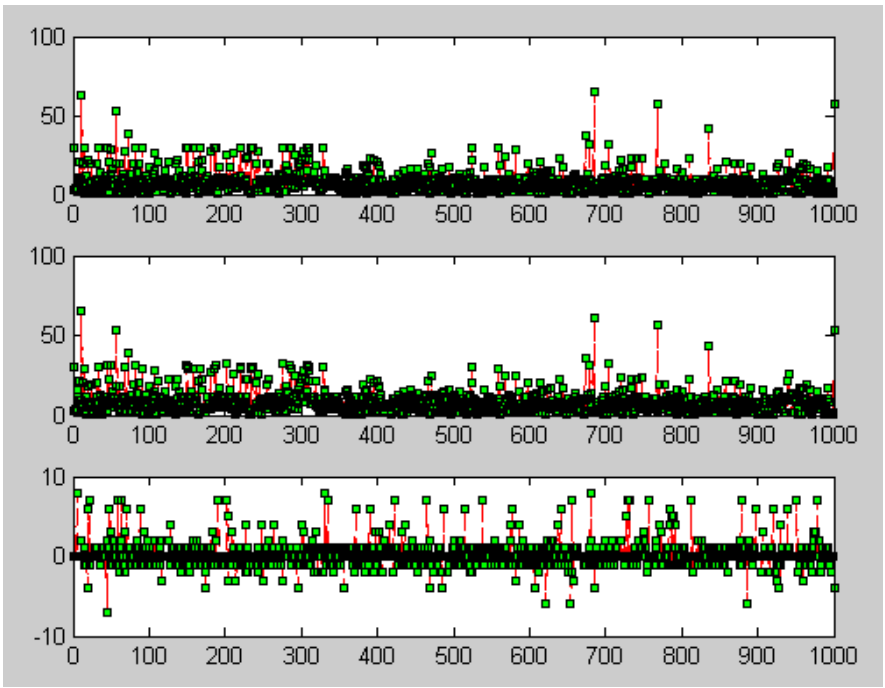

图 3.13 CTR 的实际值、预测值与差值图

对 PV、CTR 的预测值及其预测差值进行统计分析，其统计值如表 3.11 所示。

表 3.11 各变量预测值与预测差值的统计表

变量	最小值	最大值	均值	标准差
实际 PV	2	7	4.33	0.80
预测 PV	2	7	4.32	0.79
PV 预测差值绝对值	0	1	0.05（1.02%）	0.21
实际 CTR	1	65	8.58	7.64
预测 CTR	1	65	8.56	7.56
CTR 预测差值绝对值	0	8	0.94（10.99%）	1.72

本表中的偏差均使用绝对值进行统计，之所以采用绝对值，是为了防止出现偏差正负相抵的情况使的错误高估预测准确程度，预测结果无论正负都会对实际结果产生相当巨大的影响：若预测值较实际值高，则会使人高估广告的商业价值。若预测值较实际值低，一方面可能使搜索引擎公司放弃实际存在的收入，另一方面也会使广告主低估流量的冲击，无法提前进行准备。对偏差值取绝对值再进行均值计算，则结果如表 3-9 所示，广告展现量预测估计值的均值偏差绝对值为 0.05，占实际展现量均值的 1.02%，点击率预测估计值绝对值的均值为 0.94，占实际点击率均值的 10.99%。我们可以看到，展现量的偏差较小，而点击率的偏差较大，这样的结果可能是因为展现量对点击率的影响并不是最重要的或者两者之间存在隐节点，也可能是展现量的分级可能不够精确，不足以对点击率的分级进行更精确的划分。总的来讲，模型在对搜索开放平台广告点击率的预测上偏差较大，说明模型仍有进一步改进的空间。

在本文中，因我们的目标是对点击率进行预测，在贝叶斯网络中属于预测推理，因此在本文中训练生成的模型并没有完全得到应用。如果我们希望进行诊断推理，“由果溯因”也是可以的。如假设我们已知搜索开放平台广告点击率为 30%，则可以将该已知信息输入函数，进而得出该搜索开放平台广告的排序 Rank 分布如下表 3.12 所示。

表 3.12 点击率为 30%的搜索开放平台 Rank 分布

Rank 位	1	2	3	4	5	6
出现概率	57.5%	23.2%	8.3%	5.0%	3.3%	2.8%

在本文中因为节点参数过多，对完整生成的转移概率表不做更多展示，事实上，上表给出的就是 P（Rank|CTR=30）的转移概率。

3.4　引入用户体验参数的贝叶斯网络预测模型

本文在前一节中建立了所有节点均为客观参数的静态贝叶斯网络模型，初步实现了点击率的预估。但在互联网公司中存在一个耳熟能详的词——用户体验，我们说一个产品好一般都会说好使，并不会说因为它有几张图所以好使好用。好使好用，即用户体验好，和上一节中的使用的参数不同，这是一个完全主观的变量。在搜索开放平

台广告中，相同的参数，可能因为布局好亦或图片选择更贴切会促成用户体验好这一事实，从而使得点击率提升。从逻辑上讲，用户体验肯定与前文中的图片 Pic、输入项 Input 及按钮 Button 有关，而它又会对结果产生影响，所以将完全主观的用户体验节点纳入到模型中。图 3-14 中展示的就是本文所构建的纳入用户体验节点的贝叶斯网络结构模型。

