概率统计， 魅力无限

中国科学院院士 马志明教授

编者按：马志明院士于 7 月 12 日在我校做了题为《概率统计， 魅力无限》的大众报告。 在本次报告中， 马院士论述了概率统计近年在数学学科取得的丰硕成果，特别是近 10 年的菲尔茨奖每届都有概率；随机分析与经典数学的交汇以及其诞生的点滴轶事；概率统计在 AlphaGo、DNA 序列分析、Google 搜索引擎和金融中的成功应用， 展示了概率统计的无穷魅力。

概率统计的思想和方法正渗透到当代人类社会的众多科技领域和社会领域。 概率统计在现代科学技术和社会经济领域的应用日益广泛深入，它与其它学科，以及与数学的其它分支相互交叉、渗透，取得了极其丰富的成果, 展现了概率统计学科的无限魅力。当然，虽然概率统计的魅力无限，但我自己的学识却是有限。在今天的报告中，我将与各位分享我的一些点滴体会。我先说一说统计学科已发展成为当今科学与社会应用非常广泛的重要学科。在我国更是有特点， 成立了统计一级学科。 统计与其它领域交叉产生许多重要分支，如金融统计、保险精算、商务统计、计量统计、生物统计、保险统计和应用统计等。由于我的研究方向是概率与随机分析领域，因此在下面的报告中对概率与随机分析讲的多一些。

一．概率统计方法近年在数学学科取得的标志性成果

近年来概率统计日益渗透到数学的其它分支，取得了极其丰硕的成果，并且不断地产生新的学科分支。比如： 随机偏微分方程、随机动力系统（这两个正是楼上本次学术会议的内容）、随机微分几何、随机共形理论、随机图与随机复杂网络、随机算法、倒向随机微分方程、非线性数学期望，等等。概率统计与数学其它分支相融合，促进了数学学科的发展，最有代表性的事实就是近年来多项国际数学大奖都与概率统计有关：从 2006 年至 2016 年这十年中的菲尔茨奖（曾被誉为数学中的诺贝尔奖）， 每届都有概率， 而且非常多： 2006 年四位菲尔茨奖得主中， 有三个半与概率有关， 其中 Werner 与 Okounkov 可算是概率科班出身，Terance Tao 的许多研究涉及概率与随机矩阵，Perelman 的研究工作用到对数 Sobolev 不等式，也与概率有关； 2006 年的 Nevanlinna 奖颁发给了 Kleinberg, 他的研究工作是关于随机图和随机复杂网络及其算法；Gauss 奖设立于 2006 年， 以奖励对人类其他领域做出突出贡献的数学家， 首届 Gauss 奖颁发给了 Itô, 奖励他发明的随机积分对人类的贡献； 2007 年 Abel 奖（与诺贝尔奖奖金相同）奖给了国际著名概率学家 Varadhan; 2010 年菲尔茨奖四位得主中， Villani, Smirnor, 和 Lindenstruss 三位的工作都与概率有关； 2014 年 Martin Hairer 由于在随机偏微分方程的杰出贡献获得了菲尔茨奖, 他创造的正则性结构， 建立了新的框架， 统一了 Rough Path 理论和经典的 Taylor 展开理论。这一理论可以用来研究随

机偏微分方程和数学物理方程，预期在数学和物理的许多领域都有应用。 用这个新的数学框架可以对原来不适定的一些随机偏微分方程给出了严格的数学意义， 比如界面运动产生的 KPZ 方程， 统计力学中临界状态的宏观行为等。

二．随机分析与经典数学的交汇

讲到 Taylor 展开和随机分析， 如果从 Martin Haire 的正则性结构出发， 就太深奥了。 我讲一点最简单的： 用随机方法求解关于 Laplace 发展方程的初值问题:

$$\begin{cases} u_t = \Delta u, t > 0, x \in R^n \\ u(x,0) = f(x) \end{cases}, \qquad (1)$$

