明略讲堂 | 清华马少平教授详解 “人工智能能做什么？”

原创 2017-08-03 马少平 明略数据

前一段时间，马少平老师在高铁上写了一篇”人工智能能做什么？”的博客，结合自己多年的实践给出了自己对于现在人工智能如此火热现象的思考。马老师自己也说，这一篇新浪博客是随笔，只是把当时的想法写出来了。

7月下旬，我们有幸请到了马老师为明略数据内部做了一个分享，基于这篇随笔的一些思考进行扩展，马老师结合多年经验，为我们分享了一堂别开生面的人工智能讲堂。



AlphaGo的出现，让很多人对人工智能产生了不切合实际的幻想，究竟人工智能能做什么？1个半小时的分享，整理成如下内容，包含以下几个部分：

- 人工智能发展历史
- 人工智能典型应用
- 当前行业中人工智能的应用情况
- 如何指导科研工作



作者：马少平

清华大学计算机系教授，博士生导师，中国人工智能学会副理事长，中国中文信息学会副理事长。主要研究方向为智能信息处理，包括文本信息检索、网络用户行为分析、个性化推荐、社交媒体分析等。

本文经马少平教授授权，如需转载请联系后台。

以下为讲座文字整理版

（全文共9100字，阅读时间约15分钟。）



各位朋友下午好，非常高兴，今天可以一起交流。近三十年来我一直从事人工智能有关的工作，，近十几年来，我研究的比较“专”，主要从事搜索引擎相关的工作。今天我想跟大家谈一谈人工智能能够做些什么。

其实早在八十年代，国外就有一本书讨论了计算机能做什么，不能做什么。我今天也想和大家讨论一下类似的问题，首先一定要知道，有些事情人工智能技术今天不能做，未来也许不是问题。但是还有些题目今天的人工智能难以完成，从技术上来看，未来的人工智能可能也无法解决。今天和大家谈一下如今的人工智能到底能够做到哪些？

一、人工智能发展历史



我们先简单地回顾一下人工智能的发展历史。人工智能的发展大概可以被分为四个阶段。

人工智能第一阶段--前期阶段

关键词：通用问题求解 定理证明 游戏 机器翻译等

1956年在达特矛斯会议上，人工智能的概念被首次提出来。到去年已经是整整60年了。当时的概念已经不是凭空提出的了，在1950年，图灵就已经阐述过图灵测试了。那个时候第一台电子计算机已经问世十年了，所有人都期望用计算机为工具去实现很多人工智能的设想。

在当时达特矛斯会议上，大多数与会者都是二十几岁的年轻人，他们对实现人工智能这一想法非常的乐观，也很有自信。在人工智能提出的早期阶段，就有人研究“**通用问题求解**”，简称GPS。当时的设想是通过一个通用的方法去求解所有的问题。除此之外，在那个时代，人们用当时的计算机做了很多的数学定理证明，甚至是证出了《数学原理》一书上的所有定理。这在当时的条件下是十分不易的。还有很多人研究游戏，因为游戏是可以反映人类智能的。当时就有很多人研究棋类游戏。在1956年达特矛斯会议上就有人演示了机器下棋的过程。

还有一个方面是**机器翻译**，在当时人们认为有了计算机这个强大的工具，在后台存储一本庞大的电子词典，就可以解决自然语言的翻译问题了。但是，这些先驱者们很快就陷入了困境，发现这件事情不是那么简单。他们失败的关键点就在于“知识”。比如说我们要翻译一本专业性很强的人工智能方面的书，把它交给一个没有科技翻译经验的、学语言的学生，他肯定会翻译得漏洞百出，即使他的外语和汉

语能力都没有问题。为什么呢？因为我们在做这方面的工作的时候，一定要先掌握相关的“知识”。**人工智能技术其实也是这样，要实现“智能”，必须要依靠“知识”。**

人工智能第二阶段——知识处理时代

关键词：知识工程 专家系统

当人们意识到了这点的时候，人工智能就发展到了第二个时代——知识处理时代。这个时代主要的特征就是知识工程和专家系统。专家系统是先被提出来的概念，然后才是知识工程。专家的特点就是掌握相关领域的知识，如果知识能够被总结出来，那么我们就可以用计算机来替代这些专家，去解决相关领域的问题了。继专家系统之后，又提出知识工程，就是说我专家之所以是专家，就是因为他们掌握了这方面的知识，如果我把他的知识给总结出来，那么我就可以用计算机来代替专家来解决这方面的问题。

当时造了很多专家系统，最著名的大概就是六十年代中期MIT做过一个用于做血液病的诊断的专家系统MYCIN，这之后还有很多专家系统问世。虽然在这个过程中人们认识到了知识的重要性，但是知识获取始终是一个难题。当时虽然机器学习的概念也被提了出来，有基于归纳法的学习、基于解释的学习、基于演绎的学习等等各种理念，但是都没能取得成功。所以知识获取仍然是一个很大的瓶颈。

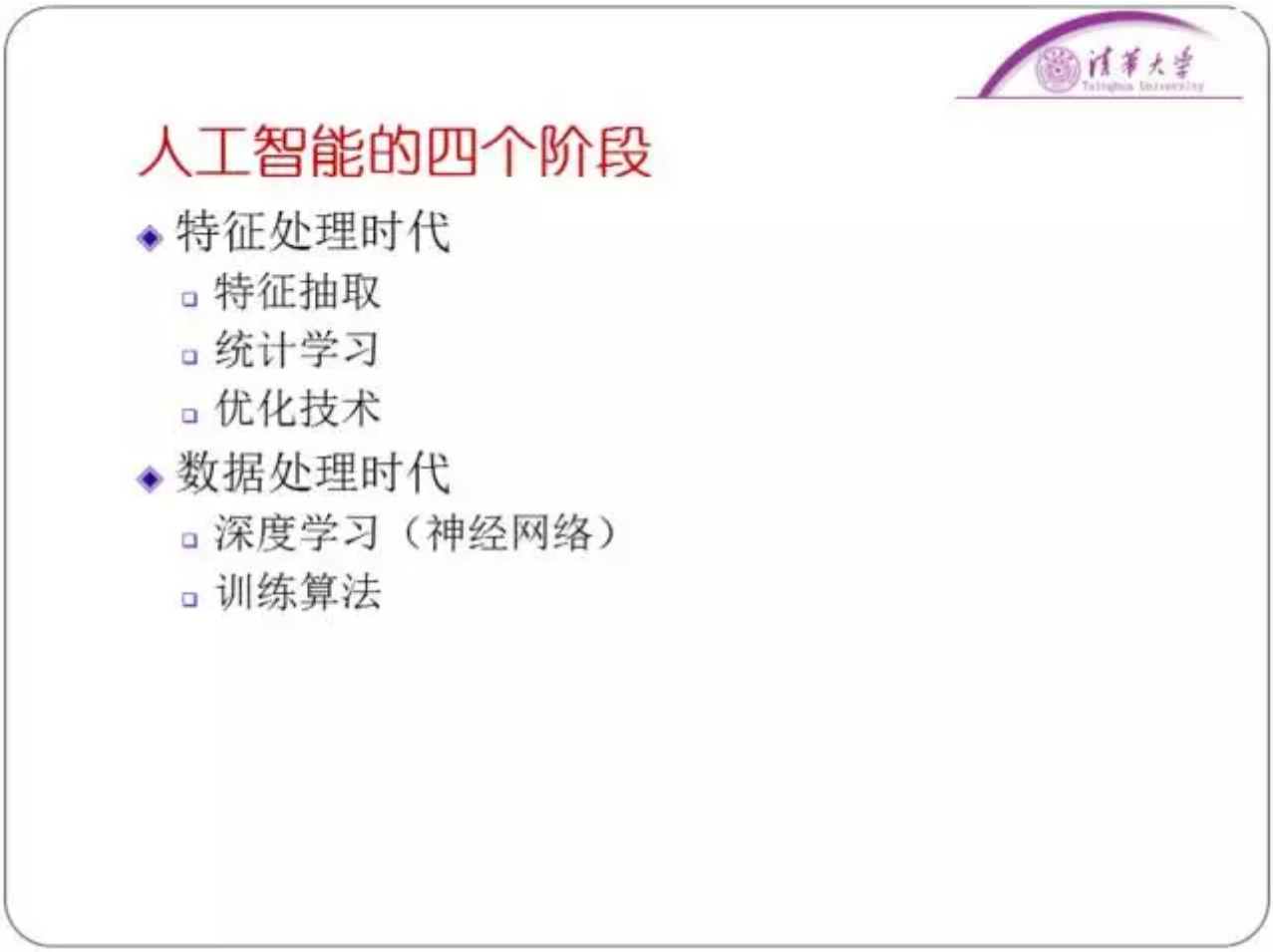
当时，也就是在大概80年代的时候，我也曾经参加制作过几个专家系统。那是非常幸苦的工作，为了能和专家没有障碍的交流，必须要学习和研究专业领域的专业知识，否则根本连专家在说什么也不知道。但是，即使是这样，这种获取知识的方式仍然是非常艰难的，因为很多知识是很难提炼和归纳的。比如说我不会骑自行车，一上去就会倒。我问会骑自行车的怎样骑车不会倒。他肯定告诉我这种事情没法口头上教，只能找个没事的礼拜天扶着练练才能会。这也就说明了人工去做知识的提取不光困难，而且效率很低。因此当年虽然做了很多的专家系统，但是真的成功应用的却并不多。

人工智能第三阶段——特征处理时代

关键词：特征抽取 统计学习 优化技术

时间进入了90年代末期，这是一个特征处理的时代。其实这个时代最主要的就是统计学习，试图用统计机器学习的方法来让机器自动地学习。不过机器学习的材料并不是原始数据，是我们从数据抽取出来的特征。比方说那个时候如果说要识别一个猫，那可能就得找各种各样的特征，想做语音识别，也需要各种各样的特征，然后在得到特征之后再用统计学习方法进行处理。

这个过程中难点其实更多的是找特征。比如那个时候做汉字识别，大家用的统计学习方法其实都差不多，关键在于怎么找特征，到底哪些特征才能把这汉字给描述出来，并且计算机还能处理。比如“横竖撇捺”是汉字的特征，我们人脑就是这么识别汉字的，但是那个时候计算机抽取不出来。所以怎样找特征才是一个真正的难点。



人工智能第四阶段——数据处理时代

关键词：深度学习（神经网络）训练算法

我们现在处在一个数据处理时代，我们应用各种数据进行深度学习。不过这些数据不再是我们抽取的特征了，而是原始的数据。让机器自己从原始数据中进行学习。做语音识别的，只要把语音的采集信号交给机器就好。做图片识别的，只要给机器图像，让它自己去判断，也就不用抽取特征了。所以这个时代在技术上已经是更加进步的了。

这是人工智能的一个进阶，从知识到特征再到数据处理，人的参与越来越少。在专家系统时代一定要专家级的人物参与才行。现在数据时代，从我们准备的学习对象、处理对象来说，把原始数据交给系统就行了。这个程度上来说，人工智能是一点点进步的，人的干预程度了越来越少了。



二、人工智能的典型应用

应用场景1: 对机器和人类都很容易

现在的人工智能的一些典型的应用可能大家也都是有一些了解的。比如说，现在我们这个阶段深度学习领域，做的比较好的像是语音识别，比如2011年的微软，首先将深度学习应用到语音识别中，一下把错误率拉低了30%。2016年谷歌也是这样，它首先做了一个很完整的系统，就是基于神经网络的机器翻译。还有一个就是名声大噪的围棋人工智能AlphaGo，在比赛中连续战胜顶尖高手李世石、柯洁。这个里边一个很重要的东西都是就所谓的深度学习，也就是说神经网络模型，这些是比较成功的应用。

应用场景2：对机器和人类都很难

我们再看一下比较难求解的领域，人工智能哪些事情做不了。比如说让人工智能提出一个新的概念，或者是创立一个新的科学体系。或者是在数学领域开创一个新的分支等等，这些事情对于计算机或者是人工智能而言都是无法完成的。至少现在还没有办法去处理。上面说到的几件事情不仅仅是对于计算机，对于我们人类其实也是比较难以完成的。

应用场景3：对人类很容易，对机器很难

那么还有一些事情，对于我们而言是非常容易的，可是人工智能处理起来依然比较困难。比如说理解幽默、比如联想能力、比如说漫画识别等等。这些方面上至老人下至三岁小孩，我们的水平都比人工智能高很多。比如我举一个例子，这是一个真实的事情，有一天我买东西，下了一个单，客服说一杯咖啡后，就会有人与我联系，害得我赶紧下楼买了一杯咖啡，很显然我做了一件十分搞笑的事情，我们正常人都能明白，但是机器就不知道幽默点在哪里。



