# 在互联网上利用人的计算能力

程蔚蔚
马尔堡大学数学与计算机系
www.chengweiwei.com

互联网被认为是人类有史以来最伟大的发明之一。它革新了人类交流与思考的方式。不过人们在享受到互联网所带来的便利的同时，也受到其中一些问题的困扰。由于在互联网上发布信息是如此的简单与直接，制造不良信息的成本也大为减少。这使得人们在接受信息的同时，往往不得不小心考量信息的真实性。不良信息的一个重要类型是垃圾信息（spam）。最为人们所熟知的垃圾信息类型是垃圾邮件；当然，除了垃圾邮件之外，垃圾信息还有其他很多种不同的形式。一般讲来，发送垃圾信息的可能是人，也有可能是事先编辑好的计算机程序。下面来举一个简单例子。

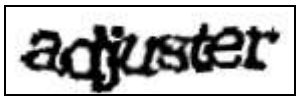

图一：一个简单的网络投票

一个大学生想要选修大学计算机课程，他于是在网上的一个论坛发布了一个如图一所示的投票，询问全美哪个大学的计算机系最好。要知道，在互联网上搞这样的投票是一件很危险的事情：斯坦福的学生一看到这个帖子，立马就开始编写计算机程序专门给斯坦福投票。第二天，麻省理工的看到这个帖子，不想输给斯坦福，于是也写了一个程序来给麻省理工投票。接下来，伯克利的学生们看到这个情况，也开始写程序给自己的学校投票……

这样的情况，当然不是希望获取真实信息的用户所愿意看到的——我们希望看到真正的人所投的票，而不是计算机程序投的。有没有办法防止计算机程序自动发布信息呢？答案是肯定的。

我们常说，开发具备人类智能级别的机器是计算机科学矢志不移的究极目标。说它是究极目标，潜台词其实就是说我们目前的水平离这个目标还相差很远。人工智能的研究者往往惊叹于人脑那叹为观止的识别、计算能力。有必要对这里说的“计算能力”做一个解释：这里说的计算能力是一个相对广义的概念。计算一个五位数的十次方，这是一种计算能力。计算机对于这种简单定义的精算，是很擅长的。一个五岁的孩子，在簇拥的人群之中把自己的母亲鉴别出来，只需要0.1秒级别的时间。这也是一种计算。人脑这种进行复杂、没有精确定义、条件模糊的计算能力，计算机目前为止还无法企及。有很多任务，对人类来说尤其简单，电脑却束手无策。这样的例子有很多。

有了这个思路，解决刚才提到的投票问题就有希望了——在投票之前先让想要投票的实体做一个事先设计好的简单的测试：人可以通过这样的测试参加投票，而计算机则不能通过测试进而无法投票。



Please type the word you see: _____

图二：一个CAPTCHA测试的例子

能够实现这个目的的测试有很多种，其中最为著名的是由卡内基梅隆大学计算机系教授Luis von Ahn等率先提出的CAPTCHA测试（图二）。CAPTCHA测试的原理很简单：从英语字典里随机挑出一个词，通过图像处理步骤将这个词进行一定程度变形（一般是模糊化或者扭曲），然后这个词被呈现给用户，用户如果能够成功的输出这个词，就算通过测试；否则系统将不能进入下一个环节。实验证明，CAPTCHA对于人类来说相对简单，而目前的计算机识别技术却几乎很难通过测试。

经常上网的朋友，在注册网站、参加投票的时候，大概都会有遇到CAPTCHA测试的经历。事实上，互联网上每天都要进行几乎2亿次CAPTCHA测试，可见其流行与重要程度。对于CAPTCHA的开发团队来说，这个数字当然是一个令人惊喜的成绩。然而，他们却在这一数字背后看到了不足：假设每个CAPTCHA测试需要花费一个用户大约10秒钟的时间，那么每天光这些测试就需要花掉全球互联网用户累计大约50万小时！说起来也可笑：全球的网民每天要付出如此大的努力来证明自己是真正的人，而不是机器。因为CAPTCHA测试是防止垃圾信息传播的一个重要手段，这每人每次10秒钟的时间显然无可避免。那么，能不能利用用户花费在CAPTCHA测试上的精力来实现一些其他有利于社会发展的目的呢？这正是reCAPTCHA项目的初衷——致力于纸质媒体的数字化。

