《空间向量及其坐标表示》课后阅读材料：余弦定理和新闻的分类

选自吴军《数学之美》

序言

吴军是自然语言处理和搜索技术领域的重要华人学者，早年毕业于清华大学，后获美国约翰斯·霍普金斯大学计算机科学博士学位，曾在 Google 从事搜索、自然语言处理和反垃圾邮件等相关研究，后来也曾任腾讯负责搜索业务的副总裁。《数学之美》最初源于他在 Google 黑板报上连载的科普文章，写作目的并不是把数学包装成抽象技巧，而是说明统计、线性代数、信息论、概率论等工具如何真正进入搜索引擎、自然语言处理、机器翻译和新闻分类等工程系统。本文选段正体现了这种写法：从中学熟悉的余弦定理出发，过渡到空间向量、坐标表示、内积与文本相似度计算，使读者看到“空间向量”并不只存在于几何题中，也可以成为信息处理的基础语言。

世界上有些事情常常超乎人们的想象。余弦定理和新闻的分类看似八竿子打不着，却有着紧密的联系。具体来说，新闻的分类很大程度上依靠的是余弦定理。

2002 年夏天，Google 推出了自己的“新闻”服务。和传统媒体的做法不同，这些新闻不是记者写的，也不是人工编辑的，而是由计算机整理、分类和聚合各个新闻网站的内容，一切都是自动生成的。这里面的关键技术就是新闻的自动分类。

1 新闻的特征向量

所谓新闻的分类，或者更广义地讲任何文本的分类，无非是要把相似的新闻归入同一类中。如果让编辑来对新闻分类，他一定会先读懂新闻，然后找出其主题，最后根据主题的不同对新闻进行分类。但是计算机根本读不懂新闻，虽然一些商业人士和爱炫耀才学的计算机专家宣称计算机能读懂新闻。计算机本质上只能做快速计算。为了让计算机能够“算”新闻（而不是读新闻），就要求我们先把文字的新闻变成一组可计算的数字，然后再设计一个算法来算出任意两篇新闻的相似性。

首先让我们看看怎样找一组数字（或者说一个向量）来描述一篇新闻。新闻是传递信息的，而词是信息的载体。新闻的信息和词的语义是联系在一起的。套用俄罗斯文豪托尔斯泰在《安娜·卡列尼娜》开篇的那句话来讲：“同一类新闻用词都是相似的，不同类的新闻用词各不相同。”当然，一篇新闻有很多词，有些词表达的语义重要，有些词相对次要。那么如何确定哪些重要，哪些次要呢？首先，直觉告诉我们含义丰富的实词一定比“的、地、得”这些助词，或者“之乎者也”这样的虚词重要，这是肯定的。接下来，需要进一步对每个实词的重要性进行度量。回忆一下我们在“如何确定网页和查询的相关性”一章中介绍的单文本词汇频率/逆文本频率值 TF-IDF 的概念，在一篇文章中，重要的词 TF-IDF 值就高。不难想象，和新闻主题有关的那些实词频率高，TF-IDF 值很大。

现在我们找到了这样一组来描述新闻主题的数字：对于一篇新闻中的所有实词，计算出它们的 TF-IDF 值。把这些值按照对应的实词在词汇表的位置依次排列，就得到一个向量。比如，词汇表中有 64000 个词，其编号和词如表 14.1 所示；在某一篇特定的新闻中，这 64000 个词的 TF-IDF 值分别如表 14.2 所示。

表 1: 统计词汇表

单词编号	汉字词
1	阿
2	啊
3	阿斗
4	阿姨
···	···
789	服装
···	···
64000	做作

表 2: 某篇新闻对应的 TF-IDF 值

单词编号	TF-IDF 值
1	0
2	0.0034
3	0
4	0.00052
···	···
789	0.034
···	···
64000	0.075

如果单词表中的某个词在新闻中没有出现，对应的值为空，那么这 64000 个数，组成一个 64000 维的向量。我们就用这个向量来代表这篇新闻，并称为新闻的特征向量（Feature Vector）。每一篇新闻都可以对应这样一个特征向量，向量中每一个维度的大小代表每个词对这篇新闻主题的贡献。当新闻从文字变成了数字后，计算机就有可能“算一算”新闻之间是否相似了（见图 14.1）。