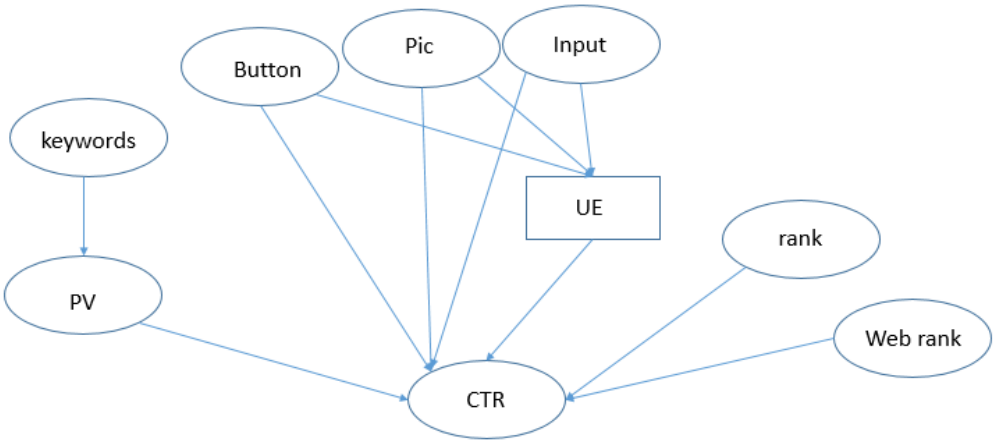

图 3.14 引入用户体验的贝叶斯结构模型图

上图模型的其他参数均与图 3.11 中的模型相同，因此不做过多解释。唯一新增变量为用户体验 UE，该数据主要是通过人工对广告 Demo 图打分的方式获得，详见表 3.13。

表 3.13 引用用户体验的模型参数解释说明

变量名称	变量描述	变量类型	取值范围
Keywords	广告关键词数量	离散	1,2,3,4,5,6(分级处理，方法参见表 3.2)
PV	广告展现量	连续	1,2,3,4,5,6,7(分级处理，方法同 keywords)
Button	广告按钮数量	离散	1 至 31 之间的整数
Pic	广告图片数量	离散	1-19 之间的整数
Input	广告输入项数量	离散	1,2,3,4,5
Rank	广告搜索结果排名	离散	1-12 之间的整数
Webrank	数据来源网站流量排名	离散	1-11 之间的整数（分级处理，分级方法参见表 3.1）

续表 3.13 引用用户体验的模型参数解释说明

UE	主观变量用户体验	离散	1,2,3,4,5
CTR	预测目标点击率	连续	1-28 之间的整数，单位 1 代表 1%

3.4.1　对拥有用户体验节点的贝叶斯网络模型的假设

根据上一小节中对用户体验参数的的描述，我们在 3.2 节的基础上得出如下假设：

假设 4：搜索开放平台广告点击率随搜索用户对其的用户体验评价提升而提升。

表 3-14 中给出的是引入的用户体验参数与前一节提到的其他客观节点之间的皮尔森系数检验结果，由表中数据我们可以看到 UE 与 Button、Input、Pic 这些外观参数相关，且对 CTR 存在影响。

表 3.14 引入用户体验参数的模型参数统一皮尔森相关系数检验

		PV	Rank	Webrank	Button	Input	Pic	CTR	UE
PV	皮尔森(Pearson)相关	1	.000	.012	.012	-.014	.013	.247**	-.012
	显著性（双尾）		.966	.226	.240	.153	.205	.000	.222
	N	372	372	372	372	372	372	372	372
Rank	皮尔森(Pearson)相关	.000	1	.007	-.014	-.003	.003	-.051**	.006
	显著性（双尾）	.966		.494	.166	.758	.796	.000	.519
	N	372	372	372	372	372	372	372	372
Webrank	皮尔森(Pearson)相关	.012	.007	1	-.005	-.002	-.025*	-.280**	-.010
	显著性（双尾）	.226	.494		.587	.808	.014	.000	.338
	N	372	372	372	372	372	372	372	372

续表 3.14 引入用户体验参数的模型参数统一皮尔森相关系数检验

Button	皮尔森(Pearson)相关	.012	-.014	-.005	1	.003	-.013	.029**	-.026**
	显著性（双尾）	.240	.166	.587		.788	.201	.003	.004
	N	372	372	372	372	372	372	372	372
Input	皮尔森(Pearson)相关	-.014	-.003	-.002	.003	1	.004	-.024*	-.032**
	显著性（双尾）	.153	.758	.808	.788		.668	.018	.000
	N	372	372	372	372	372	372	372	372
Pic	皮尔森(Pearson)相关	.013	.003	-.025*	-.013	.004	1	-.007	-.018**
	显著性（双尾）	.205	.796	.014	.201	.668		.514	.005
	N	372	372	372	372	372	372	372	372
CTR	皮尔森(Pearson)相关	.247**	-.051**	-.280**	.029**	-.024*	-.007	1	.064**
	显著性（双尾）	.000	.000	.000	.003	.018	.514		.000
	N	372	372	372	372	372	372	372	372
UE	皮尔森(Pearson)相关	-.012	.006	-.010	-.026**	-.032**	-.018**	.064**	1
	显著性（双尾）	.222	.519	.338	.004	.000	.005	.000	
	N	372	372	372	372	372	372	372	372

**. 相关性在 0.01 层上显著（双尾）。

*. 相关性在 0.05 层上显著（双尾）。

3.4.2 引入用户体验参数的贝叶斯网络参数学习及预测分析

在上一节的参数学习方法一样，使用一样的函数重新构造包含用户体验参数的贝叶斯网模型，并定义节点。对用户体验这一节点的定义为离散型，即 tabular 类型，该节点大小为 5（1—5 的离散打分值）。

本文所收集到历史数据集中没有缺失项，因此在参数学习过程运用的是完备的最

大似然估计法 MLE，与上一个模型的参数学习方法相同。

因为用户体验这一数据只能通过调查问卷的方法进行调查，所以搜集到的历史数据有限，总样本为 372 个。因此利用 300 个历史数据进行已知结构的参数学习，利用 72 个样本数据对建立的模型进行估计来验证所建立的模型是否有效。图 3.15 中所示的根据给出的 Button、Pic、Input 对 UE 的预测，和图 3.11 不同的是，本图的上半部分在一张图上给出了 UE 的实际值和预测值，其中实际值为绿色，预测值为红色，在图中可以明显看到有许多点重合，因此预测比较准确。为了更细地看清偏差，下半部分为实际值与预测值的差值图。