这是一个最常见和最简单方程， 数学系的每个学生都知道这一方程。求解这一方程有一个随机方法， 那就是，放一个布朗运动 X(·) 从 x 出发， f 与这个布朗运动复合后，在时刻 t 求数学期望，即

$$u(x,t) = E^x f\big(X(t)\big)。$$

$u(x,t)$就是这个初值问题（1）的解。 这个确定性的偏微分方程， 用随机方法求解就能比较直观地想象出这个解是怎么得来。

再者， 求一个区域 D 内的调和函数， 使得在边界上等于指定函数 f：

$$\begin{cases} \Delta u = 0, \ x \in D \\ u(x) = f(x), \ x \in \partial D \end{cases}。 \qquad (2)$$

学偏微分方程的学生都会解这一方程，但是要求区域边界具有一定的光滑性。 如果用概率方法解就很直观。 在区域内放一个布朗运动 X(·)， 让它按照布朗运动的规律跑， 跑到边界让它停止， 跑到边

界的时刻记为τ， 为初次跑出该区域的时间。 那么， f 与布朗运动在时刻τ的值复合之后再求数学期望就是这个调和方程边值问题（2）的解, 即

$$u(x) = E^x f(X(\tau))。 \qquad （3）$$

我这里讲的两个例子都是很简单的情况，楼上学术会议讨论的都是非常复杂的情况， 复杂的多， 但原理都是一样的。 概率统计的魅力， 如果我只这样说， 你们也许没有感觉， 因为用确定性方法也能解出这些方程。我再给你们举一个例子， 更能体现出概率直观思维的独特优势。刚才我们曾经提到关于 Laplace 算子调和方程的边值问题，即方程（2）。 这个方程从分析的角度来看，是一个适定的微分方程 。意思是，只要区域边界足够好，方程（2）就存在唯一解。现在我们把 Laplace 算子换成Δ^α (应理解为$-(-\Delta)^\alpha$)，$0<\alpha<2$，即考虑如下微分方程的边值问题：

$$\begin{cases} \Delta^\alpha u = 0, & x \in D \\ u(x) = f(x), & x \in \partial D \end{cases} 。 \qquad （4）$$

你会遇到极大的困惑，因为方程（4）的提法不适定，它有无穷多个解。为什么方程（2）和方程（4）有这样的不同？从纯分析的角度很难理解。但从概率和随机分析的角度，就可以很直观地明确回答这个问题。事实上，方程（2）和方程（4）的解都可以用概率的方式表达为（3）。所不同的是，对于方程（2），表达式（3）用到的随机过程 X(・) 是布朗运动，它是以拉普拉斯算子作为无穷小生成元的马氏过程。而对于方程（4），表达式（3）用到的马氏过程 X(・) 是以Δ^α作为无穷小生成元的α-稳定过程。由概率论的知识，我们知道**布朗运**

动的轨道是连续的，因此 $X(\tau)$ 的值集中在区域 D 的边界上。但 α-稳定过程的轨道是纯断的。 一个纯断的过程碰到边界， 很可能在碰到边界的时刻有跳，可能跳到区域 D 外，使得 $X(\tau)$的值分散在区域 D 的外部。因此，对于方程（2）， 只要指定 f 在边界的值，就可以由（3）唯一确定它的解。而对于方程（4），则需要指定 f 在区域 D 外的所有值，才能由（3）唯一确定它的解。由此不难理解，关于算子Δ^{α}的调和方程的边值问题的适定性的提法，不是（4）而是如下的（5）：

$$\begin{cases} \Delta^{\alpha} u = 0, & x \in D \\ u(x) = f(x), & x \in R^d \setminus D \end{cases} 。 \qquad (5)$$

从这里可以看出： 概率论和随机分析有它独到的优点， 这个独到的优点， 做纯分析的人， 如果他没有这方面的训练， 没有这方面的知识背景是想不到的， 或者是感觉不到的。 这就是为什么现在越来越多的做数学、做纯分析的人， 他也要用随机方法。 这里只举几个简单的例子， 还有很多其它的例子。

概率统计的魅力不局限于自然科学，在社会经济领域甚至我们的日常生活中，概率统计也有重要影响。举例来说，资产定价理论的 Black-Scholes 公式，必须要用到 Itô公式和随机分析。这几年金融数学非常热， 我的好几个学生毕业后都到了金融行业、到了银行和 J. P. Morgan 公司等单位工作。其实我本人并没有研究金融数学。但我们的学生的优势就在于他们会随机分析，会用 Itô公式，因此经济金融行业愿意录用他们。