这张图片，大家的第一反应就是一个头发花白的老人。可是如果我提示你，图片里面有字。我们都能很快地找出“长命百岁”来。



还有就是漫画识别，我们从来没有见过这张漫画。但是让我们认识陈佩斯。我从来没看过它的漫画，第一眼看见，我也知道它画的是是陈佩斯。但是机器怎样通过这张漫画识别呢？



这张画，显然和真实的陈佩斯还是差了很多的。这个方面也有很多人在研究，但是目前的水平还不尽如意。

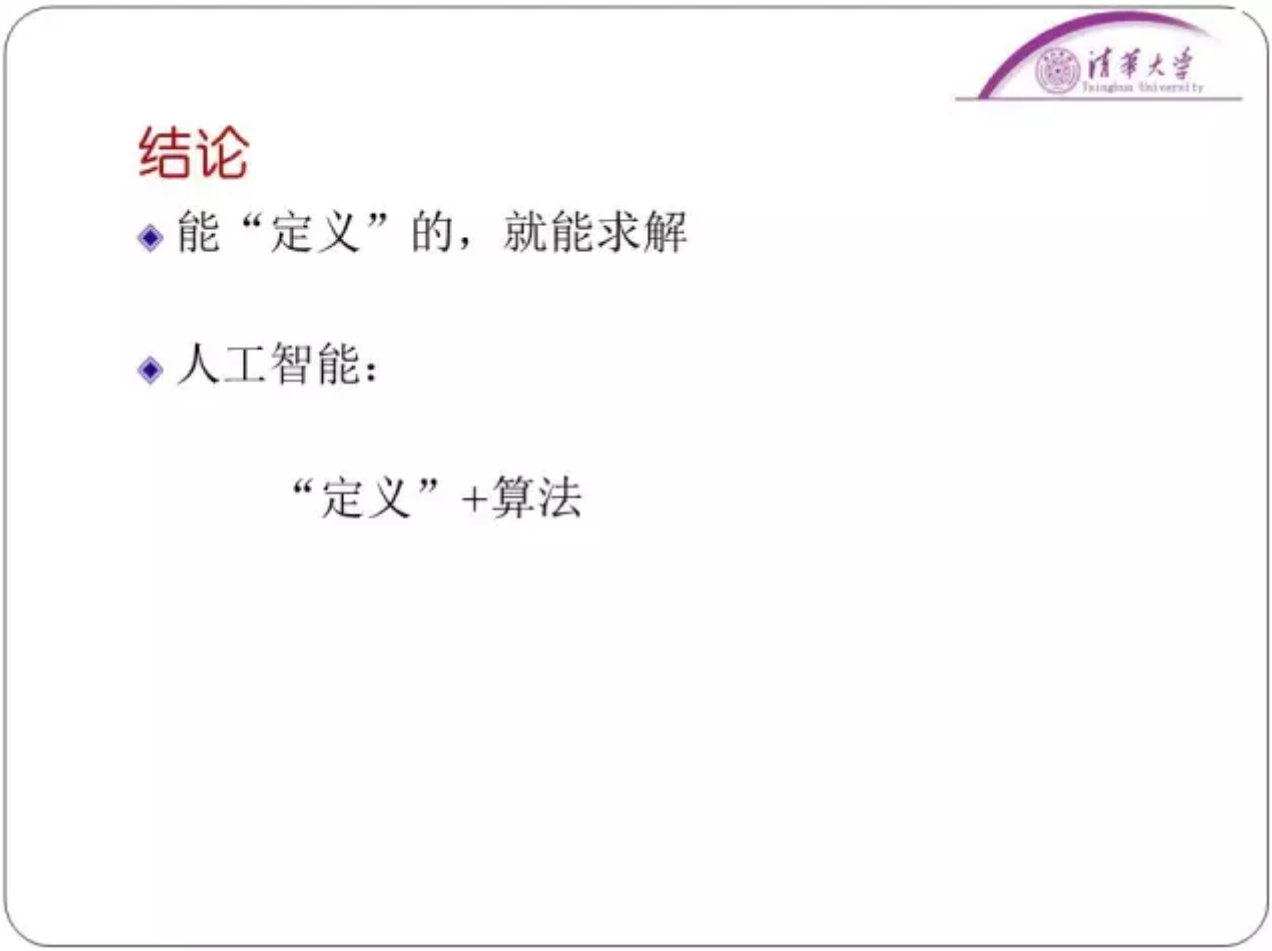
所以为什么有的方面人工智能的水平能够轻松超过人类，但是有一些人类看来很简单的事情，机器却难以完成呢？我们来分析一下。像刚刚讲过的语音识别这些技术，我认为它之所以在人工智能领域里面能做，是因为“说得清”。这个说得清并不是指是别的材料说话很清楚，或者是通过什么数学公式能给它描述得清。这个“说得清”的意思是能够明确地定义。这里所说的定义，并不一定指形式化的定义，可以是举例、可以是假设等。比如说我们做猫狗等动物的识别，可以给出大量的动物照片，说这些是猫，这些是狗。可以很明确的用数据给出动物的“定义”。但是让机器理解幽默就很难了，我们很难去说清楚为什么这句话是幽默，幽默点在哪里。



结论：机器能做“说得清”的事情，不能做“说不清”的事情。

人工智能就是通过“定义”+算法，来求解问题。

这里面对于人工智能领域提炼出来一个原则就是，能说清楚的，能做；说不清楚的，至少现在还做不好。但是技术是在发展的。人工智能的发展就是由“说不清”变得“说得清”，也就涉及到了描述的问题和相应的算法求解的发展。比如说围棋，过去没找到方法，“说不清”，但是谷歌突然间有了这种方法，把围棋说得清了，计算机围棋问题也就解决了。也就是说能“定义”的就能求解。这个定义不是说简单的一个文字，一个形式化的东西。**现在的人工智能就是一个“定义”再加上算法来求解问题。**那么这些就是目前我们人工智能能够做的。



首先是能够定义，当然这个定义一定是以计算机能够理解的方式，并且有可以求解的算法。可以是形式化的定义比如说是什么定理，或特征定义，也可以是数据定义或者假设定义。比如这个东西很复杂，要限定条件，在假设下去求解。如何下定义，这是一件和复杂的事情。

三、当前行业中人工智能的应用

1. “猫”脸识别

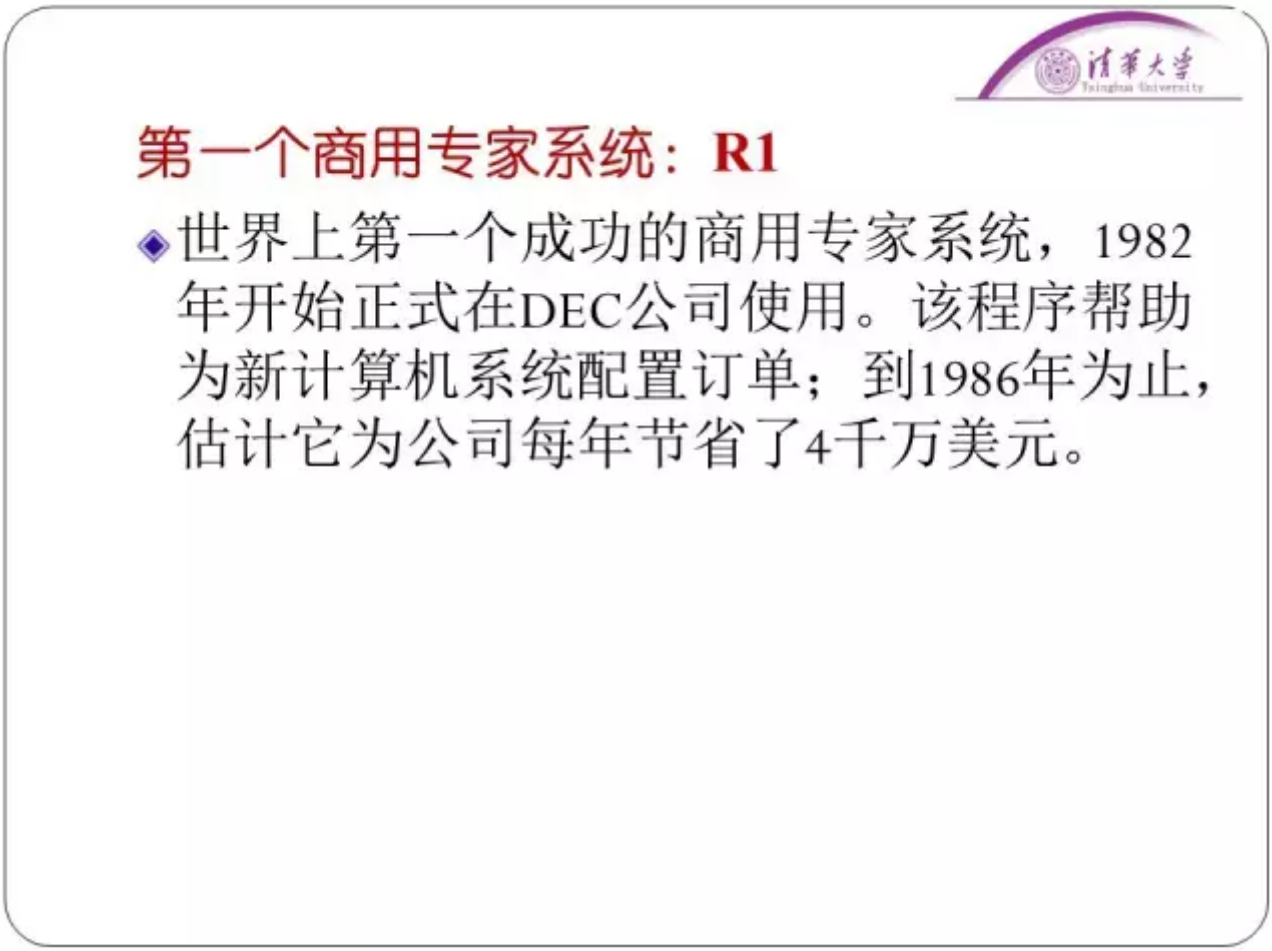
我们结合具体的例子来看，比如我们做识别猫的应用，那么什么是猫？我们查了一下搜狗百科："头圆、颜面部短，前肢五指，后肢四趾，趾端具锐利而弯曲的爪，爪能伸缩。夜行性。以伏击的方式猎捕其他动物，大多能攀缘上树。"，这个东西我们人能看懂，但是，我们自己识别猫的时候肯定不会按照这个去做。可能我们看到一个稀有的品种，或者是研究猫的专家会去研究这个定义。但是我们让计算机去做，这是不行的，按照这个定义来，肯定识别不出猫。那我们现在怎么定义？给些图片就行了，这就是现在大家做的。我给一大堆图片，告诉系统类似于这个的都是猫，类似于这些的肯定不是猫。这就是我们下的一个定义。加上机器学习算法，可能就能识别猫了。



2. 专家系统

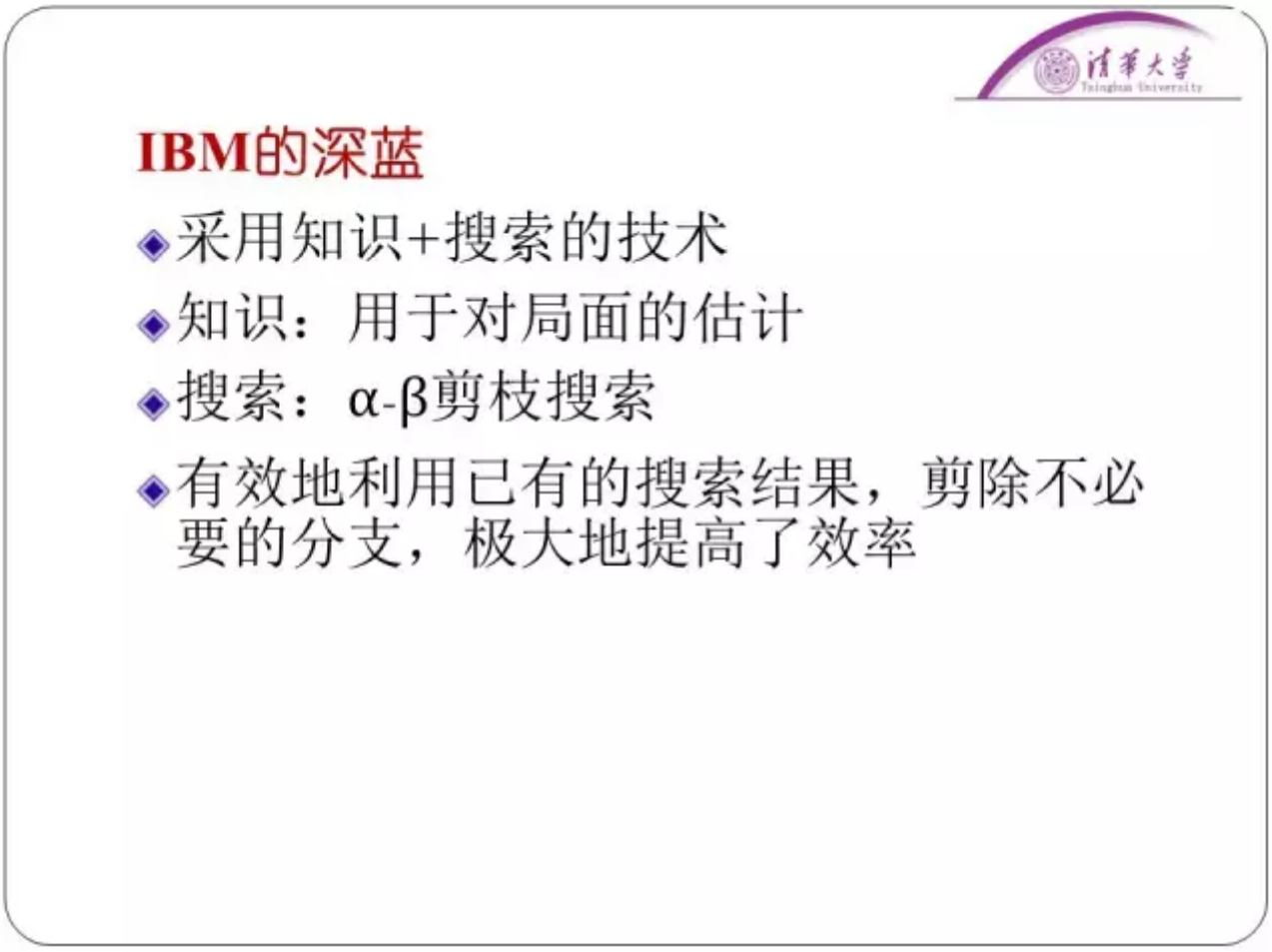
比如说专家系统中，实际中应用专家系统并不多，这里面有各种各样的原因，比方说、诊断血液病系统实际上就涉及到一个法律上的问题，一旦出了医疗事故，责任如何划分。还有我们国内有人总结了中医治肝病专家的经验，开发了一套中医的专家系统，但是这个也是有问题的，比如说中医讲号脉，但是当时号脉没有这样的机器。就必须把脉相手动的输入到系统中。普通人又不会号脉，还得找个号脉好的中医，人家号脉好的中医又用不到这样的系统。由于种种原因，这个系统就失败了。

还有比如世界上第一个成功的商用专家系统R1。82年开始在DEC公司使用，是用来代替向厂家订货的工作的。当时据说可以节省4000万美元一年，它通过几千条的规则，覆盖了所有可能出现的情况。