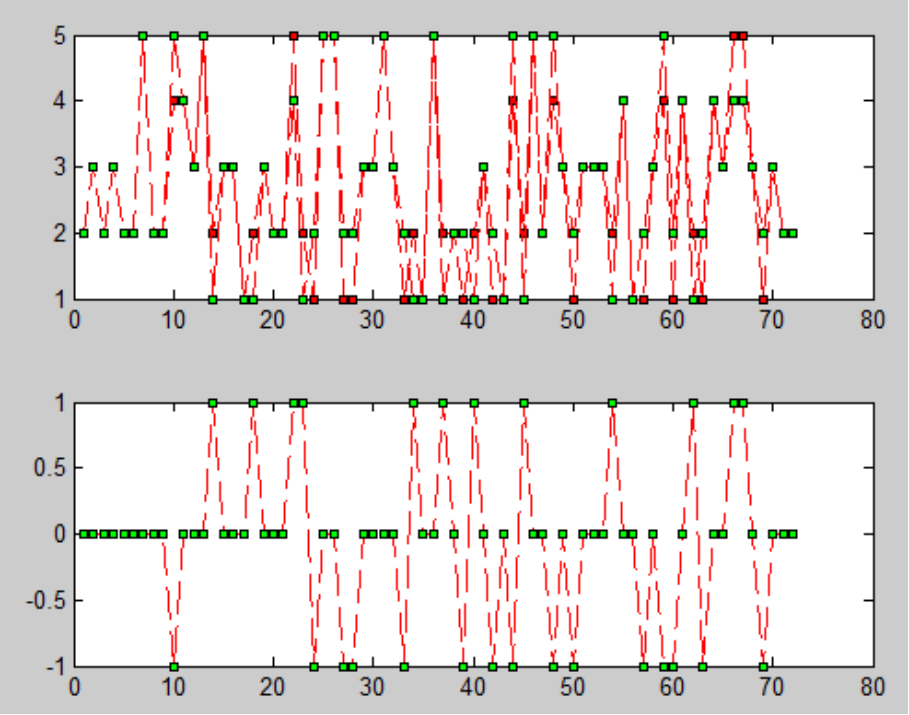

图 3.15 UE 的实际值、预测值与差值图

在给定 keywords、rank、webrank、button、pic、input 的前提下，利用模型对广告点击率 CTR 进行预测估计。图 3-16 和图 3-15 一样，将实际 CTR 用绿色标出，预测 CTR 用红色标出，差值由下半部分给出。

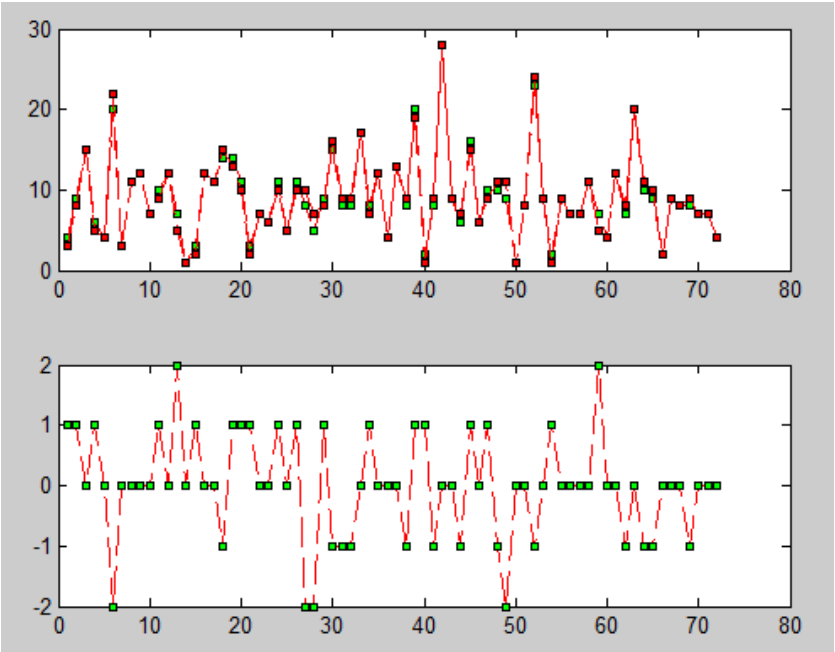

图 3.16 CTR 的实际值、预测值与差值图

对 UE、CTR 的预测值及其预测差值进行统计分析，其统计值如表 3.15 所示。其中差值均已绝对值呈现，避免正负偏差的抵消。

表 3.15 各变量预测值与预测差值的统计表

变量	最小值	最大值	均值	标准差
实际 UE	1	5	2.72	1.30
预测 UE	1	5	2.68	1.31
UE 预测差值绝对值	0	1	0.375（13%）	0.49
实际 CTR	1	28	9.15	5.12
预测 CTR	1	28	9.15	5.30
CTR 预测差值绝对值	0	2	0.58（6.34%）	0.64

根据表 3-15 中的数据分析得出，广告用户体验 UE 预测估计值的均值偏差为 13%，点击率预测估计值绝对值的均值偏差为 6.34%。相较于表 3-10 中的数据，对于我们关注的重点目标广告的点击率，模型预测的准确度有很大的提高，已经小于 10%，达到我们可以接受的偏差范围内。由此我们可以得出，在基于客观变量建立的贝叶斯网络模型的中，引入主观评价参数——用户体验 UE 确实比单纯依赖客观参数的预测效果更准确。

3.5 本章小结

本章首先根据多位互联网从业人员的实际操作经验对搜索开放平台广告的上线流程选择了一些关键参数并对它们进行了描述；其次结合实际经验通过统计方法对这些

参数之间的关系进行了相关性的假设检验；随后在确认相关性的前提下，笔者参考了参数的因果逻辑及多位业内专家的建议确立了贝叶斯网络的结构模型；最终通过历史数据对模型参数进行了学习及验证。在结构模型的确立环节，本章根据是否引入主观评价节点——用户体验 UE 建立了两个模型，实证分析结果证明引入主观评价节点的贝叶斯网络模型比节点完全为客观参数的贝叶斯网络模型的预测效果更精确，为第 4 章的搜索开放平台广告上线前的流程研究提供了基础。