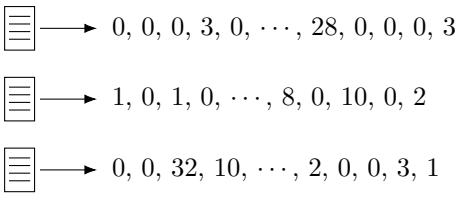


图 1: 一篇篇文章变成了一串串数字

2 向量距离的度量

世界各国无论是哪门语言的“语文课”（Language Art），老师教写作时都会强调特定的主题要用特定的描述词。几千年来，人类已经形成了这样的写作习惯。因此，同一类新闻一定是某些主题词用得较多，另外一些词则用得较少。比如金融类的新闻，这些词出现的频率就很高：股票、利息、债券、基金、银行、物价、上涨。而这些词出现的就较少：二氧化碳、宇宙、诗歌、木匠、诺贝尔、包子。反映在每一篇新闻的特征上，如果两篇新闻属于同一类，它们的特征向量在某几个维度的值都比较大，而在其他维度的值都比较小。反过来看，如果两篇新闻不属于同一类，由于用词的不同，在它们的特征向量中，值较大的维度应该没有什么交集。这样我们就定性地认识到，两篇新闻的主题是否接近，取决于它们的特征向量“长得像不像”。当然，我们还需要定量地衡量两个特征向量之间的相似性。

学过向量代数的人都知道，向量实际上是多维空间中从原点出发的有向线段（见图 14.2）。

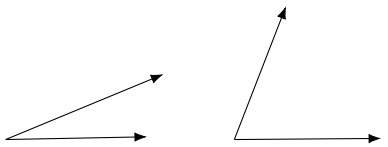


图 2: 向量的夹角是衡量两个向量相近程度的度量。图中左边的两个向量距离较近，右边两个向量距离较远

不同的新闻，因为文本长度的不同，它们的特征向量各个维度的数值也不同。一篇 10000 字的文本，各个维度的数值都比一篇 500 字的文本来得大，因此单纯比较各个维度的大小并没有太大意义。但是，向量的方向却有很大的意义。如果两个向量的方向一致，说明相应的新闻用词的比例基本一致。因此，可以通过计算两个向量的夹角来判断对应的新闻主题的接近程度。而要计算两个向量的夹角，就要用到余弦定理了，比如图 14.2 中，左边两个向量的夹角小，距离就较“近”，相反，右边两个向量的夹角大，距离就“远”。

我们对余弦定理都不陌生，它描述了三角形中任何一个夹角和三个边的关系。换句话说，给定三角形的三条边，可以用余弦定理求出三角形各个角的角度（见图 14.3）。假定三角形的三条边为 a, b, c，对应的三个角为 A, B, C。

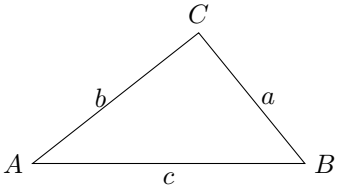


图 3: 余弦的计算

那么 $\angle A$ 的余弦

$$\cos A = \frac{b^2 + c^2 - a^2}{2bc} \tag{1}$$

如果将三角形的两边 b 和 c 看成是两个以 A 为起点的向量，那么上述公式等价于

$$\cos A = \frac{\langle b, c\rangle}{|b| \cdot |c|} \tag{2}$$

其中，分母表示两个向量 b 和 c 的长度，分子表示两个向量的内积。举一个具体的例子，假如新闻 X 和新闻 Y 对应的向量分别是

$$x_1, x_2, \cdots, x_{64000} \quad 和 \quad y_1, y_2, \cdots, y_{64000},$$

那么它们夹角的余弦等于

$$\cos\theta = \frac{x_1y_1 + x_2y_2 + \cdots + x_{64000}y_{64000}}{\sqrt{x_1^2 + x_2^2 + \cdots + x_{64000}^2} \cdot \sqrt{y_1^2 + y_2^2 + \cdots + y_{64000}^2}}. \tag{3}$$