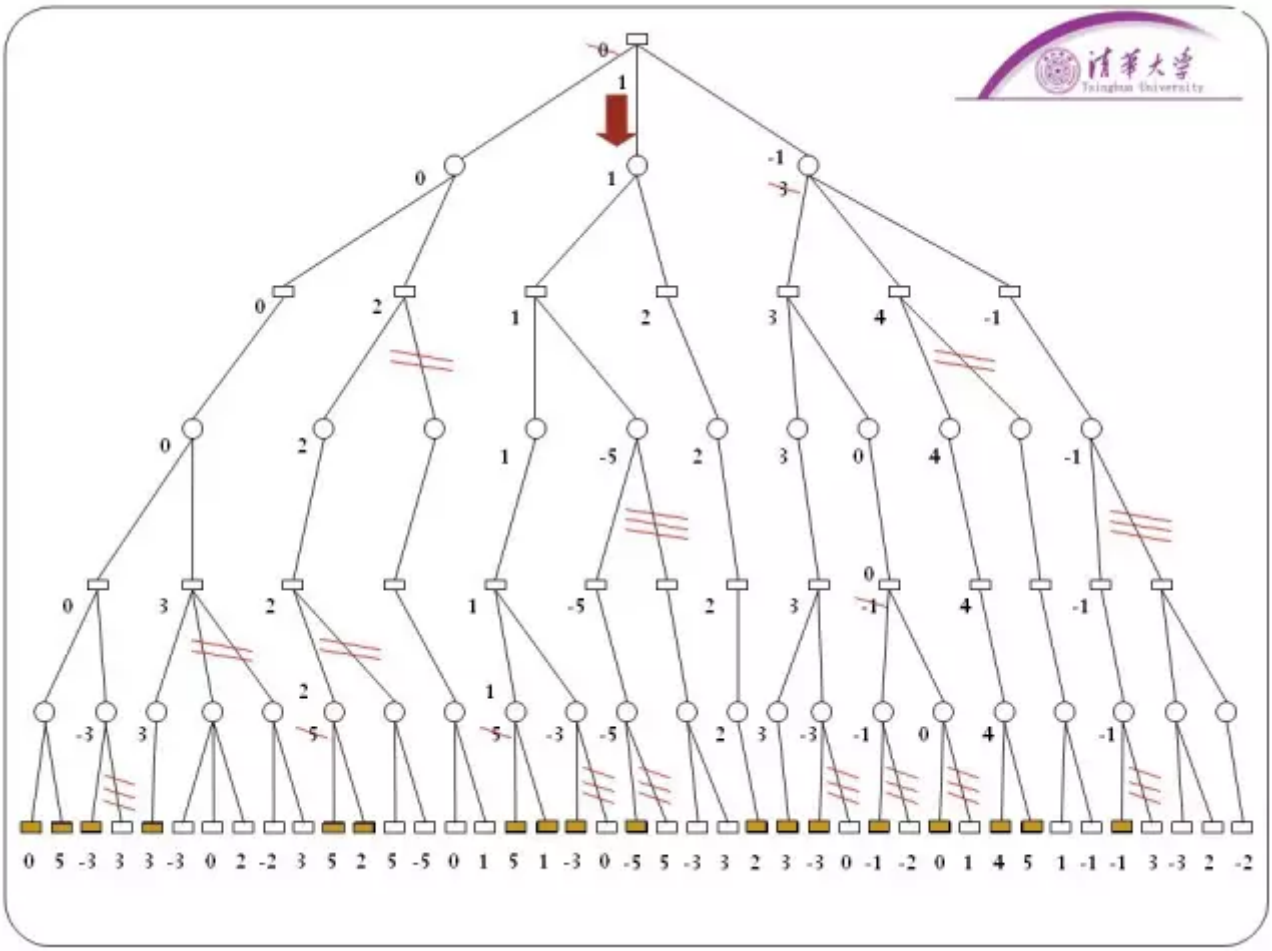
3. IBM深蓝系统

再有一个是IBM的深蓝系统，1997年，在国际象棋比赛当中战胜了国际象棋大师卡斯帕罗夫引起了世界的轰动。其实深蓝系统的主要算法叫做“alpha-beta剪枝算法”，这个算法很早被提出来了，但是直到深蓝系统才获得了很大的成功。这个算法依赖于局面的评估，在当前的局面下是对我有利还是对对方有利。IBM聘请了很多的象棋大师来为他总结经验，用于对局的局面评估。运用这个算法一方面能剪掉不必要的分支，提高系统的效率，另一方面，它利用知识把局面的评估定义的非常清楚，到底什么样的局面才是对我方有利的。



那为什么这种方法在20年前就达到了让计算机国际象棋达到了世界大师的水平？就是因为无论是中国象棋、国际象棋都有这样的特点，它相对来说容易说得清楚。比方说，残局阶段，比对方多个子就可以认为赢是没问题的，或者说马的位置可能就决定了谁胜负，然后再加上它很强的这个搜索能力，搜索深度可以达到十几步。

由于象棋这种易于“说清”的特殊性，在二十多年前计算机就打败了人类的顶尖选手，国内也是，十几年前计算机也打败了国内的顶尖高手。但是看到深蓝在很多年就获得了成功之后，很多人对象棋产生了误解，认为象棋的可能情况不多，通过暴力穷举就能解决。但其实不是，深蓝的成功很大一部分也在于“Alpha-Beta剪枝”算法的应用。



95年主持IBM深蓝项目的工程师到清华做过讲座。当时我就问过他剪枝算法相比暴力穷举的效率到底提高了多少。他们做过测算，如果单纯是穷举的话，要达到深蓝的水平，每下一步棋需要十七年，用了剪枝算法之后，只需要几分钟。所以说除了“说清楚”外，算法也是很重要的。

4. AlphaGo

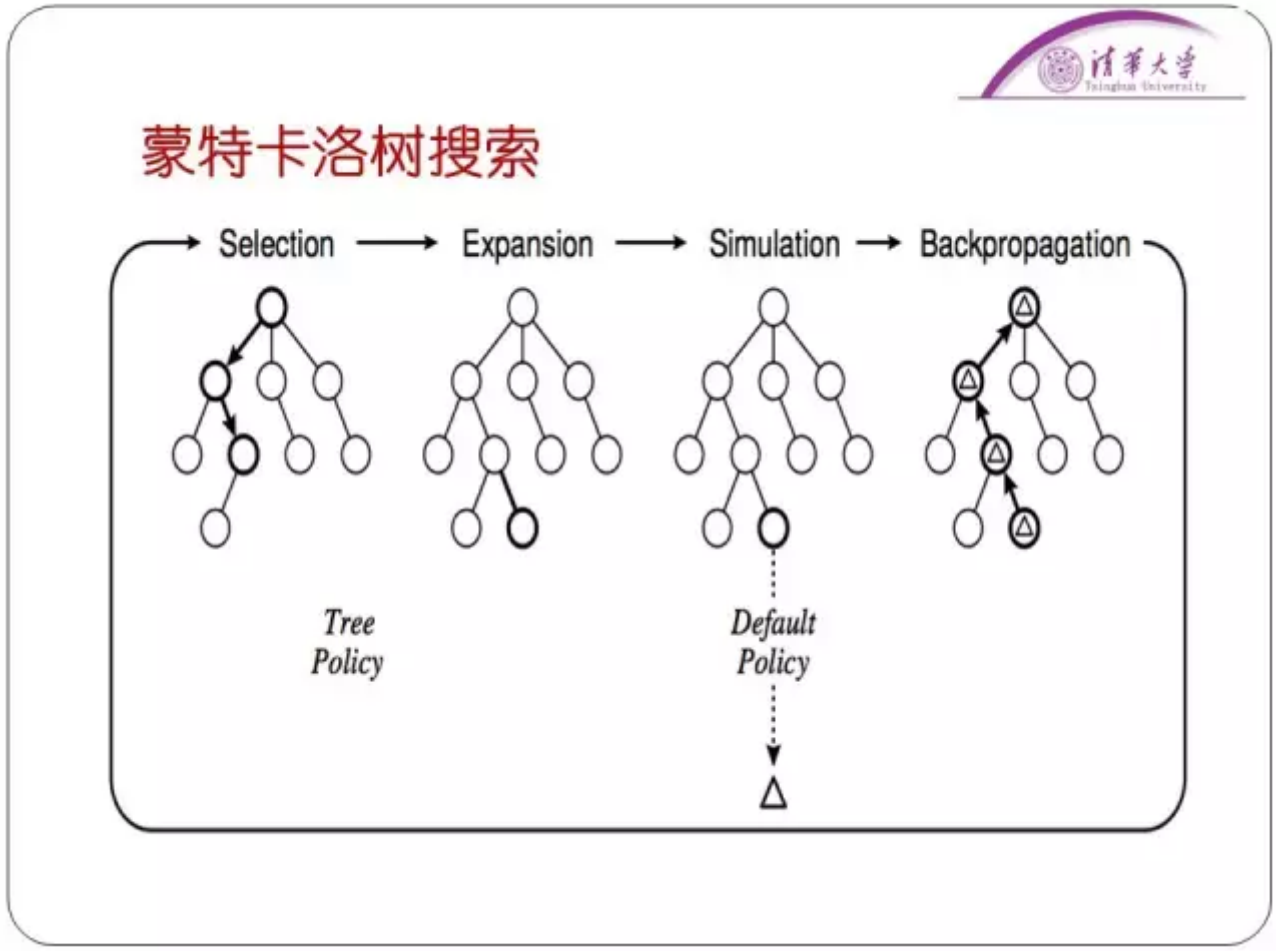
那为什么AlphaGo要到20年后才成功了呢？

其实即便是20年后计算机围棋就取得了成功，也是出乎很多人的意料，很多人没有想到2016年人工智能就可以达到这样的水平了。为什么Alpha-Beta剪枝算法用在围棋上不行呢？是因为“Alpha-Beta剪枝”算法严重依赖于局面的评估。原有的局面评估方式是建立在总结专家经验的基础上的。偏偏围棋是很靠感觉的，几个不同的围棋棋手之间往往很难有一致的看法。在几年前中、日、韩曾经举办过围棋组团对抗赛这种友谊比赛，每个队伍选出三名棋手组成团队一起下棋。结果韩国队配合最好。据说他们的队伍里一个人只负责摆棋，一个人只负责买饭准备饮料，只有一个人在下棋。而中国、日本队那边往往还没怎么下呢自己就吵起来了。所以说围棋是很难总结专家经验的，这也就是应用旧的“Alpha-Beta”剪枝算法的计算机围棋，一直处于一种水平不高的状态。

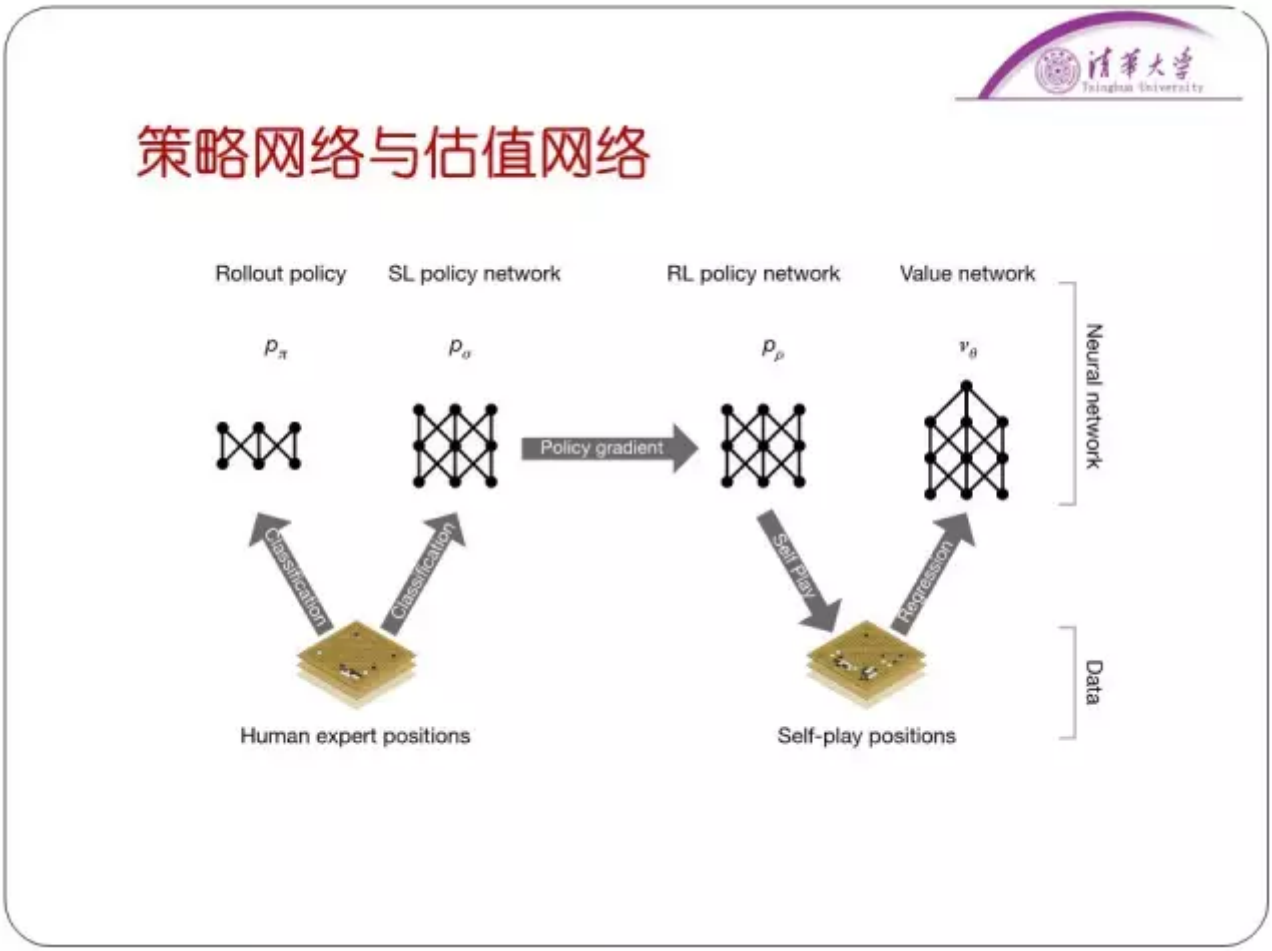
计算机围棋的第一个里程碑在于蒙特卡洛树搜索的引入。2006年法国的一个团队首先把蒙特卡洛树搜素的方法引入到了计算机围棋中，这种方法也是为了解决局面评估的问题。