刚才我们说过， Martin Haire 2014 年获得了菲尔茨奖， 他的一

个很大贡献是创建了正则性结构的数学框架。他自己说： 他的正则性结构是 Taylor 展开的一个推广。 学数学的人都知道什么是 Taylor 展开。借这个机会，我利用 Taylor 展开来普及一下随机分析中的 Itô 公式， 说明如何从 Taylor 展开的观点来理解牛顿公式（Newton-Leibnitz 公式）和 Itô公式的联系和区别。

在黎曼积分意义下，只要 f 光滑，就有 Newton-Leibnitz 公式：

$$f(t) - f(0) = \int_0^t f'(s)ds。$$

一般地, 对复合函数 $f(X_t)$ (如果函数X_t的性质适当好), 也同样可以积分：

$$f(X_t) - f(X_0) = \int_0^t f'(X_s)dX_s，$$

这就是 Newton-Leibnitz 公式。回忆一下 Newton-Leibnitz 公式的推导。记

$$0 = t_0 < t_1 < \cdots < t_i < \cdots t_n = t, \qquad \Delta X_i = X_{t_{i+1}} - X_{t_i}。$$

由 Taylor 展开，我们有

$$f(X_{t_{i+1}}) - f(X_{t_i}) = f'(X_{t_i})\Delta X_i + o(|\Delta X_i|)。$$

如果曲线X_t是可求长的，即

$$\lim_{\max_i |\Delta X_i| \to 0} \sum_{i=0}^{n-1} |\Delta X_i| < \infty,$$

那么我们可以得到

$$f(X_t) - f(X_0) = \lim \sum f'(X_{t_i})\Delta X_i + o(\sum |\Delta X_i|) = \int_0^t f'(X_s)dX_s。$$

这就是通常黎曼积分的思想。但是，对股票价格这样的曲线，或者布朗运动的轨道，你要这样求 $f(X_t) - f(X_0)$， 就求不出来。 为什么？因为这样的曲线是不可求长的 即

$$\lim_{\max_i |\Delta X_i| \to 0} \sum_{i=0}^{n-1} |\Delta X_i| = \infty \text{ 。}$$

这种曲线，当（时间）区间分割划分，对应的折线和收敛于无穷大，对再小的（时间）区间分割划分，这一极限都是无穷大。对这种曲线按照普通的（黎曼）意义是不可能求出长度的，不能定义黎曼积分。但是 Itô 注意到， 对布朗运动的轨道，虽然折线长度之和取极限是无穷， 但是，折线长度的平方之和取极限却是有限：如果X_t是布朗运动，则几乎必然地有

$$\lim_{\max_i |\Delta X_i| \to 0} \sum_{i=0}^{n-1} (\Delta X_i)^2 = t - s,$$

对 s<t 成立。因此，在 Taylor 展开中， 不是只展一项， 而是再展一项：

$$f(X_{t_{i+1}}) - f(X_{t_i}) = f'(X_{t_i})\Delta X_i + \frac{1}{2} f''(X_{t_i})(\Delta X_i)^2 + o((\Delta X_i)^2) \text{。}$$

这样一求和，当分割划分越来越细时，高阶无穷小消失了，除了第一项是通常的 Newton-Leibnitz 这一项以外，还出现了 1/2 f''(X_s) ds。这就得到了 Itô公式：

$$f(X_t) - f(X_0) = \int_0^t f'(X_s)dX_s + 1/2\int_0^t f''(X_s)ds \quad \text{。}$$

Itô公式现在讲起来比较轻松，当时发现是非常不容易的。这是 Itô非常重要的发现。 Itô公式在我们人类的无论自然科学还是社会科学中都用的非常广泛。 因此， 他得了首届高斯奖。下面一段是给 Itô颁奖的英文颁奖词的中文翻译。

“伊藤清发展出一个全新的数学形式体系——随机分析，让数学家们

能够用随机偏微分方程来表示随机的组合和其决定的力量。如今，伊藤清的理论已经应用到股票分析、生态系统中人群数量的测算以及复杂生物学的测算之中。随机分析成为数学领域中一个重要而富有成果的分支，并对技术、商业和日常生活产生了重要影响。”

那是 2006 年, 在马德里的国际数学家大会上。2006 年给 Itô颁奖的时候我在场。 那时我正是国际数学联盟执委会的副主席， 就在主席台上。给 Itô颁奖时，他本人已经 90 岁了。Itô的女儿代替父亲来领奖，她念了一页他父亲写的感言。其中一段话为 （还有别的话）：“我自己关于随机分析的研究是纯数学的。 因此， 把应用数学的高斯奖颁发给我的确出乎意外， 我深深的感谢（extremely unexpected and deeply gratified)!” 。这可见数学的力量!

