第132封信 | 计算机的角色和信息的作用（下）

吴军 今天 00:00

第132封信 | 计算机的角色和..
08:56 4.18MB

信件朗读者：宝木

小师弟，你好！

今天再和你分享一下朱雪龙教授对我做研究的一次点拨，并由此谈谈信息的作用。

接着昨天的话题，我在清华电子工程系后来就找到了语音识别这个研究方向，应该讲工作做得还不错。有一天我比较早地到实验室，实验室里也没有别人，朱教授也到了，他不是我的导师，平时从来没有和我聊过我的研究工作。这天因为实验室里就我们俩人，他来和我聊天，问我在做什么。

我就讲自己在寻找一种更好的模式分类方法，将发音相似的几个元音识别得更清楚些，并且取得了一些进展，降低了大约10%左右的相对识别错误率（原来错5%，现在错4.5%，就是相对错误率降低了10%），还想再试图改进一下方法，再降低一点错误。我指望着朱教授夸我一番，然后说，不错，继续努力吧。谁知道他给我泼了点冷水，说道，你的做法虽然有效，但是也做到头了，别在这条路上再浪费时间了。

我当时就很纳闷，问他为什么？他对我讲，减少语音识别的错误，就等于要消除不确定性，而消除不确定性，就要使用新的信息。你使用的仅仅是你前面学生们使用过的老的信息，他们的模型做得不是很准确，使得你能够有一个小的提升空间，但是你能得到的油水就这么多，再怎么玩，也玩不出什么新花样。要想进一步提高，就需要寻找新的信息来源。

朱教授是研究信息论的人，对于信息的作用有非常深刻的理解。其实信息论的发明人香农早就指出，要想消除一个系统的不确定性，就必须使用信息。当你没有收集到足够多的信息时，不确定性就是一种客观事实，无论采用什么方法，都不可能消除。

朱教授的观点其实是转述香农已经严格证明的理论。不过，他的这番话我最初也没有太在意，还是沿着自己原来的思路试了试其它方法，看来似乎油水也就这么大，便不再试了，回想起他的话，才觉得非常有预见性。以后，我越来越体会，在IT领域做事，要想获得好的结果，就需要挖掘先前别人挖掘不到的信息，如果使用的还是别人已有的信息，不论把模型建得多么准确，取得的进步都非常有限。

后来我到了约翰·霍普金斯大学读博士，继续从事语音识别的研究，对朱教授的这个观点有了更深刻的认识。今天的语音识别从本质上讲，只有两个部分，利用声学信息提高语音的识别率，利用自然语言中的信息，消除语音的错误，提高文字的识别率，我研究的就是后者。

在很长的时间里，大约从上个世纪80年代初到90年代初，全世界学术界对这个问题的看法是，我们的模型不够准确，因此很多识别错误消除不了，但是十多年下来，其实全世界学术界没有什么拿得出手的成果。

90年代初，贾里尼克教授从IBM的高管回归到学术界，他发现大家的路完全走错了，因为如果不挖掘新的信息，很容易就遇到天花板。于是1996年夏天，他召集了全世界这个领域顶级的科学家到约翰·霍普金斯大学工作了一个暑假，大家一起寻找之前没有利用的语言信息。

之后他和研究中心的另外三个教授一起给美国自然科学基金会写了个建议书，该基金会就给了约翰·霍普金斯大学一笔钱，利用新的信息，改进语音识别。也就是靠这笔科研，我得到了奖学金，得以在学校一直读完博士。

顺带讲一句，贾里尼克教授也是美国70、80年代信息论教科书的作者。而参与这个研究工作的另外三个教授还包括我之前提到过的我的导师库旦普教授，我过去的指导教师布莱尔（他后来开创了微软的网页搜索部门，后来主管整个eBay的研究），以及今天在自然语言处理领域颇有影响力的雅让斯基。你从很多学者在这个问题上一致的看法，或许能体会到信息的作用。

言归正传，我从那时起就开始寻找各种能够消除语音识别错误的方法。过去大家能够找到的只有上下文信息，后来我挖掘谈话内容中主题和相关话题的信息，果然能够提高语音识别率。再后来，我又对所要识别的内容进行语法分析，挖掘它的语法信息，就进一步提高了识别率。这些信息，之前的学者们都没有利用过，所以我们使用了，效果就非常好。