蒙特卡洛模拟本质上是一种随机模拟的办法。是什么意思呢？就是说该计算机下了，它就随机下，随便找一个点下，一直随便下直到能判断出一方胜利为止。然后再这个基础上反复模拟十万次、一百万次然后看哪个点的胜率最高。哪个点胜率最高，那么计算机下次就下到哪个位置。后来改进的蒙特卡洛树，每次模拟的结果可以重复利用，这就提高了搜索的效率。那么就是靠每一个可下棋的位置的胜率来解决局面评估的问题。

同时在蒙特卡洛树搜索中，采用了一种叫做“信心上限”的决策方法，也就是“多臂老虎机”决策问题。每个老虎机有一定的概率赢钱，多个老虎机，不同的老虎机赢钱的概率不一样。所以我可以站在旁边先看别人玩，先找到赢钱概率最大的决策方法。在模拟选择分支的时候，就把每一条路径看做一个老虎机。这之后选择最优路径然后继续随机模拟。模拟一轮后可能有胜有负，看胜率多少，然后把胜率最高的步骤传回。



应用了这种方式，计算机围棋可以说是从原来基本不会下棋的状态进到了业余五、六段的水平。谷歌的AlphaGo是怎么解决接下来的问题的呢？单纯是原来蒙特卡洛树的随机模拟是效率很低的。所以谷歌就把深度学习的技术引入进来用于解决搜索量大的问题。在AlphaGo中，采用策略网络减少了模拟的宽度，采用估值网络减少了模拟的深度，在整个模拟阶段不一定从头模拟到尾，这样的话可以提高模拟的效率。实际上围棋就不是靠知识了，是完全靠数据去定义了，通过人类的16万棋谱和AlphaGo自己和自己下的三千多万盘棋，再加上深度学习，就定义了下得好不好这件事。并且在算法中结合了蒙特卡洛树、深度学习等技术，把算法本身的效率提高。

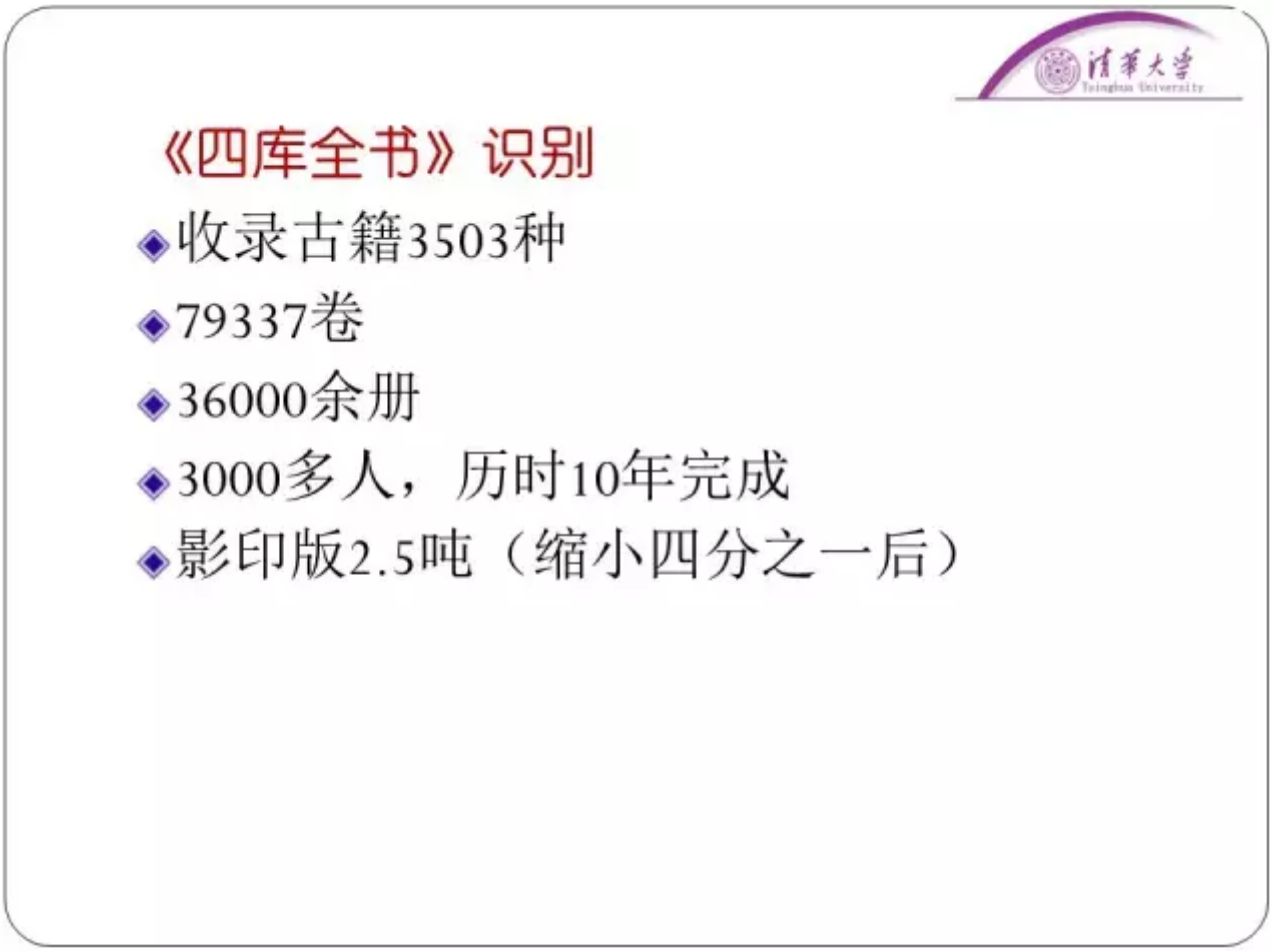


5. 汉字识别

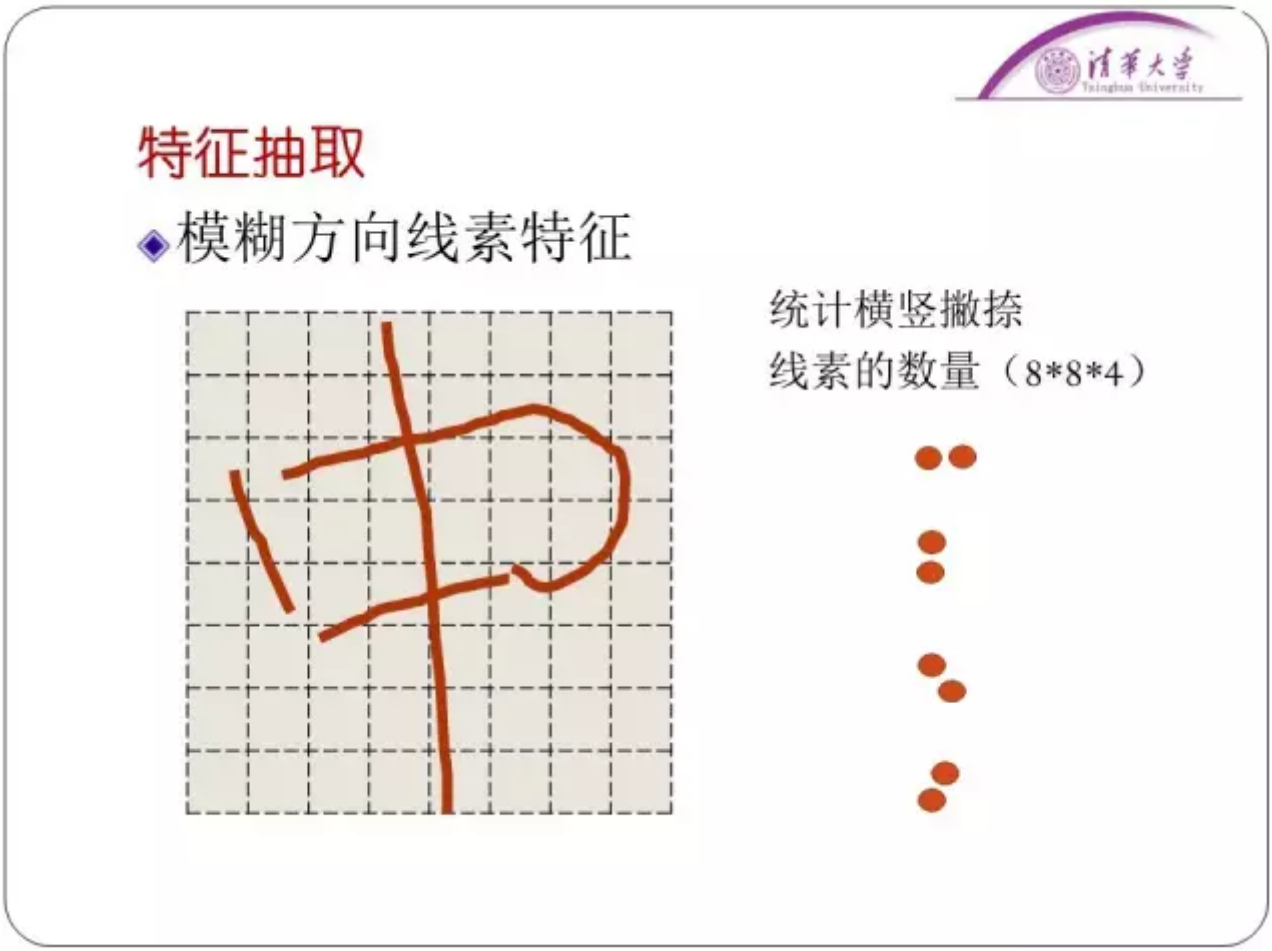
再举一个我自己做的汉字识别的例子，20多年前，九十年代的时候，当时我们做的汉字识别属于脱机汉字识别，也就是把汉字写在纸上然后扫描出来识别。联机汉字识别，比如直接在手机上写字识别相对比较简单，因为可以记录笔画顺序。但是脱机的识别就比较难了。当时我们需要把《四库全书》做成成电子版。



这个《四库全书》是当年三千多人一起抄写了十年才完成的。现存的完整版分别保存在国图和台北。但是要查询《四库全书》的内容就非常难了，当时台湾出版过影印版，整体缩小四分之一，最后一套书的重量有两吨半。因此很多人想出版成电子版。90年代的时候很多公司抢着想做这件事情。甚至还有公司在人民大会堂抢先开了新闻发布会，当然他们最后也没做成。这些人都想找人采用人工录入的方式，但是都没有成功。



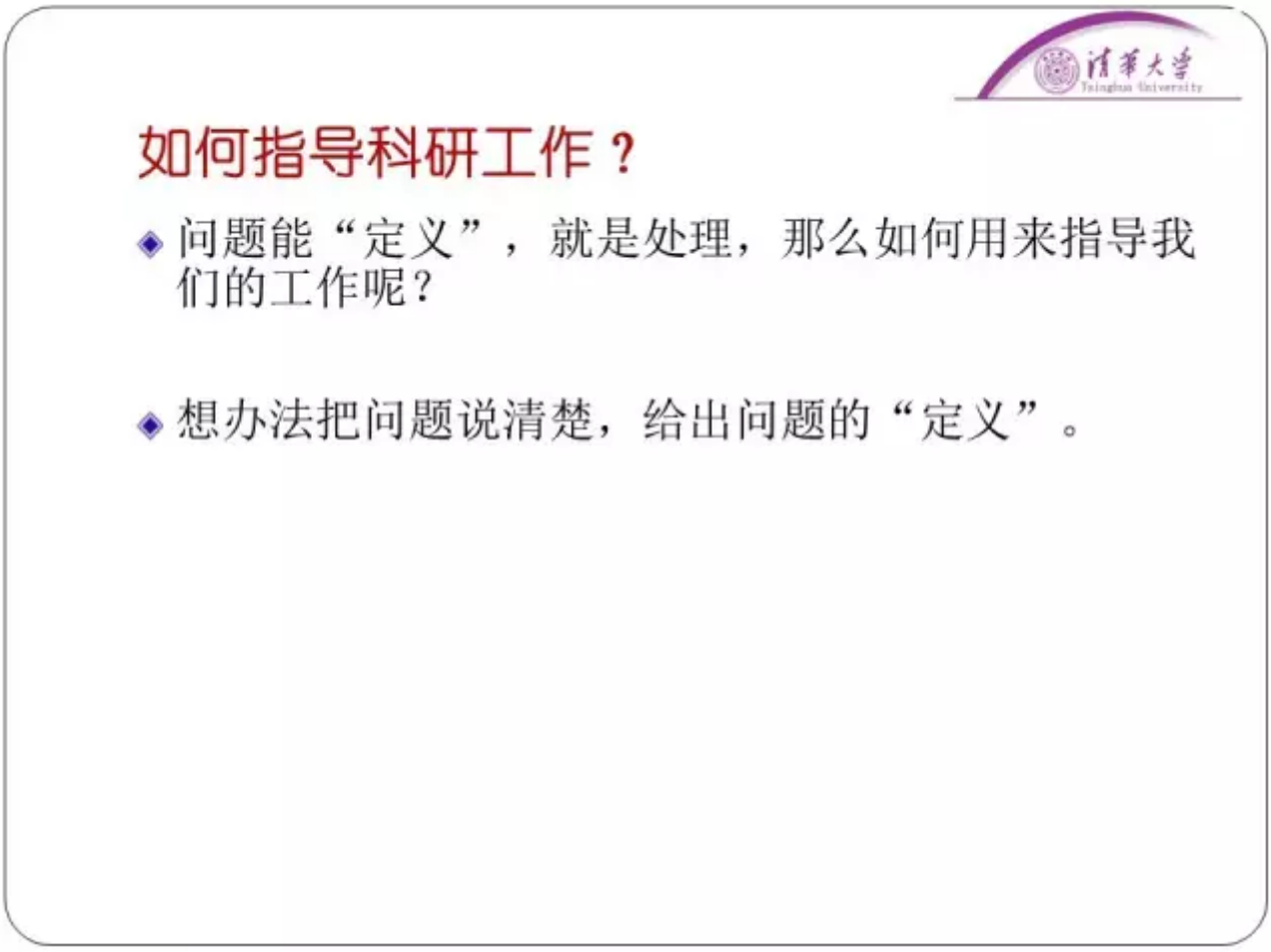
后来我们采用了一种“滚雪球”的算法让机器学习前面人工录入的内容，然后慢慢识别后续的文字。在识别的技术上我们用了模糊文字方向线素特征。这个描述的方法是把每个汉字归一化之后，然后变成网络，每一个网格里面统计字的边缘的像素，抽取不同方向的像素特征。一个汉字就是256个特征。