讲讲 Itô的一些轶事还是挺有趣的， 对年轻人很有启发。Itô 1915 年出生在日本的三重县， 他的父亲是日本文学和汉语文学的一位中学教师。他从东京大学数学系毕业以后， 并没有直接进入数学科研机构或研究所工作， 而是到了东京的政府统计局做一名职员， 直到 1943 年才到日本名古屋大学做了副教授。 Itô的最初两篇文章写于 1942 年， 第一篇论文研究 Lévy 过程的分解， 他给出了后来被称之为 Lévy-Itô分解的著名结果，这是我们现在学随机分析必须学的经典结果； 第二篇论文是用日文写的、油印的、 蜡纸刻的论文， 这篇论文包含了随机积分。 Itô积分起源于他在统计局任职员时写出来的手稿。 后来， 第二篇在二次大战之后， 在 50 年左右， 由美国数学家 Doob 推荐，又加了些内容，写了一篇很长的英文文章发表了。

再后来， 他就非常有名了。 据查， 他直系的（真传的）博士生就是 Watanabe, Kunita, 和 Fukushima 三位。 在他的 Seminar 上， 培养了一批世界一流的随机分析学家， 包括 Ikeda, Tanaka, Motoo, Hida, 和 Nisio（女数学家）等等。

顺便说说，我差点成为 Itô的学生。81 年 Itô访问中国科学院时，我在科学院做研究生，硕士论文是关于点过程。我陪 Itô爬长城时，给他介绍了我做的研究工作，他听了很感兴趣。Itô对点过程很熟悉，他的一个很有名的工作就是发现布朗运动的 Excursion 是 Poission 点过程。Itô告诉我他要推荐我到日本念博士。这是他回国后寄给我的明信片：“strongly recommend you....”， 他推荐我由日本振兴会资助到日本京都大学念博士。后来因故我没去成日本，而是获得洪堡资助到了德国。虽然没有做成 Itô的学生， 但是，Itô的学生 Fukushima 对我的帮助非常大。在德国做洪堡时，与我的导师 Albeveria 和 Rockner 一起把 Fukushima 的狄氏型研究工作推广到拟正狄氏型。初稿写成以后， 我们把手稿寄给 Fukushima（以前在德国见过他）， 邀请他提提意见或和我们一块合作。 Fukushima 看了我们的手稿之后， 甚至来不及写信， 发了一封电报给我， 说：他相信我们的工作将来肯定会成为马氏过程被经常引用的文献， 叫我们赶快发表。 Fukushima 非常肯定的话对我帮助非常大。 那是 90 年的事。 后来，Fukushima 为我筹到经费， 去京都参加 90 年的国际数学家大会， 这是我第一次参加国际数学家大会。94 年我作为 45 分钟邀请报告人在国际数学家大会上介绍我们关于拟正则狄氏型的工作。

关于随机分析， 也还有一些有趣的话题。后来， 法国人发现有一位叫 Wolfgang Doeblin 的犹太数学家，他出生在德国柏林，1933 年去巴黎入了法国籍， 25 岁在与德国交战时阵亡。 他生前最后两年在军中服役，同时，写下了不少珍贵的数学手稿。在与德国交战之前，他把手稿用密封信件送到巴黎科学院存档。60 年后的 2000 年，经他的兄弟同意才解密。法国人呼吁：他们在手稿中惊异地发现， 在 Doeblin 潦草地写在学生练习本的笔记中， 已经隐藏有用 Itô随机方法求解 Kolmogorov 抛物偏微分方程的思想。历史会有很多偶然。如果 Doeblin 手稿早点流传出来， 也许现在不叫 Itô公式了， 有可能叫 Itô- Doeblin 公式，历史会有一些不完善的地方。