第四章　搜索开放平台广告上线策略研究

前文的研究使搜索开放平台广告在上线前的点击率预估成为可能，使得搜索引擎公司对该类广告效果变得更易把控，使流程更容易变得标准化、规范化，也更易使公司的商业流程运作变得有效。在本章，将把重点放在模型的应用环节，并根据搜索开放平台广告点击率影响因素讨论其上线前后如何使广告效果达到最优。

4.1　设计环节优化建议

前面提到搜索开放平台广告上线前需要进行展现效果设计，在这一环节，根据之前的模型研究，存在三个会影响广告最终上线效果的变量：图片数量、输入项数量、按钮数量。

1.图片数量。在搜索开放平台广告的展现效果上，有图片必然会使广告变得更加醒目，但是更多的图片则使点击率下降（见第三章的相关性分析）。尽管本文没有研究图片质量对点击率的影响，但是笔者认为图片对搜索开放平台广告点击率的影响可以归结为“在精而不在多”。使用一个足以使搜索用户对广告内容理解的图片，而不是使用铺天盖地的图片使搜索开放平台广告展现结果广告化严重到使搜索用户对广告选择性屏蔽无疑是设计人员在图片添加这一环节应当遵守的原则。目前，受限于网速原因，大部分的搜索开放平台广告还未将 GIF 图片甚至是 FLASH 作为常规化，但是一个形象生动的动态图片的动态变幻效果对比搜索结果页其他结果以静态字符栏目展现毋庸置疑地是更吸引人的。所以，搜索开放平台广告在设计环节对图片应遵守减少数量，加强质量的原则。

2.输入项数量。输入项对搜索用户意味着时间成本，越多的输入项，点击率越低。但是本规则有一个前提，即输入项需满足搜索用户的最基本功能需求。例如线路搜索，搜索用户需要了解从起点到终点的线路前进，因此需要两个输入项来进行输入项与输出项的填充。但是设计人员刻意地使输入项减少，只有一个输入项，那么结果很可能是搜索用户搞不明白如何使用该平台，最终没有点击，即点击率为零。诚然，搜索开放平台广告可能无法在现有设计的模式下对输入项直接减少，但是却可以从设计逻辑的角度规避输入项。如下图示例，同样是机票的搜索开放平台广告查询，当用户搜索词包含了起点终点时，直接使平台抓取搜索词，尽管仍然是两个输入项，但对搜索用户来说却无需进行人工输入，两项输入项就被成功地消除掉了。

图 4.1 从设计逻辑上消除输入项示例

归纳起来，在满足搜索用户需求功能前提下，想方设法地减少输入项，是设计人员在搜索开放平台广告设计流程中应当遵守的原则。

3. 按钮数量。在第三章中我们已经证明，按钮数量增多意味着一个广告的点击率下降。产生这种情况的原因很可能是按钮使广告的广告性质变得更加明显，但是笔者认为在小区间中可能存在着点击率随按钮增多而增大的情况。本次论文研究并未发现想象中的按钮恰当值，因此只能建议在确保按钮在满足功能的前提下竟可能地减少按钮数量。在按钮设计环节，笔者的分析建议是让按钮在可以布置“伏笔”的位置添加，即把按钮添加在搜索用户最可能产生好奇需要进一步探索的位置。这样按钮对搜索用户就是“恰好”存在，满足用户所需，而这样的按钮肯定会提升该广告的点击率不会产生按钮增多点击率下降的情况。

除了上述三个单独的环节，在设计流程上还应综合考虑搜索结果，即搜索结果上下文。对于某个汽车的搜索词，如果已经存在多个搜索开放平台广告结果进行覆盖，当针对一个新的数据内容进行设计样式效果时，设计人员很有可能就需要考虑其它的结果是否存在图片，按钮数量与布局是否太过雷同等因素。把新的样式设计结果设计成为更搭配之前的搜索结果，使广告更自然更像是一个整体，是现在设计人员乃至搜索引擎公司都应努力的方向。

整体地讲，笔者更愿意把上述的因素都归纳为用户体验。搜索用户在使用搜索引

擎的过程中，每一个细微地环节都在影响着用户对服务的评价，无论是图片、按钮、输入项还是搜索结果的展现形式亦或触发形式都在影响着用户体验。如果考虑搜索引擎这一个大整体的因素，那么影响用户体验的因素则更多，很可能因百度的搜索结果更多更全，好搜的搜索结果链接更安全，而搜狗的搜索结果更多的覆盖了社交信息而导至搜索用户对不同的搜索引擎打分也不同。

用户体验，一个完全主观的概念，却无限地提升着生产者或者服务者的产品（服务）水平.

本节将引用某一个培训类的搜索开放平台广告在样式上多次优化的例子来证明模型点击效果预估的作用在设计优化环节所起到的作用，对点击率的预估将可以使搜索引擎公司避免在点击率较低的广告样式上进行人力投入。

该项目背景为某在线教育客户 A 计划对其在线培训项目计划在搜索引擎公司 B 进行搜索开放平台广告推广，在审核其数据质量过关的前提下，于是有了如下图 4.2 所示的第一个 Demo 图。在上线点击率并不足以满足 A 预期的情况下，B 设计了新样式 b，但是结果并未令人满意，甚至点击率下降。最终 B 设计了样式 c，结果差强人意，A 最终付费。整个优化过程一个半月，消耗人力 4（A 公司 2 人，B 公司 2 人）。

图 4.2 培训类搜索开放平台广告样式 a

图 4.3 培训类搜索开放平台广告样式 b

图 4.4 培训类搜索开放平台广告样式 c

从样式 a 到样式 b 的逻辑其实比较简单，产品设计逻辑未变，希望通过增多图片和增多按钮的方式增加单次搜索的点击可能性，从而增大点击率。而从 b 到 c 则放弃

了原有样式的设计布局，重新进行了规划。

上述案例是某搜索公司中真实的搜索开放平台广告上线案例，这也是在未有点击率指导的前提下进行设计优化的典型案例。然而，上述结果其实完全可以通过提前的点击率预估进行规避。