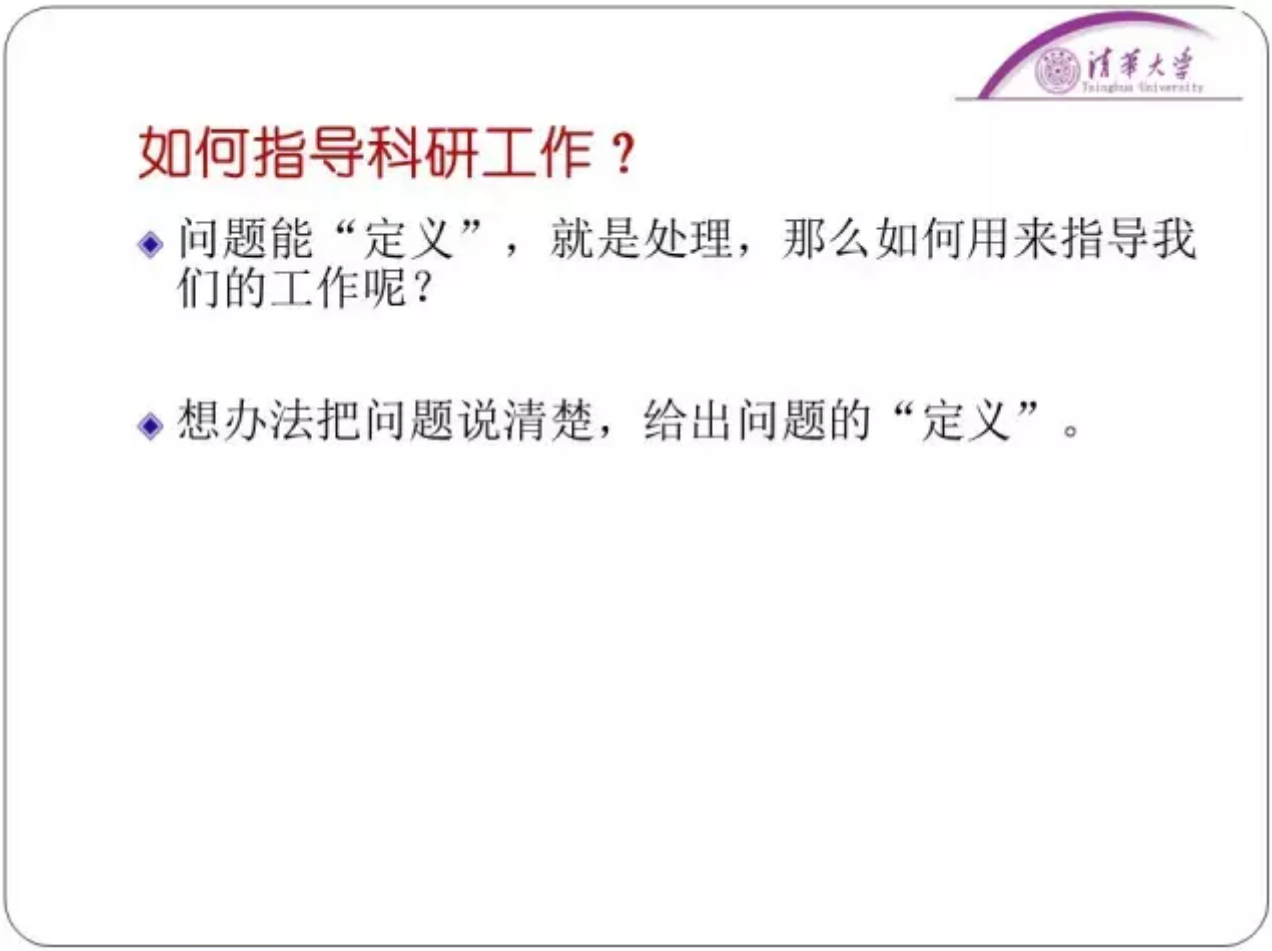
我们当时的正确率能够达到95%，后来再通过人工的一些方法，把错误率下降到了万分之一的水平。这里我们等于做的也是一个对于汉字的描述。我们在做这些项目的时候，**首要就是要想办法把问题说清楚、给出它的定义。**

四、如何指导科研工作

上面分享了一些行业中的成功实践，这些实践如何知道我们的科研工作。要让机器理解人类希望他们处理的“问题”，还是要回到如何把问题说清楚，给出问题的“定义”。



下面我结合自己的三个例子说一下。比如我们曾经做过一个垃圾网页识别的项目。


清華大學
Tsinghua University
如何指导科研工作？
问题能“定义”，就是处理，那么如何用来指导我们的工作呢？
想办法把问题说清楚，给出问题的“定义”。

1. 垃圾网页识别

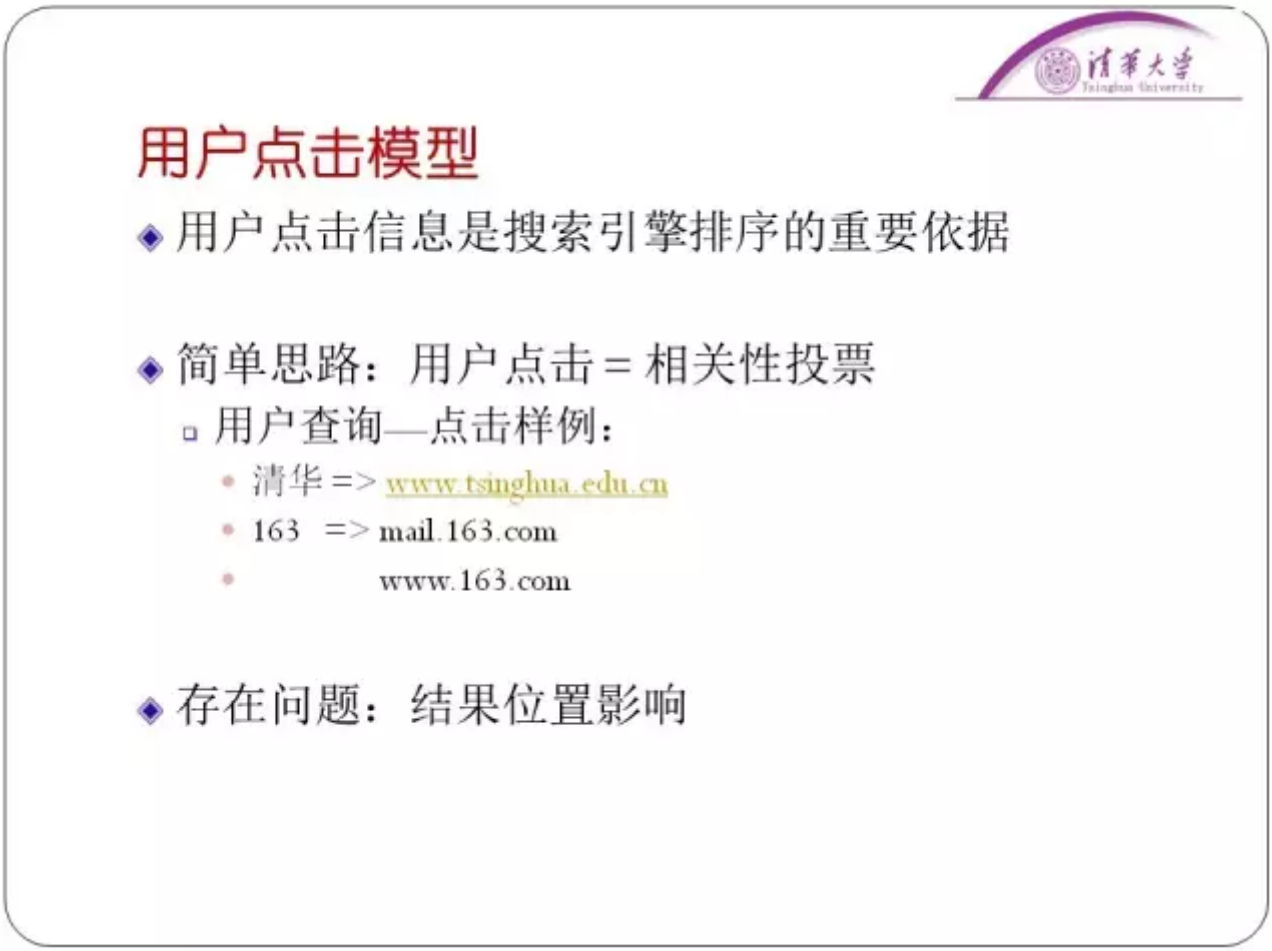


垃圾网页指的就是通过欺骗搜索引擎，用户搜索的时候，即使是搜的与这个垃圾网页没有关系的词，它也会通过欺骗搜索引擎，把自己的网页排到前面去。非常影响用户体验。

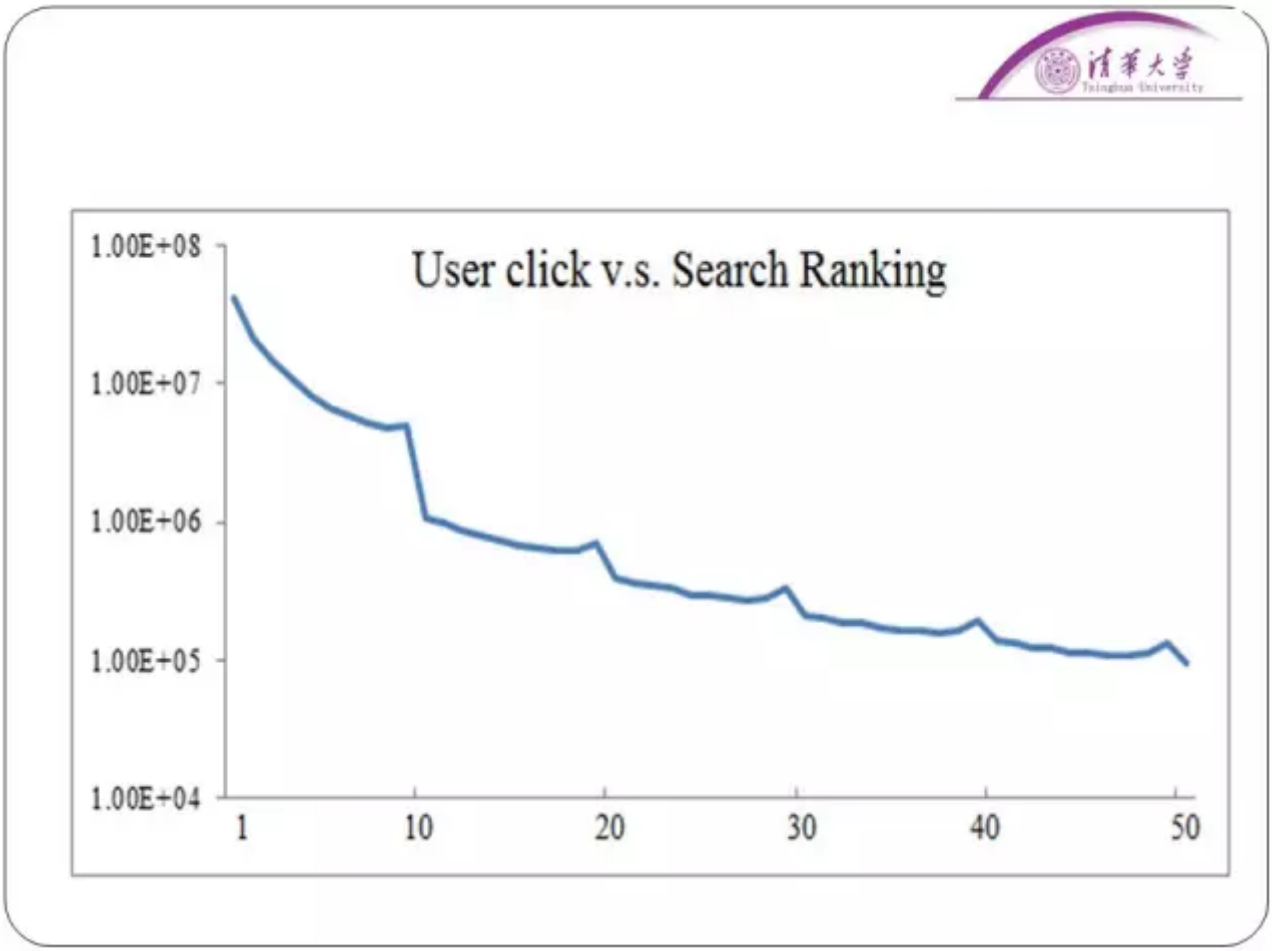
但是识别这种垃圾网站非常难。因为作弊手段层出不穷，只能“一事一议”的进行处理。比如发现了一种作弊方法，我们提取了规则，说清了这个垃圾网站的特征，可能它稍微修改一下，我们就识别不到了。而且这种方法很慢，也很没有效率，很多时候用户点开了垃圾网站就直接关掉了，根本懒得反馈。后来我们通过了另一种思路，就是分析用户打开网页之后的行为。如果是秒关，那很有可能这是一个与搜索内容无关的垃圾网站。或者某个网站的所有打开来源都是搜索引擎，几乎没有人通过访问URL打开网页。我们共总结了20多个类似的特征，用于描述垃圾网站。我们就通过这些特征去描述、去给垃圾网站下定义，再结合一些统计学习的算法，差不多就解决了垃圾网站的识别问题，实现了作弊手段无关的垃圾网站识别。