关于概率统计与经典数学的交汇， 我就讲这些。 更多的， 楼上还有很多参加随机动力系统会的专家， 你们可以去问他们。 现在做纯数学的很多人都在学习随机方法， 这方面确实是魅力无限！

三．深度学习和强化学习中的概率统计

给大家讲讲比较有趣的深度学习和强化学习中的概率统计。之所以选取这个题材，是因为四个月前， AlphaGo 战胜世界围棋冠军、韩国九段围棋手李世石， 在人类社会掀起了不小的波澜。AlphaGo 算法设计的主要工具就是深度强化学习和蒙特卡罗树搜索，这里面用到大量的概率统计。 下面我主要讲讲 AlphaGo 用到的概率统计。在讲述之前， 我公开申明： 我要感谢微软亚洲研究院的贺迪。 起因是中国科学院大学的一二年级大学生做科创计划， 他们选择了学习

AlphaGo 的科创计划， 研究 AlphaGo 的概率统计原理， 希望我做他们的导师。我就通过我在微软工作的过去的学生邀请到贺迪，请他给我们作报告介绍 AlphaGo 的原理。下面介绍的内容部分取自贺迪的报告，部分取自查阅互联网获得的资料，不一一注明知识产权的出处。

人工智能下棋已经有很长历史， 过去 IBM 有一个深蓝团队， 用“深蓝”计算机下国际象棋。 国际象棋所有棋局可能性约10^{47}， 围棋的所有棋局的可能性大约是2×10^{170}, 而全地球的原子总数也只有10^{80}。围棋所有棋局远比地球所有原子数目多， 这真是一个大数据。过去 IBM 团队用“深蓝”同人类下国际象棋时， 可以把人所有下国际象棋的步骤穷举。但是， 围棋做不到， 围棋不能穷举！你想，这么大的天文数字怎么能穷举？！ 围棋只能用随机方法、只能用概率方法，这正是体现了概率统计的重要性。

谷歌的研发团队用深度学习和强化深度学习为 AlphaGo 训练了四个神经网络， 用通俗的语言， 这四个网络分别是： 快速走子网络、走棋网络、强化学习网络和估值网络。 他们先用 3 千万局人类下棋的棋谱来有监督地学习出两个模型：其一是用 13 层的卷积神经网络学出来的走棋网络，另一个是用逻辑回归学出来的快速走子网络。这两个网络都可以近似理解为基于 3000 万个有标注的数据< s , a >，评价在当前局面 s 下，棋子落在某一位置 a 的概率，也就是 p(a|s)。其中“快速走子网络”可以被看作是“走子网络”的轻量级版本，它能够比“走子网络”快 1000 倍，但是精确性较差。在走子网络的基础上，通过机器和机器自己对弈，由产生多达 3000 万个标注样本，每个样

本的局面 s 都来自不同的一局棋，用大量增加的样本训练出强化学习网络。而第四个网络，是在走子网络和强化学习网络的基础上训练出来的估值网络，它可以估出在当前棋局下胜算的概率值。总体来说，前三个神经网络都以当前围棋的对弈局面为输入， 经过计算后， 输出可能的走子选择和对应的概率。概率越大的点意味着神经网络更倾向于在那一点走子， 这个概率是针对输入局面下所有可能的落子点都有一个概率。 第四个神经网络是用来进行价值判断的，输入一个对弈局面，它会计算出这个局面下黑棋和白棋的胜率。我的理解，四个网络都是概率， 前三个都是概率矩阵， 第四个是一个概率值。

真正对弈的时候， 用的是蒙特卡罗树搜索(MCTS)算法， 它也是吸收了概率的思想。 现在很多的计算都是用蒙特卡罗方法， 它的中心思想是按照一定的分布去落点， 因为分布是给定的， 落点落多的时候， 自然地，原来分布所要求的函数就能够得到， 计算机也就会把它绘出来。AlphaGo 怎么下围棋？ 刚才四个网络做好了， 相当于四个大脑。 现在从当前位置的棋子出发， 它要计算不知多少遍， 才走出一个棋子。 它怎么走？ 直观地解释，它根据神经网络选出一个路径走， 走到一定程度让它扰动一下， 再继续走下去， 看它是输还是赢，最终给出一个判断这一步走子输赢的值，这个值用快速走子网络（它能很快把棋走到底决出胜负）和估值网络估出来的输赢概率按一定公式计算出来。然后返回到原来准备要走的地方。这就是蒙特卡罗树搜索的一个基本过程。这样的过程可以不断重复，一直算到电脑认为最佳为止，或者算到规定下一步必须走子的时间为止。电脑根

