Homework 12

孙锴

June 11, 2012

练习(7.21)**.** 设*A*与*B*为两个长度为*n*的由*0*到*9*组成的随机序列。
在$n \to \infty$情况下，所有长度为*k*的子序列可以近似看为相互独立，因此可将其分布看为近似的高斯分布。
设x为$|A \cup B|$的期望值，则有
$1 + \frac{10^k}{10^k-1} + \frac{10^k}{10^k-2} + \ldots + \frac{10^k}{10^k-(x-1)} = 2n$
解得$x = 10^k - \frac{10^k}{e^{\frac{2n}{10^k}}} + 1$。
设y为$|A|$的期望值，则有
$1 + \frac{10^k}{10^k-1} + \frac{10^k}{10^k-2} + \ldots + \frac{10^k}{10^k-(x-1)} = n$
解得$x = 10^k - \frac{10^k}{e^{\frac{n}{10^k}}} + 1$。
同理$|B|$的期望值也为y。
则$\frac{|A\cap B|}{|A\cup B|} = \frac{|A|+|B|-|A\cup B|}{|A\cup B|} = \frac{2y-x}{x} = \frac{1-\frac{2}{e^{\frac{n}{10^k}}}+\frac{1}{e^{\frac{2n}{10^k}}}+\frac{1}{10^k}}{1-\frac{1}{e^{\frac{2n}{10^k}}}+\frac{1}{10^k}}$ *(*)*
更进一步，类比*increasing property*，不难证明当$n \to \infty$时，存在且仅存在一个界点（*threshold*），使得在渐进意义比界点小的点相似度为*0*，比界点大的点相似度为*1*。下面通过计算证明界点为$k = lg(\frac{n}{ln3})$。
用$t = \frac{1}{e^{\frac{n}{10^k}}}$换元，则*(*)*式$= \frac{(1-x)^2+\frac{1}{10^k}}{1-x^2+\frac{1}{10^k}}$。
令$\frac{(1-x)^2+\frac{1}{10^k}}{1-x^2+\frac{1}{10^k}} = \frac{1}{2}$ *(**)*，断言$n \to \infty$且*(**)*式成立时$\frac{1}{10^k} \to 0$。下面令$\frac{1}{10^k} = 0$，则证明断言成立只须证明此时确有$\frac{1}{10^k} \to 0$。
由于此时方程化为一二次方程$3t^2 - 4t + 1 = 0$，因此不难解得$t_1 = \frac{1}{3}$与$t_2 = 1$（舍）。从而得到$k = lg(\frac{n}{ln3})$，易见此时$\frac{1}{10^k} \to 0$，从而断言成立，进而证明了界点为$k = lg(\frac{n}{ln3})$。当在渐进意义下$lg(\frac{n}{ln3}) < k$时相似度为*0*，$lg(\frac{n}{ln3}) > k$时相似度为*1*。通过程序模拟验证，印证了以上计算的界点是正确的。
综上，$\frac{|A\cap B|}{|A\cup B|} = \frac{1-\frac{2}{e^{\frac{n}{10^k}}}+\frac{1}{e^{\frac{2n}{10^k}}}+\frac{1}{10^k}}{1-\frac{1}{e^{\frac{2n}{10^k}}}+\frac{1}{10^k}}$，且当在渐进意义下$lg(\frac{n}{ln3}) < k$时相似度为*0*，$lg(\frac{n}{ln3}) > k$时相似度为*1*。

练习(7.22). 不难看出，当序列不再完全随机时，序列的分布更集中于某一部分序列（即随机序列集的一个子集），于是两个序列相似的概率将增大，即结果将变得更糟糕。

练习(7.23). 根据对题目的两种不同的理解，有以下两种解答：
*(i)*若是计算随机序列中存在冲突的概率，即存在至少两个长度为*k-1*的子序列相同，则有以下解：
(a)$P \approx 1 - \Pi_{i=1}^{9997} \frac{100^{3-1}-i}{100^{3-1}}$，通过近似计算可以得$\Pi_{i=1}^{9997} \frac{100^{3-1}-i}{100^{3-1}} \approx 0$，从而$P \approx 1 - 0 = 1$
(b)$P \approx 1 - \Pi_{i=1}^{9995} \frac{100^{5-1}-i}{100^{5-1}}$，通过近似计算可以得$\Pi_{i=1}^{9995} \frac{100^{5-1}-i}{100^{5-1}} \approx 0.6$，从而$P \approx 1 - 0.6 = 0.4$
*(ii)*若是计算对于某一个给定的长度为*k*的子序列，计算与其它长度为*k*的子序列存在冲突的概率，则有以下解：
(a)$P \approx 1 - (\frac{100^{3-1}-1}{100^{3-1}})^{9996} \approx 0.63$
(b)$P \approx 1 - (\frac{100^{5-1}-1}{100^{5-1}})^{9994} \approx 0$

练习(7.24). 按照*7.21*所推结论取定合适的*k*。每篇论文可以看成一个字符串，用它的所有长度为*k*的子串的集合代表这篇论文。设*A*与*B*分别为两篇论文的代表集合，则两篇文章的相似度为$\frac{|A \cap B|}{|A \cup B|}$，设两篇文章的相似度$\geq$ *l*时定为抄袭，则*A*与*B*对应的论文存在抄袭当且仅当$\frac{|A \cap B|}{|A \cup B|} \geq l$。

练习(7.25). 设网页数为*n*，网页的最大长度为*s*，按照*7.21*所推结论和所需的精确度取定合适的*k*和*m*（*m*的意义见下文）。每个网页可以看成一个字符串，用它的字典序前*m*小的长度为*k*的子串的集合代表这个网页（复杂度为$O(nslogs)$），则这个集合可以看成是这个网页的一个很好的哈希值。从而可以认为在*k*与*m*设定的精度范围内，两个网页相同当且仅当代表它们的集合相同。因此可以将所有网页按照代表他们的集合以复杂度$O(nmlogn)$排序（准确地说是分类，*sort*），然后可以用$O(nm)$的复杂度扫描一遍分类后的网页，完成去重工作。因此总时间复杂度为$O(nslogs + nmlogn)$。（如果采用线性复杂度的排序算法，则总时间复杂度可以降为$O(ns + nm)$）

练习(7.27). 首先，不难推得*k*取决于文章中的最长重复字串（可重叠）。而歌词中最长重复字串为“*you'll never walk alone*”，有*24*个字符，从而得到$k = 24 + 2 = 26$。