2. 用户点击模型

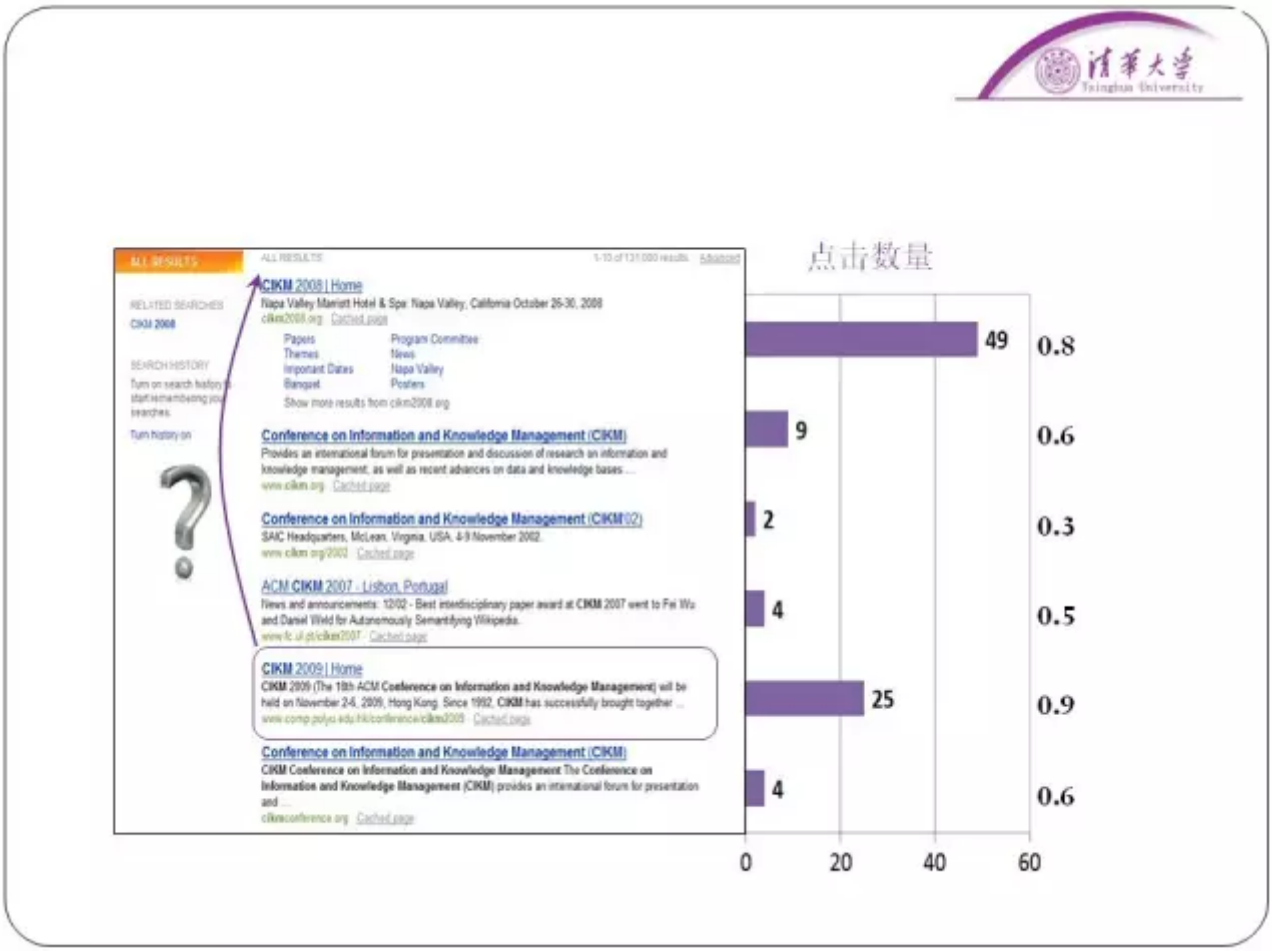
另外一个例子，就是用户点击模型。这其实是一个“前人栽树，后人乘凉”的工作，把搜索引擎上面的用户反馈情况记录下来分析用户需求。有一个简单的想法点击量越高越有可能是用户想要找的东西。



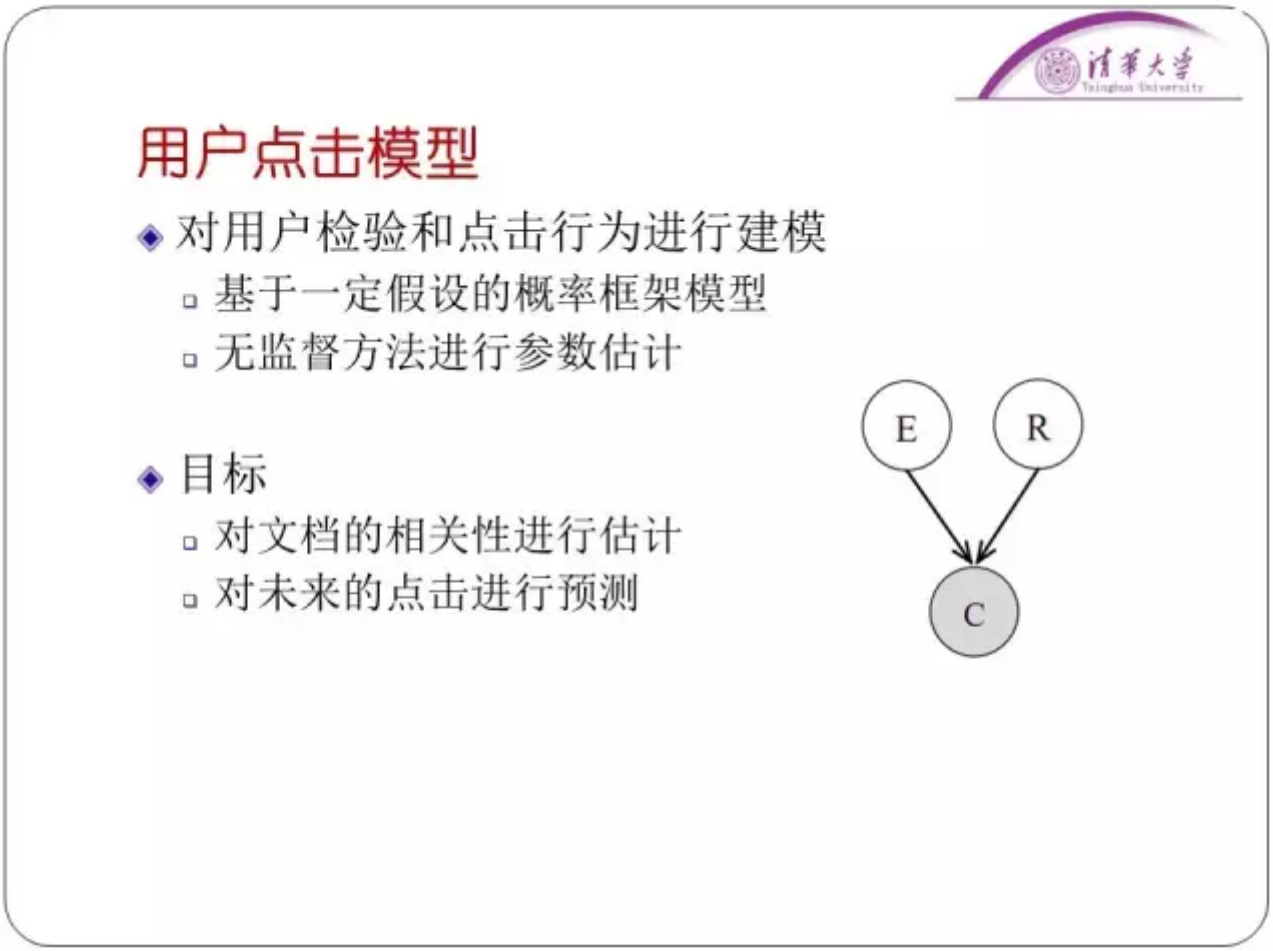
但是完全按照点击量来判断是不行的，因为网页在搜索引擎的不同位置是影响这个结果的，从而影响客户需求分析的准确性。



用户有可能就点击前几个，然后随手滚轮一滚动就点击到第九、十个去了，我们要做的工作就是把这个影响消除掉。我们找一个简单的例子来示意，比如这个CKIM会议的搜索结果，比如今年是2009年，今年的CKIM会议网站新出来，如果直接按点击量排序，在搜索引擎上搜索，2008年的会议肯定会被排序到前面。



但是实际上大多数用户想找的肯定是最新的会议上的消息。但是搜索点击模型的目的是对文档的相关性进行估计以及对于用户的点击进行预测。那么上面的这些因素都会影响用户点击模型的分析。我们要做的是消除这种影响，帮助搜索引擎更好地排序。这件事情比较复杂，因为每个用户的点击情况都不太一样，所以我们需要引入一些假设把这个问题说清楚。



最早的点击模型叫做单一点击假设，他就认为用户输入后，用眼睛去找有关结果，然后只去点击一个他认为的最有关的结果。显然，这个假设不符合我们正常的使用习惯，我们往往要点击很多次。很快就有人提出了多点击假设。就是第一次点击之后，用户还回到搜索界面，以一个固定的概率去检验下面的条目。但是其实，用户向下面检验的概率并不是固定的，它还受很多因素的影响。

后来有人提出了UBM模型，这个模型目前也被大量使用着。

已有的点击模型

- Cascade model
  - “单一点击”假设 $P(E_{i+1}=1|E_i=1,C_i)=1-C_i$
- Dependent click model
  - 多点击假设，假设用户点击后以固定概率继续检验

$$P(E_{i+1}=1|E_i=1,C_i=0)=1$$

$$P(E_{i+1}=1|E_i=1,C_i=1)=\lambda_i$$

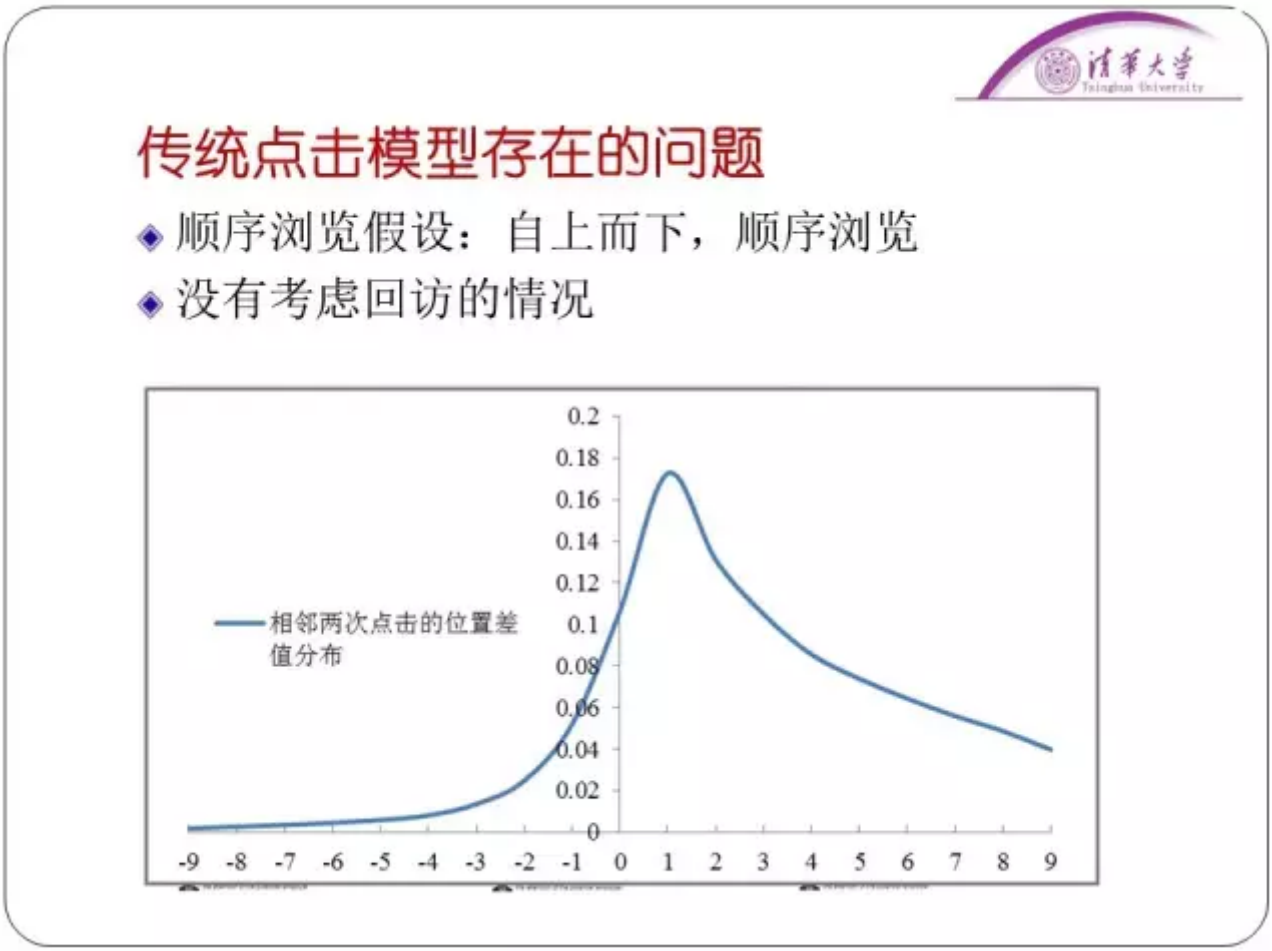
- User browsing model
  - 用户检验一个文档的概率依赖于文档位置(ranking position)和与上一次点击位置之间的距离

$$P(E_i=1|C_{1\ldots i-1})=\lambda_{r_i,d_i}$$

  - 其他点击模型：DBM, CCM, …



它的分析方法就是用户会不会继续向下检验是和当前文档的位置有关，也和前次被点击文档的距离有关。虽然这个模型效果已经不错了，但是里面仍然存在着问题。比如说这个模型是建立在“顺序浏览”假设上的。“顺序浏览”假设是指用户从上到下按顺序浏览，不考虑有“回访”的情况。