据在这之前的所有计算信息综合出一个值来，然后决定下一步在哪落子。

我们现在看来， 人工智能下围棋把世界冠军下赢， 除了电脑计算速度非常快之外， 它的算法中概率统计是离不开的，功不可没！这是概率统计魅力无穷的一个实例。

四、 概率统计在 DNA 序列分析中的应用.

下面讲讲概率统计在 DNA 序列分析中的应用。这部分内容与我们目前的研究方向有关。 今年 7 月 3 号我在上海财大举行的国际生物统计中国分会做了大会报告， 下面我将取自那里的一些材料， 来说明概率统计的作用。 这几年我们做应用， 一方面与微软合作， 另一方面与生物学家合作。 我们一直在念 Rick Durrett 的《Probability Model and DNA Sequence Evolution》 和杨子恒最近的一本书《Molecular Evolution: A Statistical Approach》（2014 年出版）。 这一学期， 我们学生都在念他这本书。去年， 北京召开了国际工业与应用数学大会，我是大会程序委员会主席，挑选了 27 个大会报告。同时， 我和杨子恒共同组织了一个小的 Symposium《Mathematics in Population Genetics and Evolution》, 其主题有下面的一段话：“This symposium will focus on probabilistic modeling and statistical analysis of modern genetic and genomic data, and the statistical and computational challenges that we face。” 随着当代基因和基因组数据的迅速增加，DNA 序列分析越来越需要生物学、数学、统计学和计算机科学的共同

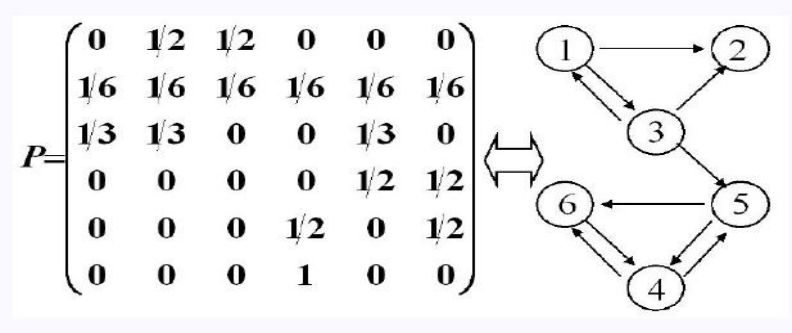
参与和交叉合作。这方面研究成果也很多， 也很活跃。近年来我们研究组与中科院基因组所、上海马普生物研究所等单位的生物学家合作，也做了一些研究工作。我们的研究成果包括：基于同源一致片段推断人口迁移历史 ，基于祖先片段推断人口混合历史，带有重组的溯祖新模型，等等 。另外，我的学生朱天琪与杨子恒等人用真实的DNA 数据，结合化石提供的校准区间信息，估计生物进化的时间。他们最主要方法是概率统计的贝叶斯分析，由于改进了贝叶斯分析的初始分布，他们得到相对准确的哺乳类动物的分化年代。这些结果都充分展示了概率统计的魅力。

五、 搜素引擎中的概率统计.