为了便于知识更快更广的传播，大量的书籍以及报刊有待数字化，

从而可以在互联网上发布。数字化的好处不用多说：用户在时代杂志的网站（www.time.com），通过搜索，可以在几秒钟之内查阅到该杂志从1923 年到 2009 年之间的文章。在美国，很多知名的杂志报刊和国立图书馆都已经实现了完整的电子备份。图书数字化所遇到的最大挑战是如何正确识别年代比较久远的文字，尤其是手写体的文字。虽然近些年光学字符识别（OCR，optical character recognition）技术取得了长远的发展，但是对于手写文字，其正确识别率仍然较低，远不及正常人的识别率。这种时候，想要进行数字化就不得不雇用员工来进行识别，极大地增加了图书数字化的成本。


图三：一个reCAPTCHA测试的例子

大家可能也都想到了：既然我们有 CAPTCHA 测试，能不能把这些机器无法识别的字符交给想要注册网站、参加网络投票的网民去识别呢？这样的话，全球的网民就都在为图书的数字化作贡献了。不过这里依然有一个问题：网民写出来的答案，系统不知道正确与否，因为系统也没有正确的答案。对于这个问题，reCAPTCHA 测试是这样处理的（图三）：它会要求用户在一次测试当中识别两个单词，其中一个从想要数字化的图书中截取出来，另外一个系统知道其正确答案（当然，用户并不知道哪个单词是有待数字化，哪个是系统已知的）。用户给出的两个答案之一与已知正确答案匹配，即算通过测试。那个有待数字化的单词还会被随机的发送给其他很多个测试。在大量的测试结果被聚合之后，我们就能得到这个单词非常精准的数字化结果。

自从reCAPTCHA项目推出之后，很多著名的网站，比如Facebook、Twitter、TicketMaster，甚至是美国白宫的网站（www.whitehouse.gov）都已经开始从原来CAPTCHA测试转到reCAPTCHA。截至 2009 年五月份全球已经有超过 12 万家网站主动参加到了reCAPTCHA项目之中来（reCAPTCHA测试的下载与使用是免费的）。全球的网民从每次花 10 秒证明自己是人，转变为每次花 15 秒不但能防止垃圾信息传播还能帮助书籍数字化，何乐而不为？虽然只是每人每次一个词，在不到 11 个月的时间里，reCAPTCHA帮助The New York Times（www.nytimes.com）完成了其从 1851 到 1980 一百三十年间所有文章内容的数字化!reCAPTCHA项目资料显示，在这 11 个月间，一共有 4 亿人次参与了The New York Times的数字化工作！这大约就是我们常说的网络的力量。

## 相关资料与阅读

- reCAPTCHA官方网站：http://recaptcha.net/
- 维基百科：http://en.wikipedia.org/ 相关条目包括：Human-based computation、reCAPTCHA、CAPTCHA等
- Luis von Ahn个人主页：http://www.cs.cmu.edu/~biglou/

## 作者简介

**程蔚蔚**　　德国马尔堡大学(University of Marburg)数学与计算机系Knowledge Engineering & Bioinformatics实验室成员，科研助理、博士研究生；德意志银行(Deutsche Bank)数据挖掘项目咨询。主要研究方向为机器学习、数据挖掘，并在相关领域的国际重要期刊及会议上发表论文多篇。担任多个国际学术期刊、会议委员会成员，审稿人。德国马格德堡大学(University of Magdeburg)计算机硕士学位，郑州大学计算机与工商管理双学士学位。曾获得2009年第十九届欧洲机器学习大会最佳学生论文奖、2009年第二十六届国际机器学习大会奖学金、2008年马格德堡大学最佳毕业生奖、2008年国际机器学习Summer School奖学金、2006年德国萨哈森安哈特州教育与文化部优秀国际学生奖学金、2002年中国河南大专辩论赛最佳辩手、1999年建国五十周年演讲比赛安徽省安庆市三等奖。

电子邮件：roywwcheng@gmail.com 个人主页：www.chengweiwei.com