比如说用户点击了第一条、第三条然后是第五条。但是经过我们的统计，甚至是用“眼动仪”监测眼睛在看搜索引擎界面时的位置，我们得出结论：实际使用中“回访”的行为是大量存在的。往往用户点击第一条后直接点击第五条，然后可能会最后再看看第二条的内容。所以这种情况是必须考虑的。因此我们需要对这种情况建立进一步的假设。为此我们引入了两个假设，一个是局部有序性假设，也就是在两次点击之间，人眼不是从上往下看、就是从下往上看，人眼的观察浏览是有顺序的。这个就是局部有序性。

部分顺序点击模型

- 局部有序性假设: Locally Unidirectional



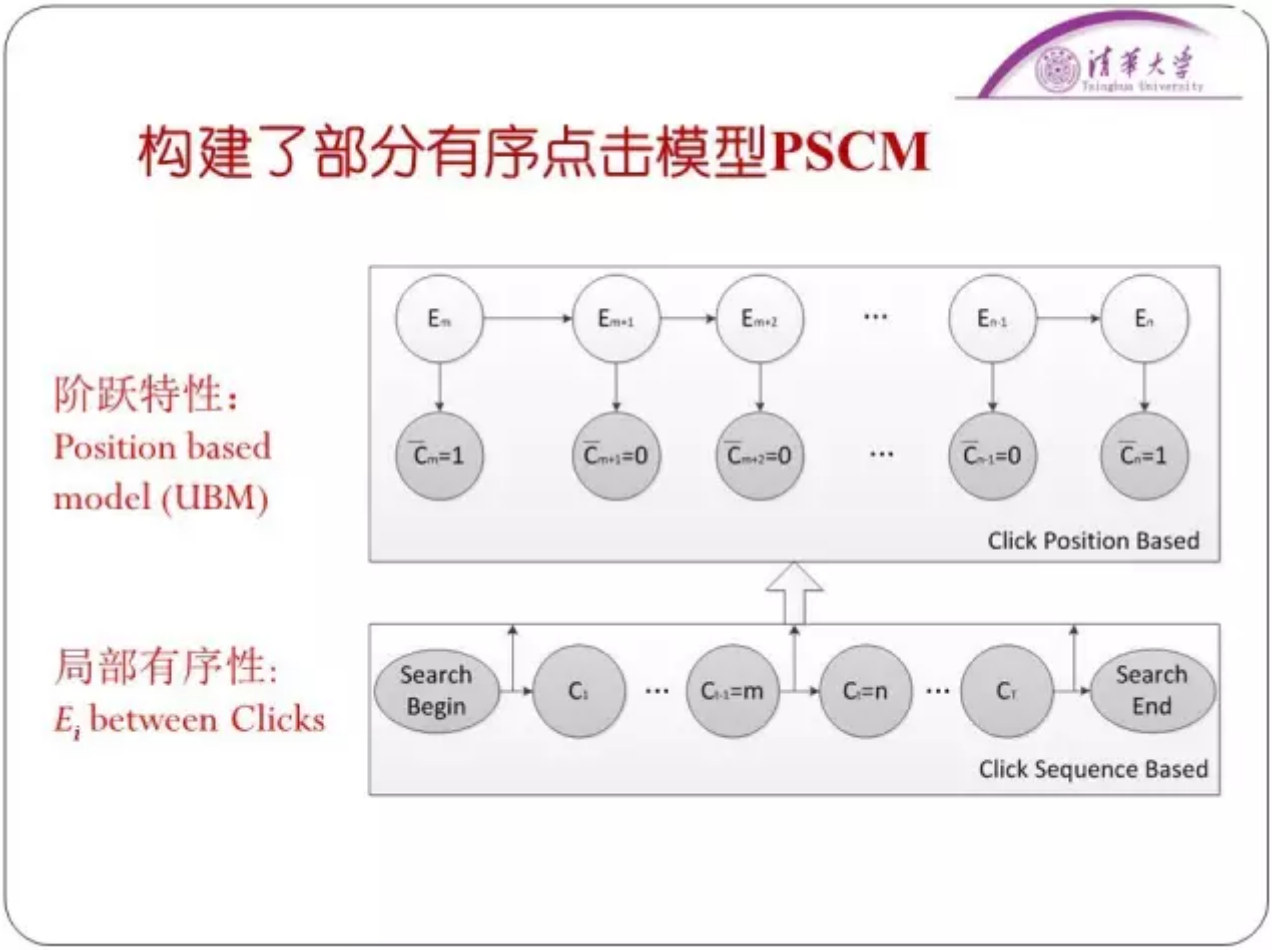
部分顺序点击模型

- 阶跃检验假设： Non-first Order



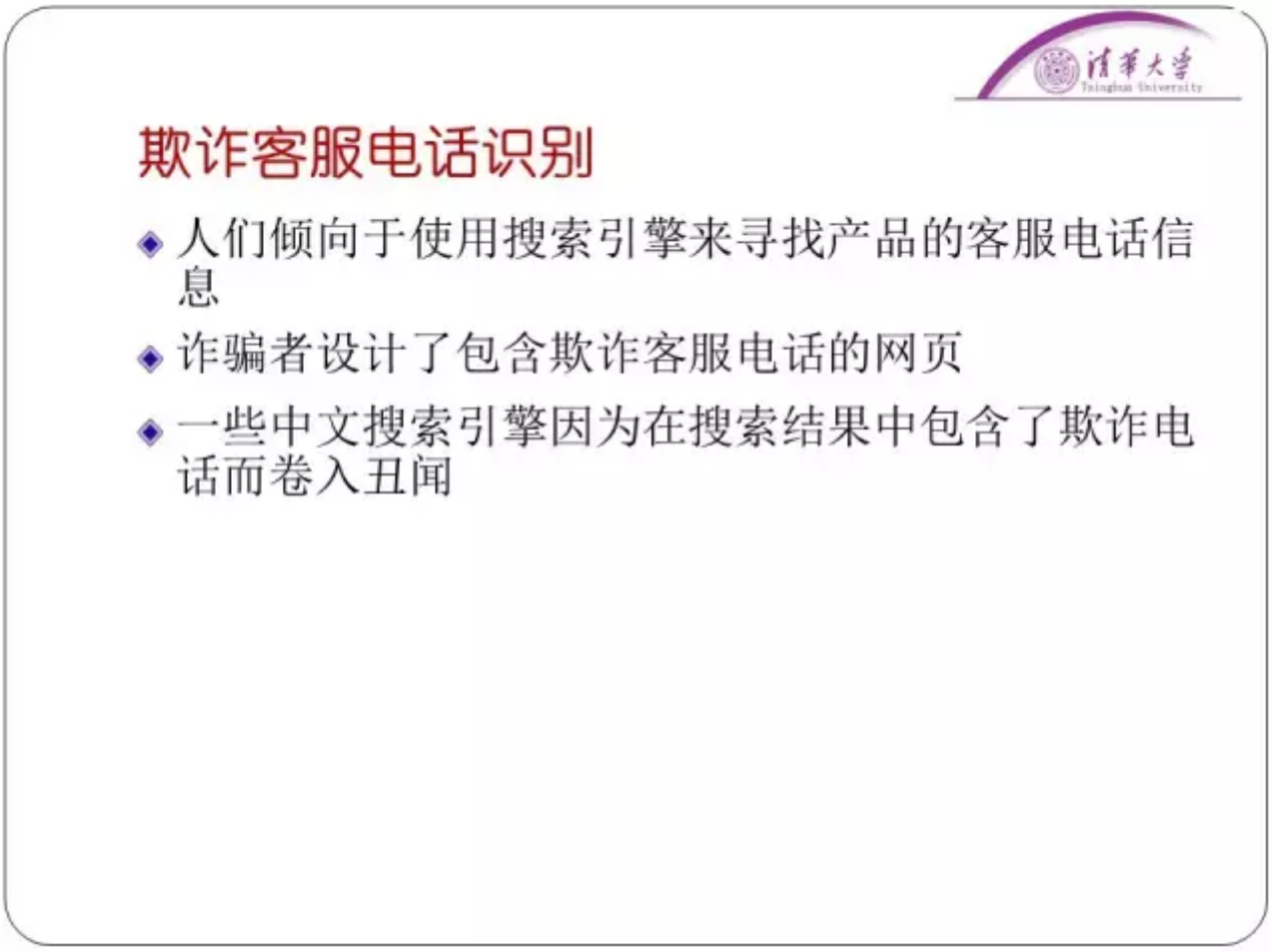
还有一个假设叫做阶跃检验假设，人眼浏览的时候不一定是按照搜索引擎的结果一条一条顺序看的，可能会出现中间跳跃的可能。这些都是符合我们眼动仪实验结果的。

在这两个假设的基础上，再结合已有点击模型的合理假设，我们构造了一个部分有序点击模型PSCM。这样，我们就在原有的UBM模型上，把用户点击模型的问题描述得更好、讲得更清楚了。实验也取得了非常好的效果，相关论文获得了SIGIR2015最佳论文提名奖。



3. 欺诈客服电话识别

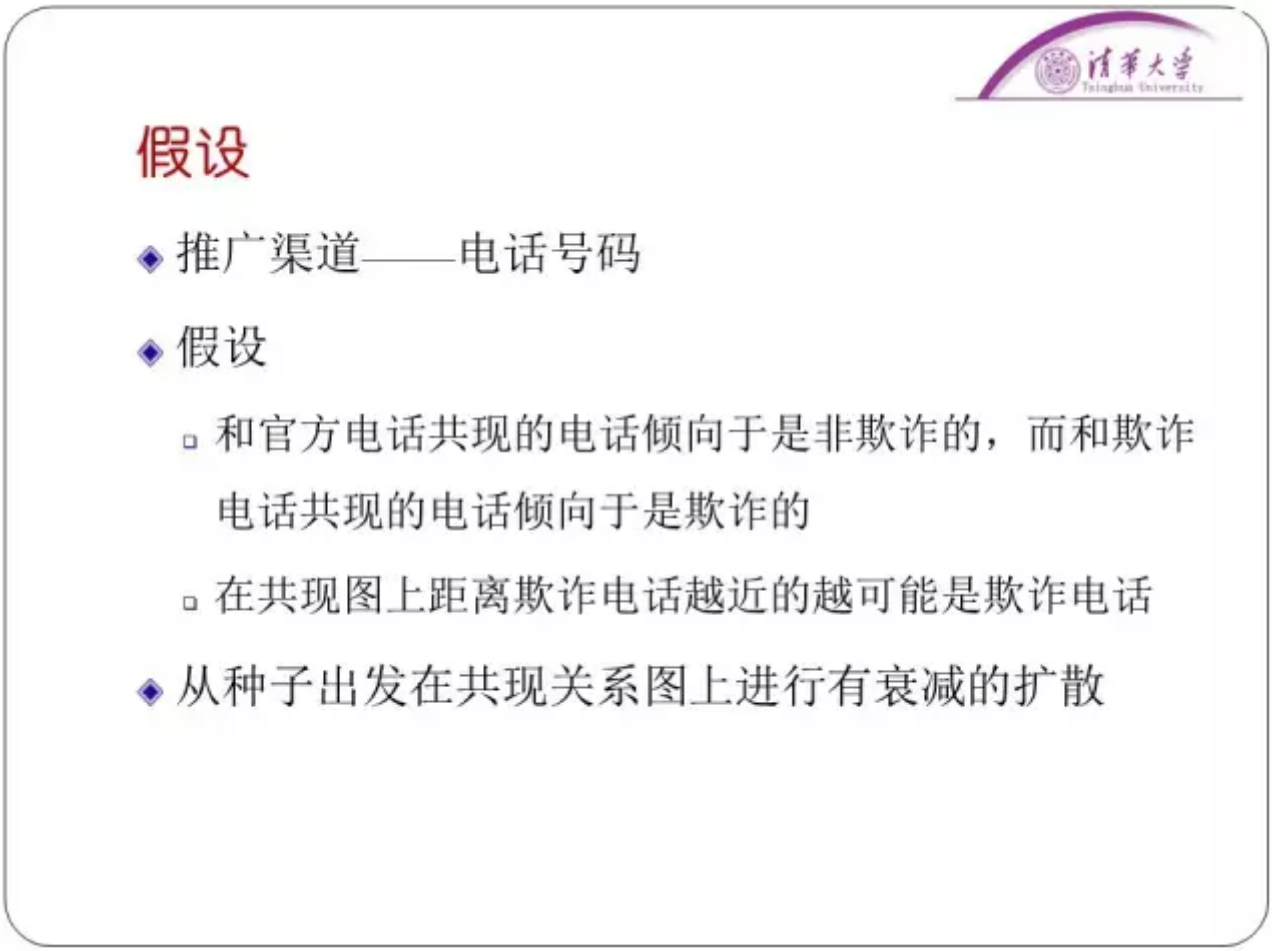
比如说冰箱坏了，从搜索引擎上搜了一个厂家的客服电话，但是打开的网页是个欺诈网页，打的电话也是个诈骗电话。最后假的厂家客服上门了之后，才发现自己被诈骗了。