概率统计与信息领域的交叉也是一个非常有说服力展示概率统计魅力的例子。我前些年在各地作公众报告时经常讲这个例子。

这是我在 Google 中搜寻中国科学院出现的页面。 页面上标记有 874 万条结果，用时 0.15 秒。 计算机很聪明，并没有把 874 万条结果不排序地全部列出，而是把最重要、 最相关的结果排在前面。 计算机怎么会识别哪些结果比较重要，哪些结果比较不重要呢？它能读懂这些页面的内容，然后根据内容来确定页面的重要性吗？显然

不可能，现在的计算机还没有发展到那么先进。实际上很多搜索引擎公司做的一件主要的事，就是网页的排序。 网页排序包括重要性排序和相关性排序，都要用到概率统计。 相关性排序我今天可能没时间讲，我就讲讲网页的重要性排序，下面我用概率论和马氏过程理论来说明网页重要性排序的原理。



这里右边是我们的互联网，当然里面有上万上亿个网页，为了能够说明清楚，这里就假定我们有 6 个网页。 假如你现在浏览页面 1，页面 1 有两个超链接，一个指向 2,一个指向 3，下一步你很可能点一个超链接就到页面 2，或另一个超链接到页面 3，也就是说从页面 1 出发，可能有 1/2 的概率到页面 2, 1/2 的概率到页面 3。同样的道理假如从页面 3 出发，页面 3 有三个超链接，所以在浏览页面 3 的时候，可能有 1/3 的概率到页面 1, 1/3 的概率到页面 2, 1/3 的机率到页面 5，以此类推。 如果你现在浏览的页面没有向外的超链接，比如页面 2，那么在浏览页面 2 时，下一步也许有相同的概率到任何一个其它页面。 当然我这样描述的上网动作并不全面，因为你也可能不顺着超链接到下一个页面，而是通过输入一个关键词或者是一个网址进入下一个页面。假定有概率 α 顺着超链接到另外一个页面，同时有 1−α 的概率通过输入一个网址或是关键词去到另外一个

页面，这两个动作综合起来就是我们上网冲浪的动作。 这是两种随机游动组合成的一个随机游动，连续上网冲浪的动作构成一个马氏链，它的转移概率由我们刚才描述的两个上网动作来确定。这是一个不可约马氏链，它有唯一平稳分布。 Google 把马氏链的平稳分布称作 PageRank，并以此来为页面重要性排序。 一个页面的 PageRank 值越高，即平稳分布在一个页面的值越大，就认为这个页面越重要。用概率的理论上可以严格证明，平稳分布在一个页面的值正好等于点击这个页面的平均访问率，所以用这个值来为页面的重要性排序很合理。不可约马氏链的平稳分布在计算机上运用迭代法容易实现。但由于互联网的规模很大，实际计算时也需要很长时间。 这种计算页面重要性的算法出自 1998 年就读斯坦福大学 (Stanford University) 的博士研究生 Sergey Brin 与 Larry Page，他们把这个算法称作 PageRank 算法，并且编写了一个 PageRank 搜寻工具。 他们发现，网络越大，链接越多，这个引擎提供的结果就越准确。于是，他们将新引擎命名为 Google，这是 Googol 的变体，Googol 是一个数字名词，表示 10 的 100 次方。 Brin 与 Page 于 1998 年在第七次国际 World Wide Web 会议 (WWW98) 上公布他们的论文“The Page Rank citation ranking:Bringing order to the Web” 时，正在用自己的宿舍作为办公室初创产业，这一产业后来发展为庞大的 Google 公司，Brin 和 Page 现在已跻身世界上最有钱的人之列。 PageRank 算法是信息检索领域里一个革命性的发现，这个在信息检索领域看似很困难的问题，用一个马氏链就能就解决了，概率统计的用处有时真是不可估

量。我还要补充强调一下，现在各搜索引擎公司对页面的排序，除了用到 PageRank 算法，或类似于 PageRank 算法提供的重要性排序外，还要考虑相关性排序和诸多其它因素。

从 1998 年到现在，Google 的 PageRank 算法作为网页排序的优点已经充分显示，而缺点也逐渐地暴露出来，最大的缺点是它只利用了页面结构，没有考虑网络用户的感情。 其实现在有很多的垃圾页面，它的 PageRank 可以排得很高。 甚至有些 SPAM 公司，自己搞个服务器，让许多页面互相连结，如果对方给钱，公司就将你的页面连结上去，从而恶意提高页面排序。 这个问题，特别是在前几年，成为搜索引擎公司非常关注的问题，怎么样能够克服这个缺点，当时很多搜索引擎公司都在做。 我们跟微软亚洲研究院在这个问题上也有些合作的关系。当时是这样开始的，记得大概是 2005 年吧，我那时候对随机复杂网络感兴趣，办了一个随机复杂网路的讨论班。 微软亚洲研究院的一位年轻工作人员来找我，想请教我一些问题。我借此请他在我们讨论班作报告，他向我们介绍了 Google 的故事。以后我们跟微软亚洲研究院开始合作，我的学生也到微软作实习生，共同培养人才。有一次，一位年轻的研究员和我的学生一起来找我，把用户上网纪录数据拿给我看，问我由这些数据，能不能够判断出页面的重要性，或着说能不能挖掘出什么样的讯息来。 我们坐下来开始想这个能做什么用。 当然我们是学概率的，所以我们就想到这是个随机过程，它不是确定性的，当然它也是跳过程，一跳一跳的。 我们猜想其中比较关键的是，在这个页面上你下一步到哪个页面去，或者你

