余弦相似性

维基百科，自由的百科全书

余弦相似性通过测量两个向量的夹角的余弦值来度量它们之间的相似性。0度角的余弦值是1，而其他任何角度的余弦值都不大于1；并且其最小值是-1。从而两个向量之间的角度的余弦值确定两个向量是否大致指向相同的方向。两个向量有相同的指向时，余弦相似度的值为1；两个向量夹角为90°时，余弦相似度的值为0；两个向量指向完全相反的方向时，余弦相似度的值为-1。這結果是與向量的長度無關的，仅仅與向量的指向方向相關。余弦相似度通常用于正空間，因此給出的值为0到1之间。

注意這上下界对任何维度的向量空間中都適用，而且余弦相似性最常用於高维正空间。例如在信息检索中，每个词項被賦予不同的維度，而一个文档由一个向量表示，其各個維度上的值對應于該词項在文档中出现的频率。余弦相似度因此可以给出两篇文档在其主题方面的相似度。

另外，它通常用于文本挖掘中的文件比较。此外，在数据挖掘领域中，會用到它来度量集群内部的凝聚力。[1]

目录

定义

两个向量间的余弦值可以通过使用欧几里得点积公式求出：

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos\theta$$

給定两个属性向量， A 和B，其余弦相似性θ由点积和向量長度給出，如下所示：

$$\mathbf{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$ ，這裡的A_i和B_i分别代表向量A和B的各分量。

給出的相似性范围从-1到1：-1意味着两个向量指向的方向正好截然相反，1表示它们的指向是完全相同的，0通常表示它们之间是独立的，而在这之间的值则表示中間的相似性或相异性。

对于文本匹配，属性向量A和B通常是文档中的词频向量。余弦相似性，可以被看作是在比较過程中把文件长度正規化的方法。

在信息检索的情况下，由于一个词的频率（TF-IDF权）不能为负数，所以这两个文档的余弦相似性范围从0到1。并且，两个词的频率向量之间的角度不能大于90°。

角相似性

「余弦相似性」一詞有时也被用来表示另一個系数，儘管最常见的是像上述定义那样的。透過使用相同計算方式得到的相似性，向量之间的规范化角度可以作为一个范围在[0,1]上的有界相似性函数，從上述定义的相似性计算如下：

$$1 - \left(\frac{\cos^{-1}(\mathbf{similarity})}{\pi} \right)$$

這式子適用於向量系数可以為正或負的情況。

或者，用以下式子計算

$$1 - \left(\frac{2 \cdot \cos^{-1}(\mathbf{similarity})}{\pi} \right)$$

这式子適用於向量系数总為正的情況。

虽然「余弦相似性」一词有時會用來表示這個角距离，但實際上很少這樣說，因為角度的餘弦值只是作为一种计算角度的简便方法而被用到，本身并不是意思的一部分。角相似系数的优点是，当作为一个差异系数（从1减去它）时，产生的函數是一个嚴格距离度量，而對於第一種意義的「余弦相似性」則不然。然而，对于大多数的用途，这不是一个重要的性質。若对于某些情况下，只有一组向量之間的相似性或距离的相对顺序是重要的，那么不管是使用哪個函数，所得出的顺序都是一樣的。

與「Tanimoto」系数的混淆

有时，余弦相似性會跟一種特殊形式的、有著类似代数形式的相似系数相混淆：

$$T(A,B) = \frac{A \cdot B}{\|A\|^2 + \|B\|^2 - A \cdot B}$$

事实上，这个代数形式是首先被Tanimoto定义，作為在所比較集合由位元向量表示時计算其Jaccard系数的方法。虽然這公式也可以扩展到向量，它具有和余弦相似性頗為不同的性质，并且除了形式相似外便沒有什麼關係。

Ochiai系数

这个系数在生物学中也叫Ochiai系数，或Ochiai-Barkman系数[2][3]：

$$K = \frac{n(A \cap B)}{\sqrt{n(A) \times n(B)}}$$

這裡A和B是集合，$n(A)$是A的元素個數。如果集合由位元向量所代表，那麼可看到Ochiai系数跟餘弦相似性是等同的。

另见

- Sorensen相似性指數
- 汉明距离
- 相关
- Dice系数
- Jaccard指数
- SimRank
- 信息检索

外部链接

- 加权的余弦措施 (http://mathforum.org/kb/message.jspa?messageID=5658016&tstart=0)
- http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html#Cosim

参考文献

1. P.-N. Tan, M. Steinbach & V. Kumar, "Introduction to Data Mining", , Addison-Wesley (2005), ISBN 0-321-32136-7, chapter 8; page 500.
2. *Ochiai A.* Zoogeographical studies on the soleoid fishes found Japan and its neighboring regions. II // Bull. Jap. Soc. sci. Fish. 1957. V. 22. № 9. P. 526-530.
3. *Barkman J.J.* Phytosociology and ecology of cryptogamic epiphytes, including a taxonomic survey and description of their vegetation units in Europe. – Assen. Van Gorcum. 1958. 628 p.