提取图 4.2-4.4 所示样式图的参数，通过贝叶斯网络模型进行点击率预估，所得结果如下表 4.1 所示。其中因为数据来源一致，且触发词和展现量一致，在本表中不做展示。

表 4.1 搜索开放平台广告 a,b,c 样式点击率预估

样式名称	按钮数量	图片数量	输入项数量	用户体验分	预估点击率	实际点击率
样式 a	14	3	0	3	10%	8%
样式 b	22	5	0	2	8%	7%
样式 c	10	4	0	4	13%	13%

根据表 4.1 所给的对比结果，可以清楚的发现样式 a 和样式 b 的上线结果其实在产品设计之初就已经有了点击率较低的倾向，预测值低于 10%，而填充上线也就变得略显多余。至于样式 c 的设计，一方面更改了布局，另一方面也减少了按钮和图片的数量，使得用户体验更好，打分值更高，点击率也就有所上升。事实上，如果在设计之初可以对点击率有所预估，那么至少可以规避掉样式 a 和样式 b 的数据填充时间和人力成本，对 A 和 B 两公司的效率提升还是非常显著的。

4.2 上线环节优化建议

上线环节其实关系着本次研究模型中的三个变量：搜索开放平台广告排序位置、触发词数量以及数据来源网站流量排名，其中前两个变量都是搜索引擎公司在广告上线前决定的，也是相对可控的两个变量。而最后一个变量，尽管针对某一个客户来说是相对不变的，但是在广告主切换时却也对搜索引擎公司收益有着较大的影响。

4.2.1 搜索开放平台广告排序位置的优化

在搜索开放平台广告排序位置这一变量上，如搜索结果页只有一个搜索开放平台广告，那么逻辑相对简单，排序位置越靠前点击率越高。因此，为了尽可能地提升点击率，应将搜索开放平台广告初上线的位置排在第一位，甚至可以考虑固定排在第一位。这样充分提高点击率，向客户导入更多的流量，从而增大广告收益。

但是，在某些搜索结果中，搜索开放平台广告往往并不是一个。最典型的例子就是搜索武侠小说类的词，该类词包可能会触发三个搜索开放平台广告结果：其一为针对小说阅读需求给出的书籍阅读广告；其二为同名小说改编的电视剧或者电影的广告；

其三为同名小说改编网游的广告。当出现多个广告并存的搜索结果时，前面提到的将广告排序位置摆放越靠上越好的道理就不能直接实施了。但是，恰恰是因为有着较为精准的点击率预估，搜索引擎公司可以利用点击率进行一个较为准确的计算，计算如何排序广告可以使公司的整体收益最大化。

已知搜索开放平台广告产生的收益公式如下所示：

Return = PV(展现量)×CTR(点击率)×CPC（触发词点击均价）　　（4.1）

在上式中，触发词点击均价遵循着一定的经济规律，即词包所在行业盈利能力越高，该行业愿意付出的点击成本也就越高，因此 CPC 也就越高。CPC 的计算在搜索引擎公司已经有一套比较成熟的规则，在这里就不做展开，在本研究中可以将其当做已知研究。

综合上述公式，当同个触发词包存在 n 个搜索开放平台广告结果时，每个广告展现量相同，总收益公式最大化的函数如下所示：

Max(Return)= Max（ $PV\sum_{i=1}^{n} CTR_i \times CPC_i$ ）　　（4.2）

在该公式中，可变变量为影响 CTR 的广告位排位，通过对不同广告的广告位顺序进行排列组合，求取点击率与在该行业点击均价乘积之和的最大值，即为所求。以同一触发词包触发三个搜索开放平台广告 a,b,c 为例对点击率的应用进行解释。通过点击率预估，我们可以得到 a,b,c 三个广告排序在广告位 1,2,3 时的点击率。因为在 abc 分别三个行业的点击均价已知，对 abc 广告位置进行排列组合，共有六种情况，进而得到下表。

表 4.2 多搜索开放平台广告共存分析表

abc 排列情况	a 的 CPC	b 的 CPC	C 的 CPC	点击率	点击率	点击率	乘积之和
123	5	4	3	12	11	9	131
132	5	4	3	12	8	12	128
213	5	4	3	10	13	9	129
231	5	4	3	10	8	13	121
(*) 312	5	4	3	9	13	12	133
321	5	4	3	9	11	13	128

通过上表 4.2，分析得出当搜索开放平台结果 b 排序在第一位，c 排序在第二位，d 排序在第三位时，对公司产生的总收益最大。当 a、b、c 广告排列次序为 2、3、1 时，公司的总收益最小，最大收益较最小收益高 10%。可以解释为，在搜索引擎公司并未作出其他设计优化的情况下，仅变更各广告位的排序，即有可能增加公司广告收益 10%。

4.2.2　搜索开放平台广告触发词的优化

在模型构建过程中已经有触发词数量越多，搜索用户触发该广告的几率越大的结

论。在确保触发词与广告内容相关的前提下，触发词越多，则点击率越高。笔者推测，触发词的增多覆盖了更多的长尾词，从而使得长尾需求被覆盖。长尾词的点击率高于搜索开放平台广告的平均点击率，增加长尾词从而使得广告整体点击率被拉高。因此，运营人员应极尽可能地把相关词添加到触发词列表，特别注重长尾词的添加，从而实现点击率的提升。

4.2.3 谨慎选择流量排名较低的网站对接数据

在现实生活中，并不是广告效果满意广告主就一定会持续购买广告，广告主的预算出现问题、企业转型放弃某产品推广的例子不胜枚举。在某搜索开放平台广告上线一段时间之后，该客户如果放弃投放广告，那么搜索引擎公司需要做的事情并不是重新去设计一个新的广告样式，而是在已有广告样式的前提下寻找拥有相似数据的广告主，争取对接相似数据。这样可以避免前期样式设计成本的浪费。而这一环节也就引出了对接数据的网站选择。