由于向量中的每一个变量都是正数，因此余弦的取值在 0 和 1 之间，也就是说夹角在 0 度到 90 度之间。当两条新闻同时关联的余弦等于 1 时，这两个向量的夹角为零，两条新闻完全相同；当夹角的余弦接近于 1 时，两条新闻相似，从而可以归成一类；夹角的余弦越小，夹角越大，两条新闻越不相关。当两个向量正交时（90 度），夹角的余弦为零，说明两篇新闻根本没有相同的主题词，它们毫不相关。

现在把一篇篇文字的新闻变成了按词典顺序组织起来的数字（特征向量），又有了计算相似性的公式，就可以在此基础上讨论新闻分类的算法了。具体的算法分为两种情形。第一种比较简单，假定我们已知一些新闻类别的特征向量 $x_1, x_2, \cdots, x_k$，那么对于任何一个要被分类的新闻 Y，很容易计算出它和各类新闻特征向量的余弦相似性（距离），并将其归入它该去的那一类中。至于这些新闻类别的特征向量，既可以手工建立（工作量非常大而且不准确），也可以自动建立（以后会介绍）。第二种情形就比较麻烦了，即如果事先没有这些新闻类别的特征向量怎么办。我们在纽约时，霍普金斯大学的同学弗洛里安（Radu Florian）和教授雅罗斯基给出了一个自底向上不断合并的方法，大致思想如下。

1. 计算所有新闻之间两两的余弦相似性，把相似性大于一个阈值的新闻合并成一个小类（Subclass）。这样 N 篇新闻就被合并成 N_1 个小类，当然 $N_1 < N$。
2. 把每个小类中所有的新闻作为一个整体，计算小类的特征向量，再计算小类之间两两的余弦相似性，然后合并成大一点的小类，假如有 N_2 个，当然 $N_2 < N_1$。

这样不断做下去，类别越来越少，而每个类越来越大。当某一类太大时，这一类里一些新闻之间的相似性就很小了，这时就要停止上述迭代的过程了。至此，自动分类完成。图 14.4 是弗洛里安给出的一个真实的文本分类迭代和聚合的过程，图中左边的每一个点都代表一篇文章，因为数量太多，密密麻麻地连成了一片。每一次迭代，子类数量就不断减少，当子类的数量比较少时，就会看清楚这些子类了。

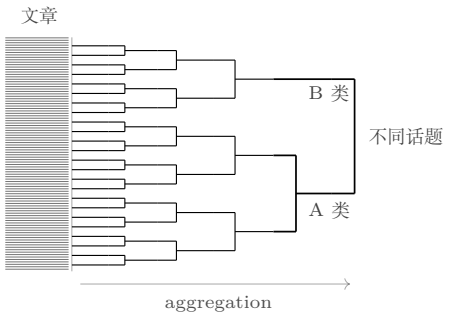


图 4: 真实的文本分类聚合过程

当然，这里面有很多的技术细节，有兴趣和基础的读者可以读他们的论文。

弗洛里安和雅罗斯基 1998 年做这项工作的动机很有意思。当时雅罗斯基是某个国际会议的程序委员会主席，需要把提交上来的几百篇论文发给各个专家评审以决定是否录用。为了保证评审的权威性，需要把每个研究方向的论文交给这个方向最有权威的专家。虽然论文的作者自己给论文定了方向，但是范围太广，没有什么指导意义。雅罗斯基当然没有时间浏览这近千篇论文，然后再去分发，于是就想到了这个将论文自动分类的方法，由他的学生弗洛里安很快实现了。接下来几年的会议都是这么选择评审专家的。从这件事我们可以看出美国人做事的一个习惯：美国人总是倾向于用机器（计算机）代替人工来完成任务。虽然在短期内需要做一些额外的工作，但是从长远看可以节省很多时间和成本。