等我到了Google后，和辛格博士一起改进Google的搜索算法。Google最初的第一版搜索算法是佩奇和布林设计的，叫做Scorer，就是打分系统的意思。它和之前Yahoo等公司用的算法不同的是，额外地使用了网页排名PageRank的信息，因此效果一下子好了很多。辛格博士在Google写了第二个版本的排序算法，叫做Ascorer，其中的A就是辛格博士名字的首字母。这个算法让Google搜索结果的点击率有非常明显的提高，而其本质并非调整了多少网页搜索排序中的数学公式，而是使用了新的信息，具体讲，就是上下文信息。

后来辛格博士指导我做第三个版本的Google搜索算法，叫做Ascorer2，其核心还是利用更多的信息，具体说，就是一些文法搭配和语义的信息。

我先完成了中日韩文的搜索算法，由于使用了更多的信息，搜索的相关性提高了十多个百分点。再往后，我们一同将这个思想用回到英语，把英语的相关性又提高了几个百分点。最后，另外一批同事将这个方法用于欧洲语言，主要包括法语、意大利语、德语和西班牙语，我们称之为"无花果"，也就是英语里的FIGS，它是上述四种欧洲语言首字母的缩写。

在辛格领导Google搜索的时期（直到2015年），绝大多数改进都是围绕信息的使用上。90%的改进来源于找到了新的有用信息，只有不到10%的改进，在于用更好的机器学习方法，把模型的参数训练得更准确。

为了找到有用的信息，辛格博士自己以身作则，虽然后来他已经做到了主管整个搜索业务的高级副总裁，但是他每天坚持分析一些因为缺乏信息量，而做不好的搜索情况，然后在群组中讨论。在他的带领下，整个部门变成了一个数据公司。当然，到后来最好用的信息已经不容易找了，Google的搜索部门也不得不花很多力气调整算法的精度。但是，后来那些改进，幅度只是当初改进的1/10，甚至1/100。

相比Google搜索，微软的Bing采取的是另一条技术路线，更多地强调模型的训练。这件事虽然和寻找有用信息不矛盾，但是当一个公司资源有限，只能保证一头时，就要做一个决策了，到底先强调哪一方的工作。

从朱雪龙教授、到贾里尼克，再到辛格博士，对我的影响都是一致的，即要消除不确定性，就需要不断寻找新的信息，这便是这几位信息论专家对信息的理解。

我后来在《智能时代》一书中讲，所谓大数据思维，本质上就是利用信息消除不确定性。当你无法获得他人所没有的信息时，你比他人也走不远。

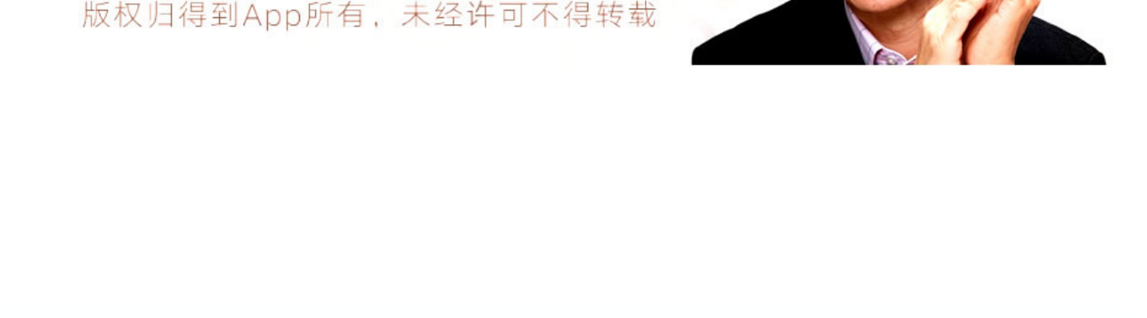
很多传统企业的人问我，我们不懂IT技术，不懂人工智能怎么办？我说，你们从业这么多年，积累下来的信息就是财富，不要做捧着金饭碗讨饭的事情，要善用这些信息。至于你们找的IT工具，使用的开源的人工智能算法是否最佳，没有太多关系。用不用信息，是面对金山银山取和不取的差异，模型比不过那些著名的IT公司，只不过是少了一两颗金豆而已。

希望这几位专家对信息的理解对你有所帮助。也希望你结合自己的工作，谈谈对朱教授关于**"要消除不确定性就必须使用信息"**的理解。

祝近安

吴军

2018年7月10日



吴军的谷歌方法论
一份智能时代的行动指南