在本文第三章的模型中有一个名为 Webrank 的变量，即网站流量的排名，但它衡量的实际是网民对这个网站的信任程度。流量越多，代表着网民对其越信任，访问的频率或者人口占比也就越高。

下表为某一理财类搜索开放平台广告在更换对接网站后，点击率的变化情况。

表 4.3 某理财搜索开放平台广告变更客户后点击率变化

状态	流量排名	点击率	点击率预估
数据变更前	400	15%	15%
数据变更后	50000	5%	6%

在表 4.3 中，因广告样式、触发词、排名等因素并未改变，所以不做描述。但仅仅是一个流量排民的变更，就已经造成了点击率的惊人变化。而这样大的点击率变化也导致了搜索引擎公司在变更对接数据后，收入锐减 2/3，对公司收入产生了不好的影响。其实，从上表也可以看到，如果在对接数据之前对流量排名这一参数进行考量，提前对点击率进行预判，这样的结果极大的可能会被规避掉。

事实上，流量排名这一变量衡量的不单单是网民对某网站的信任度，同时也体现了网站的规模及实力，流量越大，实力也就越强。在特别是理财、保险等搜索用户使用较为谨慎的互联网产品上，实力越强也就意味着更大的担保能力，因此也更容易受到搜索用户的偏爱。

4.3 本章小结

本章结合了第三章的研究内容，分搜索开放平台广告的设计流程和上线流程两部

分，分别分析了环节中涉及的变量对搜索开放平台广告点击率存在的影响，针对变量对点击率的影响提出了优化建议，使得广告的优化过程融合在了广告上线前个流程中，使搜索引擎公司商业广告运作上更具效率。

第五章　结论与展望

本文通过文献回顾，在总结前人研究成果的基础上，提出了本文研究的理论模型和相关假设。本文以搜索开放平台广告上线流程为基础，提取出了搜索开放平台广告的触发词数量、广告排名、广告展现次数、广告图片数量输入项数量、按钮数量和网站排名等相关变量，在第三章中依据是否引入主观变量用户体验评分，分别构建了两个结构模型，针对两个模型利用历史数据分别进行了参数学习。实证结果表明完全客观的数值并不足以完全解释广告点击原因，引入主观变量预测效果优于纯客观变量的模型，有效地提升了预测精度。这对搜索引擎公司利用搜索开放平台广告进行商业盈利存在很大的便利。

全文主要研究内容如下：

1.通过理论研究提取出相关变量和假设。本文在研究过程中根据搜索开放平台广告上线流程，提取出了搜索开放平台广告的触发词数量、广告排名、广告展现次数、广告图片数量输入项数量、按钮数量和网站排名等建模过程中的相关变量，并利用搜集到的数据对变量关系进行验证。

2.通过理论研究提出贝叶斯网络各节点均为客观节点的网络模型。结合前人的研究及领域专家的建议确定了贝叶斯网络的结构模型，首先对开放平台广告上线的数据进行收集整理，其次对这些整理后的数据进行参数学习及验证。

3.本人创新地将完全主观变量用户体验UE引入了对搜索开放平台广告点击率预测的模型当中。笔者在对上一模型进行检验后发现利用纯客观参数的预测模型对点击率进行预测并不能很好地解释某些实际结果后，结合互联网经常提及的用户体验这一概念，引入用户体验这一主观评价参数，将其通过调查问卷打分的形式量化。在上一个个贝叶斯网络模型基础上，结合用户体验打分变量构建了主观与客观相结合的贝叶斯网络模型。

4.根据研究结果，对搜索开放平台广告的上线流程与设计流程提出了优化建议。

因为在因素选择上，可能对广告的正常运转没有影响所以没有在数据库中存在记录，本文的研究过程中也不可避免的存在一些不足，而这些不足的改进可能意味着模型的进一步优化，在搜索开放平台广告点击率的预测上仍有进一步上升的空间：

1.网站排名这一变量是笔者对搜索用户对广告信任度的一个替代，但是不同行业的网站排名相比较低并不意味着信任度的降低，很有可能只是对该行业整体关注度不高。因此对网站排名可进一步演化为对同一行业的网站排名，或者用其他用户信任指标来进行衡量。

2.搜索开放平台广告的点击率影响因素还有加载速度、信息覆盖饱和度等。受限于数据获取难度，无法将这些参数引入模型。而作为贝叶斯网的一大优点就是可以推测一些隐含节点的参数，从这个角度，在本文的模型的基础上可以再引入隐含节点实现模型对某些不确定参数的估计，实现模型精准度的提高[33]。

参考文献

[1] 何咏梅,毛云舸.搜索引擎的发展现状与趋势研究[J].吉林省经济管理干部学院学报,2007,21（4）:65-68.

[2] 徐涛.企业搜索引擎广告策略研究[D].上海:华东师范大学情报学硕士学位论文,2010:11.

[3] Jakob Nielsen,F-Shaped Pattern for Reading Web content,2006.

[4] 中国互联网络信息中心.第 35 次中国互联网络发展状况统计报告[R]. 2015(13):2-3.

[5] 童强.搜索引擎关键字广告点击率与保留价研究[D]. 大连:大连理工大学硕士学位论文, 2011.

[6] Haibin Cheng, Erick Cantu-Paz. Personalized Click Prediction in Sponsored Search, 2010: 351-359.

[7] Bernard J. Jansen,Simone Schuster. Bidding on the buying funnel for sponsored search and keyword advertising [J]. Journal of Electronic Commerce Research,2011,12(1):1-18.

[8] 许建盈. Google 关键词广告竞价的收益率预测[J]. 科学技术与工程,2008,8（14）:3868-3871.

[9] 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.2009.

[10] Jason Auerbach , Joel Galenson, Mukund Sundararajan. An Empirical Analysis of Return on Investment Maximization in Sponsored Search Auctions [C]. ADKDD, 2008:1-9.

[11] Benjamin Edelman, Michael Ostrovsky.Strategic bidder behavior in sponsored search auctions [J]. Decision Support Systems, 2007 (1): 192-198.