余弦定理就这样通过新闻的特征向量和新闻分类联系在一起了。我们在中学学习余弦定理时，恐怕很难想象它可以用来对新闻进行分类。在这里，我们再一次看到数学工具的用途。

3 延伸阅读：计算向量余弦的技巧

读者背景知识：数值分析。

3.1 大数据量时的余弦计算

利用公式（14.2）计算两个向量夹角时，计算量为 $O(|a|+|b|)$。如果限定其中一个向量更长，不失一般性 $|a| > |b|$，这样复杂度为 $O(|a|)$。如果要比较一篇新闻和所有其他 N 篇新闻的相关性，那么计算复杂度为 $O(N \cdot |a|)$；如果要比较所有 N 篇新闻之间两两的相关性，计算复杂度为 $O(N^2 \cdot |a|)$。注意，这还只是一次迭代。因此，这个计算量是很大的。我们假定词汇表的大小为 10 万，那么向量的长度也是这么大，假定需要分类的新闻为 10 万篇，总的计算量在 10^{15} 这个数量级。如果用 100 台服务器，每台服务器的计算能力是每秒 1 亿次，那么每次迭代的计算时间在 10 万秒，即大约 1 天。几十次迭代就需要两三个月，这个速度显然很慢。

这里面可简化的地方非常多。首先，分母部分（向量的长度）不需要重复计算，计算向量 a 和向量 b 的余弦时，可以将它们的长度存起来，等计算向量 a 和向量 c 的余弦时，直接取用 a 的长度即可。这样，上面的计算量可以节省 2/3，也就是只有公式（14.2）中分子的部分需要两两计算。当然这还没有从根本上降低算法的复杂度。

其次，在计算公式（14.2）中的分子即两个向量内积时，只需考虑向量中的非零元素。计算的复杂度取决于两个向量中非零元素个数的最小值。如果一篇新闻的一般长度不超过 2000 词，那么非零元素的个数一般也就是 1000 词左右，这样计算的复杂度大约可以下降 100 倍，计算时间从“天”这个量级降至十几分钟这个量级。

第三，可以删除虚词，这里的虚词包括搜索中的非必留词，诸如“的”“是”“和”，以及一些连词、副词和介词，比如“因为”“所以”“非常”等等。我们在上一节中分析过，只有同一类新闻，用词才有很大的重复性，不同类的新闻用词不大相同。因此，去掉这些虚词后，不同类的新闻，即使它们的向量中还有不少非零元素，但是其相同的非零元素并不多。要做到乘法并不多。大多数情况下，都可以跳过去（因为非零元素乘零，结果还是零）。这样，计算时间还可以缩短几倍。因此，10 万篇新闻两两比较一下，计算时间也就是几分钟而已。如果做几十次迭代，可以在一天内计算完。

需要特别指出的是，删除虚词，不仅可以提高计算速度，对新闻分类的准确性也大有好处，因为虚词的权重其实是一种噪声，干扰分类的正常进行。这一点与通信中过滤掉低频噪声是同样的原理。通过这件事，我们也可以看出自然语言处理和通信的很多道理是相通的。

3.2 位置的加权

和计算搜索相关性一样，出现在文本不同位置的词在分类时的重要性也不相同。显然，出现在标题中的词对主题的贡献远比出现在新闻正文中的重要。而即使在正文中，出现在文章开头和结尾的词也比出现在中间的词重要。在中学学习语文和大学学习英语文学时，老师都会强调这一点——阅读时要特别关注第一段和最后一段，以及每个段落的第一个句子。自然语言处理中文本规律依然有用。因此，要对标题和重要位置的词进行额外的加权，以提高文本分类的准确性。

小结

本章介绍的这种新闻归类的方法，准确性很好，适用于被分类的文本集合在百万数量级。如果大到亿这个数量级，那么计算时间还是比较长的。对于更大规模的文本处理，我们将在下一章中介绍一种更快速但相对粗糙的方法。