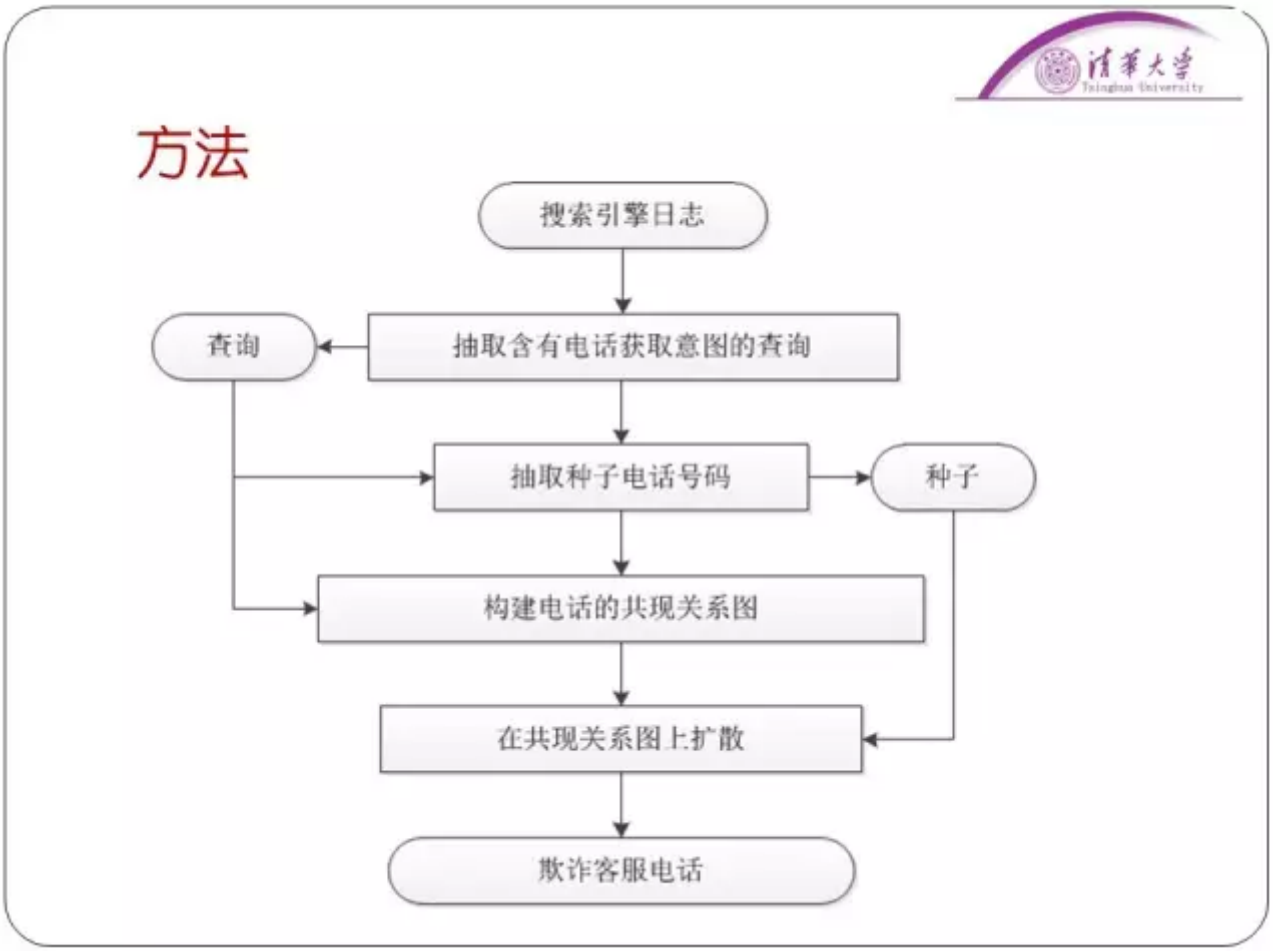
这种涉及到搜索引擎的诈骗问题也让很多公司苦恼，甚至还有很多公司因为这个问题被诉讼了。我们当时也觉得这种欺骗电话识别很难做。因为诈骗网页完全可以照抄厂家的官网，只是上面的客服电话不一样。后来调研了一番，我们发现这个问题实际上是可以被描述清楚的。这里面有哪些挑战呢？有的网站并不属于我们定义的垃圾网页范畴，所以采用垃圾网页识别的方法并不能很好的解决这个问题。

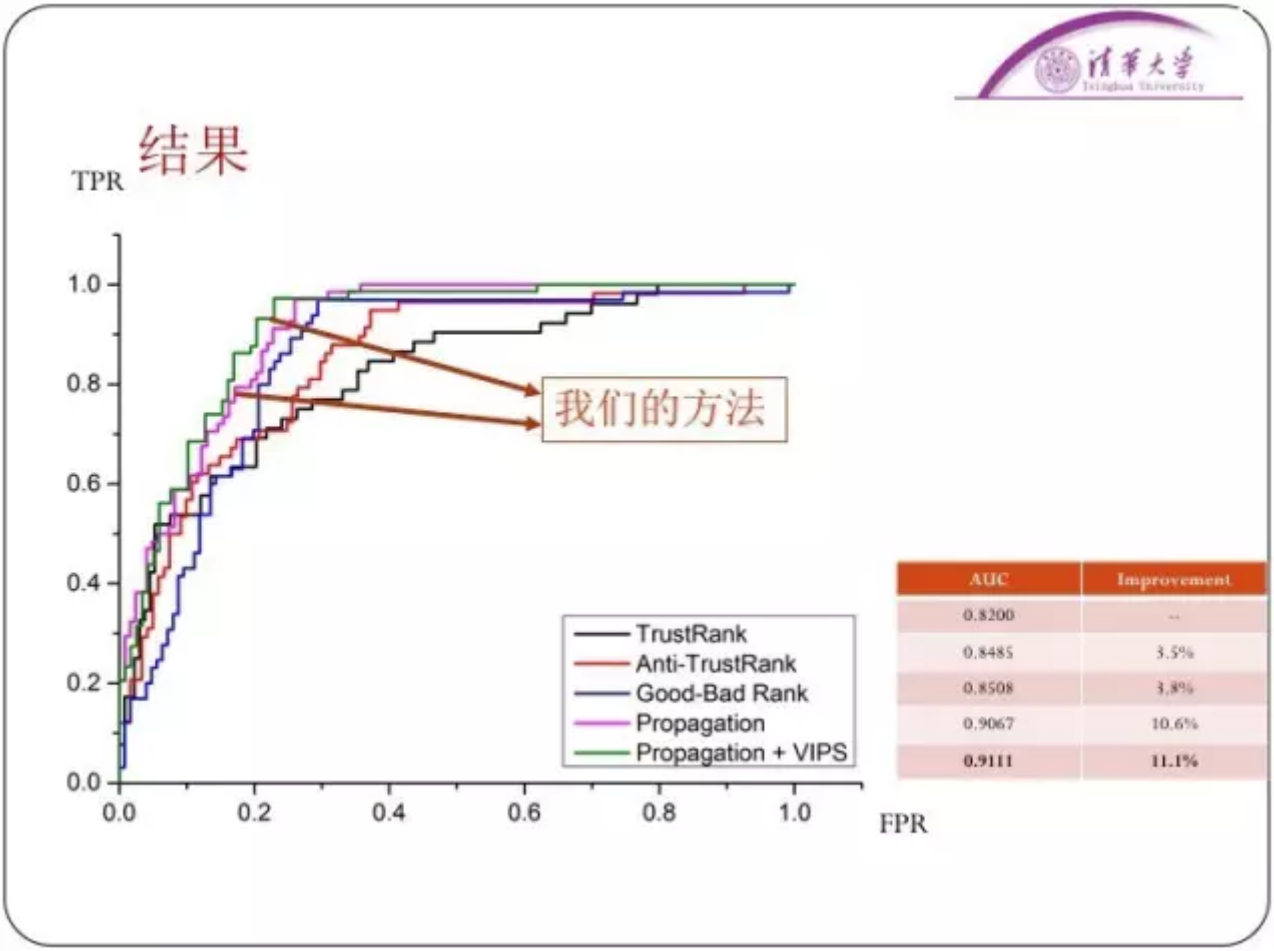


后来经过我们调查发现，欺诈网站都有一个特点就是一个诈骗电话会伪装成多个不同客服，可能在这个网页是一个厂家的，另一个网页是另一个厂家的。我们就试图用这个特征去定义欺诈客服电话。



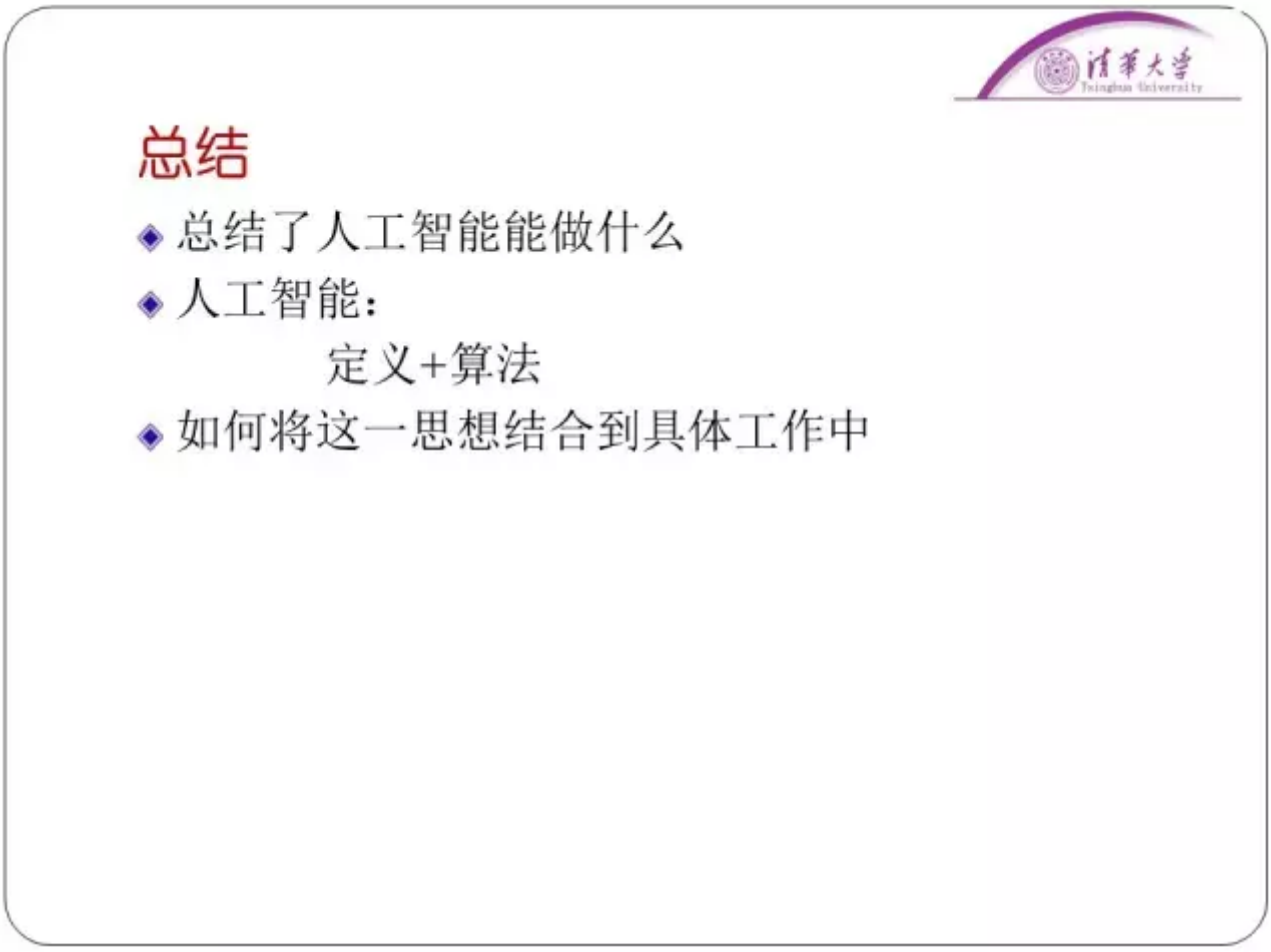
为此我们标记了一些真正的客服电话作为“正面”种子，定义了一些欺骗客服电话作为“负面”种子，我们定义了电话的共现关系图。比如说和官方客服共现的一般是可信赖的电话，和欺诈电话共现的号码一般是欺诈电话。通过在电话的共现关系图中迭代传播“正面”电话和“负面”电话，最终识别出欺诈客服电话，取得了很好的结果。


方法
搜索引擎日志
抽取含有电话获取意图的查询
查询
抽取种子电话号码
种子
构建电话的共现关系图
在共现关系图上扩散
欺诈客服电话


结果
TPR
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FPR
我们的方法
TrustRank
Anti-TrustRank
Good-Bad Rank
Propagation
Propagation + VIPS
AUC
Improvement
0.8200
--
0.8485
3.5%
0.8508
3.8%
0.9067
10.6%
0.9111
11.1%

五、总结

简单地总结一下，我们通过这些实例总结了人工智能到底能够做什么，得出来一个结论。现阶段所谓的人工智能就是通过定义加算法求解问题。关键是要想办法把目前还不能解决的问题是否可以从新的角度进行定义，无论是依靠特征，或者依靠数据，或者依靠假设，想办法把问题描述清楚，这之后再配以相应的算法，这样的话问题就可能得以解决。对于我们工作中解决不好的问题，就朝着这个方向努力去做。这就是我今天的思想，希望能够大家带来一点启发。谢谢。



本文经马少平教授授权，由明略数据整理发布。如需转载，请联系明略数据公众号后台（微信号：Minglamp_BigData）

本文编辑：赵奕、吕诚成

相关阅读

「人工智能创业30人」明略吴明辉：比起科学家，我更喜欢当工程师和创业者
明略吴明辉：行业领军企业+AI科技企业将引爆行业人工智能发展
明略人· 孔誉乾：情报研判遇上行业人工智能，神探已来！
明略人·杨昀：效率提升3000倍，可信任的AI让金融风险更可控
明略人·黄代恒：让轨道交通零延误从“预先感知故障、人机交互洞察”开始

明略人·魏明：普惠金融难？小额信贷不良率降低60%，我们是这么做的...

明略人·徐安华：笔录/研报/征信报告，有NLP加持，看我们如何让机器懂“行话”...

明略人·尤旸：让AI可感知，探索新一代交互范式下最有效的知识表达