[12] Ilya Gluhovsky. Forecasting Click-through Rates Based on Sponsored Search Advertiser Bids and Intermediate Variable Regression [J]. ACM Transaction on Internet Technology, 2010, 10(3).

[13] A. Animesh, V. Ramachandran and S. Viswanathan. Quality uncertainty and adverse selection in sponsored search markets. ACM Conference on Electronic Commerce (EC'05).

[14] W Lam, F Bacchus. Learning bayesian belief networks: An approach based on the mdl principle [J]. Computational Intelligence (S0888-613X). 1994, 10: 269–293.

[15] Don Schultz, SIVA 范式（搜索引擎触发的营销革命）[M]. 北京:中信出版社,2014.

[16] 魏平.一些预测算法的研究与应用[D].杭州:浙江大学硕士学位论文,2004.

[17] Jordan M I. Learning in graphical models. Massachusetts :MIT Press, 1998.

[18] Pearl J. Probablilistic reasoning in intelligent systems :networks of plausible inference. San Mateo CA :Morgan Kaufman Publishers, 1998.

[19] Pearl J. Graphical models for probabilistic and causal reasoning. The Computer Science and Engineering Handbook. Boca Raton, FL USA : CRC Press, 1997, Volume1.697 ~ 714.

[20] Cowell R G, Dawid AP, Lauritzen S L, et al. Probabilistic net2 works and expert

systems. New York : Springer, 1999.

[21] Jensen F V. An Introduction to Bayesian networks. New York : Springer, 1996.

[22] Jensen F V. Bayesian networks and decision graphs. New York : Springer, 2001.

[23] 李刚. 知识发现的图模型方法. 中国科学院软件所博士学位论文, 2001.

[24] Howard,R.A and Matheson,J.E.，”Influlence Diagrams”,The Principle and Application of Decision Analysis Vol,Howard,R.A.,and Matheson,J.E.(eds),strategic Decision Grooup ,Mento Park, (1979),719-762.

[25] David,A.P.,Conditional independence in statistical theory, Journal of the royal Statistical Society B43,（1979）,105-672.

[26] 赵雪雪. 面向单一搜索引擎的关键字广告竞价策略研究[D].哈尔滨:哈尔滨工业大学硕士学位论文,2013.

[27] 仰景岗. 在线关键字广告最优竞价策略效果及预算的影响研究[D]. 上海:上海交通大学管理学科硕士学位论文,2008.

[28] 欧阳波,贺赟.用户研究和用户体验设计[J].江苏大学学报（自然科学版）,2006,9（27）.

[29] 张文霖, 谁说菜鸟不会数据分析（入门篇）[M]. 北京：电子工业出版社,2013

[30] Cooper G F, Herskovits E.A Bayesian method for the induction of probabilistic

networks from data. Machine Learning, 1992.

[31] 薛万欣等. Bayesian 网中概率参数学习方法. 电子学报. 2003.11.

[32] BNT http://bnt.sourceforge.net/usage.html.

[33] Russel S, et al. Local learning in probabilistic network with hidden variables. Proc.Of the 14th Ijcai. Montreal. Canada, Morgan Kaufmann, 1995.