在这个页面上停留多少时间，这些在很大程度上，只跟页面的内容有关，而跟你以前访问过哪些页面无关。因此作为一阶近似，这个过程很可能是一个马氏过程，它将来的发展只与现在有关，跟过去无关。另一个想法，你上午看这个页面或下午看这个页面，你的动作可能差不多，所以还应该是时间齐次的。所以当时我们就分析，也许可以把所有人群上网的动作，近似的看作是一个时间齐次的马氏跳过程。当然，要判断它是不是时间齐次马氏跳过程，要用到概率知识，假如真的是时间齐次马氏过程，那么用户在一个页面停留的时间，应该是负指数分布，这是马氏过程理论的一个基本结果。我们建议微软把他们的数据拿来检验一下，于是微软亚洲研究院的相关研究组用真实资料作了大量实验模拟，由我当时在微软实习的学生刘玉婷设计算法，发现用户在网页的停留时间基本服从负指数分布。这个分析出来之后，我们相信可以用马氏过程来研究上网动作，微软亚洲研究院成立了一个小组主攻这个项目，刘玉婷当时作为微软的实习生也在这个研究小组。 这个研究小组做得非常好，在微软相关研究员的带领下，他们克服了种种难关，每一步都在课题组内反复论证，深入探讨，反复模拟实验。 这里面含有许多奇思构想和巧妙的数学。微软亚洲研究院从产品部门调来大量数据，做了大规模模拟实验。2008 年 7 月，在新加坡召开的的第 31 届国际信息检索大会上，刘玉婷报告了他们的论文:《浏览排序: 让因特网使用者为页面重要性投票》，论文获得了会议设立的唯一最佳学生论文奖。 这篇文章，据说他们修改了八十一次，在新加坡得奖之后，“Browse Rank” 成了业内的热

门话题。 最热的时候，输入关键词 Browse Rank 有 157,000,000 个结果。 当时网页的文章，有的题目是 “Browse Rank vs Page Rank”，有的说 “Microsoft Lauches Browse Rank To Compete With Page Rank”，还有 “Live Search is researching a ranking feature similar to Google’s Page Rank called Browse Rank”，等等。 网上还有一个以“Browse Rank the next PageRank” 为题目的视频介绍微软亚洲的研究人员开发的 Browse Rank。 这是前几年的事，当然了，一个新产品的开发还与许多其它因素有关，现在也没有 Browse Rank 出现，但是说明当时这个工作在讯息检索领域引起了一些关注。 我们与微软现在还有合作，现在我还有学生在微软，已经是正式的员工。从做科学研究的角度来说，我们感到高兴的是我们第一个用 Browsing Process 刻画了真实的用户上网行为。 我相信今后人们在研究用户上网行为时，一定会想到 Browsing Process，应用并发展 Browsing Process 的理论和实践。 上面说到我们发现用户上网的一阶近似可以用马氏过程来刻画，后来我们又有进一步发挥，在这个基础上提出了 web 马氏骨架过程，之所以提出 web 马氏骨架过程，是因为后来研究手机网的搜索引擎时，发现它不完全是马氏过程，最多可以算是 web 马氏骨架过程，也就是说它有一个骨架是马氏的，而它的等待时间不仅依赖当前页面，还依赖以前的页面。 由于手机上面网页的超链接，跟一般普通网页超级连接的设计不一样。

附记：本文由蒋继发根据马志明的公众报告录音整理，并经马志明修

改。