附录 A　贝叶斯网络 matlab 代码

```
%%完全客观变量的搜索开放平台广告点击率预测 matlab 代码
load'objectivedata.mat'
DAG = zeros(8);
DAG(1,7) = 7;
DAG([2 3 4 5 6 7],8) = 1; %%construct the DAG with matrix,it can be saw by input draw_graph(DAG)
obdata = cell(8,8921);
obdata(1,:) = num2cell(Keywords(1:8921));
obdata(2,:) = num2cell(Button(1:8921));
obdata(3,:) = num2cell(Pic(1:8921));
obdata(4,:) = num2cell(Input(1:8921));
obdata(5,:) = num2cell(Rank(1:8921));
obdata(6,:) = num2cell(Webrank(1:8921));
obdata(7,:) = num2cell(PV(1:8921));
obdata(8,:) = num2cell(CTR(1:8921)); %%Input all the data
nodesizes = [6 43 25 5 12 11 1 1] %%the size of nodes
disnodes = 1:6; %% discrete nodes
observednodes = 1:6; %%mark the nodes whose data can be observed from every case
bnet = mk_bnet(DAG,nodesizes,'discrete',disnodes,'observed','observednodes')
for i = 1:6
   bnet.CPD{i} = tabular_CPD(bnet,i)
end;
bnet.CPD{7} = gaussian_CPD(bnet,7);
bnet.CPD{8} = gaussian_CPD(bnet,8);
bnet2 = learn_params(bnet,obdata); %%learn from the history data
engine2 = jtree_inf_engine(bnet2); %%produce the induction engine
PV_predict = [];
CTR_predict = [];
```

```
    evidence = cell(8,1);
    for i = 1:1000
       evidence(1,1) = num2cell(Keywords(8921+i))
       evidence(2,1) = num2cell(Button(8921+i));
       evidence(3,1) = num2cell(Pic(8921+i));
       evidence(4,1) = num2cell(Input(8921+i));
       evidence(5,1) = num2cell(Rank(8921+i));
       evidence(6,1) = num2cell(Webrank(8921+i));
       [engine3,ll] = enter_evidence(engine2,evidence);
       target = marginal_nodes(engine3,7)
       PV_predict(i) = target.mu;
       target = marginal_nodes(engine3,8)
       CTR_predict(i) = target.mu;
    end;
    figure;
    subplot(3,1,1);
    plot(PV(8922:9921),'--rs','LineWidth',1,'MarkerEdgeColor','k','MarkerF
aceColor','g','MarkerSize',4);
    subplot(3,1,2);
    plot(PV_predict,'--rs','LineWidth',1,          'MarkerEdgeColor','k',
'MarkerFaceColor','g', 'MarkerSize',4);
    subplot(3,1,3);
    plot(PV_predict-PV(8922:9921),'--rs','LineWidth',1,MarkerEdgeColor','k
',MarkerFaceColor','g','MarkerSize',4);
    figure;
    subplot(3,1,1);
    plot(CTR(8922:9921),'--rs','LineWidth',1,'MarkerEdgeColor','k','Marker
FaceColor','g','MarkerSize',4);
    subplot(3,1,2);
    plot(CTR_predict,'--rs','LineWidth',1,          'MarkerEdgeColor','k',
'MarkerFaceColor','g', 'MarkerSize',4);
    subplot(3,1,3);
    plot(CTR_predict-CTR(8922:9921),'--rs','LineWidth',1,'MarkerEdgeColor'
,'k','MarkerFaceColor','g','MarkerSize',4);
```

```
%%引入主观变量用户体验评分（UE）的搜索开放平台广告点击率预测 matlab 代码
load'subjectivedata.mat'
DAG = zeros(9);
DAG(1,7) = 7;
DAG([2 3 4 5 6 7 8],9) = 1;
DAG([2 3 4],8) = 1; %%construct the DAG with matrix,it can be saw by input
draw_graph(DAG)
subdata = cell(9,300);
subdata(1,:) = num2cell(Keywords(1:300));
subdata(2,:) = num2cell(Button(1:300));
subdata(3,:) = num2cell(Pic(1:300));
subdata(4,:) = num2cell(Input(1:300));
subdata(5,:) = num2cell(Rank(1:300));
subdata(6,:) = num2cell(Webrank(1:300));
subdata(7,:) = num2cell(PV(1:300));
subdata(8,:) = num2cell(UE(1:300));
subdata(9,:) = num2cell(CTR(1:300)); %%Input all the data
nodesizes = [6 31 10 5 11 11 1 1 1] %%the size of nodes
disnodes = 1:6; %% discrete nodes
observednodes = 1:6; %%mark the nodes whose data can be observed from every
case
bnet                                                                    =
mk_bnet(DAG,nodesizes,'discrete',disnodes,'observed','observednodes')
for i = 1:6
   bnet.CPD{i} = tabular_CPD(bnet,i)
end;
bnet.CPD{7} = gaussian_CPD(bnet,7);
bnet.CPD{8} = gaussian_CPD(bnet,8);
bnet.CPD{9} = gaussian_CPD(bnet,9);
bnet2 = learn_params(bnet,subdata); %%learn from the history data
engine2 = jtree_inf_engine(bnet2); %%produce the induction engine
PV_predict = [];
```

```
    UE_predict = [];
    CTR_predict = [];
    evidence = cell(9,1);
    for i = 1:72
       evidence(1,1) = num2cell(Keywords(300+i))
       evidence(2,1) = num2cell(Button(300+i));
       evidence(3,1) = num2cell(Pic(300+i));
       evidence(4,1) = num2cell(Input(300+i));
       evidence(5,1) = num2cell(Rank(300+i));
       evidence(6,1) = num2cell(Webrank(300+i));
       [engine3,ll] = enter_evidence(engine2,evidence);
       target = marginal_nodes(engine3,7)
       PV_predict(i) = target.mu;
       target = marginal_nodes(engine3,8)
       UE_predict(i) = target.mu;
       target = marginal_nodes(engine3,9);
       CTR_predict(i) = target.mu;
    end;
    figure;
    subplot(2,1,1);
    plot(UE(301:372),'--rs','LineWidth',1,'MarkerEdgeColor','k','MarkerFac
eColor','g','MarkerSize',4);
    hold on;
    plot(UE_predict,'--rs','LineWidth',1,          'MarkerEdgeColor','k',
'MarkerFaceColor','r', 'MarkerSize',4);
    hold off;
    subplot(2,1,2);
    plot(UE_predict-UE(301:372),'--rs','LineWidth',1,MarkerEdgeColor','k',
MarkerFaceColor','g','MarkerSize',4);

    figure;
    subplot(2,1,1);
    plot(CTR(301:372),'--rs','LineWidth',1,'MarkerEdgeColor','k','MarkerFa
ceColor','g','MarkerSize',4);
```

```
    hold on;
    plot(CTR_predict,'--rs','LineWidth',1,          'MarkerEdgeColor','k',
'MarkerFaceColor','r', 'MarkerSize',4);
    hold off;
    subplot(2,1,2);
    plot(CTR_predict-CTR(301:372),'--rs','LineWidth',1,'MarkerEdgeColor','
k','MarkerFaceColor','g','MarkerSize',4);
```

致谢

犹记得自己在考研期间为成为一名合格的北大研究生做出的努力，然而时光飞逝，眨眼间已经又要到了研究生毕业的时节。内心忐忑又充满激动，忐忑在于我要离开学校走向社会，去面对一个我并未熟悉的环境。激动则在于，寒窗苦读十数载，终于到了用自己所学去为社会添砖加瓦的时刻。

衷心感谢我的导师李杰教授，在她的教导下，使我懂得了很多人生道理。我想导师之所以称为导师肯定不在于知识的传授，而在于人生的指导。知识可能几年就过时，但是人生哲理却受用一辈子。李杰导师不厌其烦地给我指导、答疑解惑以及鼓励创新、尊重学生个人选择的教育方法使我的研究课题得以深入进行并顺利完成。

感谢我的项目组长顾智博与我的直接导师宋旋婷，他们为我提供了研究条件，使我在项目研究实践中得到了锻炼和提高。他们对我的论文从选题、撰写到定稿都给予我悉心指导和教诲。同时一年多来言传身教，使我对人生和事业有了更深的感悟。

正是以上老师的培养和教诲，使我对研究问题的思路不断拓展、研究问题的方法更科学，我的理论水平、知识能力和综合素质才得以不断提高。

最后，感谢北大软件学院为我提供的学习环境。正是因为要先实习后毕业的毕业方式，使我不再恐惧毕业，不再担心“一毕业就失业”。至今为止，我听说了无数个师兄师姐在各个领域崭露头角，我听说了无数的传说。到今天，我想终于到了该我出去缔造新的传说的时候了，未来将由我们新的一届来打造，请相信我们！

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：　　导师签名：

日期：　　年　　月　　日