**向量空间模型中基于余弦和Jaccard相似性度量的信息检索**

1、阿布舍克·贾因计算机科学系，巴拉迪·维迪亚佩斯工程学院

2、阿曼·詹：印度理工学院计算机科学系

3、印度理工学院尼哈尔·乔汉计算机科学系

4、Vikrant Singh计算机科学系，Bharati Vidyapeeth工程学院

5、纳丽娜·塔库尔·霍德，巴拉蒂·维迪亚佩斯工程学院计算机科学系

**摘要**

随着网络上可供我们使用的文档呈指数级增长，对检索与给定搜索查询匹配的最相关文档的有效技术的需求变得至关重要。信息检索领域研究的是从大量数据中检索所需信息的文档相似性问题。人们提出了各种模型和相似性度量来确定两个对象之间的相似程度。本文的目的是总结整个过程，研究一些最著名的算法和方法，将查询文本与一组索引文档进行匹配。

**关键词**：权重度量，TF\/IDF，余弦相似性度量，Jaccard相似性度量，信息检索。

1. **导言**

基于输入查询的文档检索是信息检索的基本形式之一。网络搜索就是这个应用程序的完美例子。为此已经开发了许多算法，这些算法将输入查询与存储的文档或文本片段相匹配，并根据其与给定查询的相似性得分对输出进行排序。

这种算法依赖于将索引文档与单个查询词进行匹配，索引文档维护关于词频率和位置的信息。根据每个文档的相似性值为其分配分数。由于在所述文档中出现的频率较高，查询词在文档中的得分较高。

不同的算法采用不同的方法来分析这种相似性并计算分数。向量空间模型是一种被高度评价和使用的方法。它胜过布尔模型，布尔模型接受布尔查询，并仅基于布尔逻辑将文档与查询匹配，无论文档中是否包含所需的术语。这往往给人的印象太少(≈ 0）或太多（1000）个文档。

向量空间模型以及TF\/IDF加权方案在第2节中有详细说明。有关TF\/IDF方案的最新研究将在相关工作部分进行讨论。第4节解释了TF\/IDF方案中使用的两种相似性度量，即余弦和Jaccard。

1. **向量空间模型**

向量空间模型是一种基于线性代数的简单且最流行的模型，允许根据文档可能的相关性对文档进行排序。该模型将文本对象表示为n维空间中的向量，其中n表示用于构建表示文档的索引的术语数[2]。创建索引需要对文档进行条带化，并以其唯一术语的形式进行分隔。然后可以对文档进行进一步处理，将不同形式的单词简化为一个共同的词干，这有助于提高匹配两个文档的效率。可以对查询项进行加权，以考虑它们的相对重要性，从而得到更好的结果。

**2.1加权方案**

术语频率tftd是术语t在文档d中出现的次数。该术语用于计算documentquery匹配分数。由于文件的相关性不应随着术语的频率线性增加，因此对数函数会减弱其影响。文件d中术语t的对数频率权重定义为：



此外，一个罕见的术语应该比更频繁的术语更重要。对于集合中罕见的查询中的术语，与其他匹配更频繁术语的文档相比，匹配该术语的文档更有用。为此，使用了文档频率。dft是术语t的文档频率，即术语t出现的文档数。根据术语的用途，定义了反向文档频率。



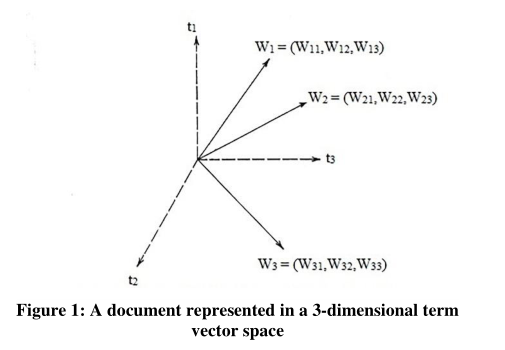
该术语的权重加在一起是其两个权重的乘积：tf权重和idf权重。tf-idf是信息检索中最好的加权方案之一。



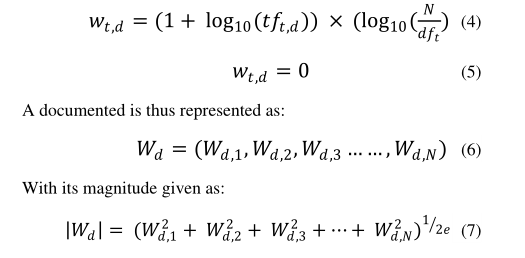
它会随着文档中某个术语出现次数的增加以及查询术语在集合中的稀有程度的增加而增加。

**2.2 代表**

由于tf idf是最佳加权方案，文档随后在n维空间中由tf idf权重的实值向量表示，其中n是用于索引一组文档的不同术语\标记的数量。



如果该项出现在文档中，则维度t中文档d的向量大小由等式（4）给出，否则由等式（5）给出。



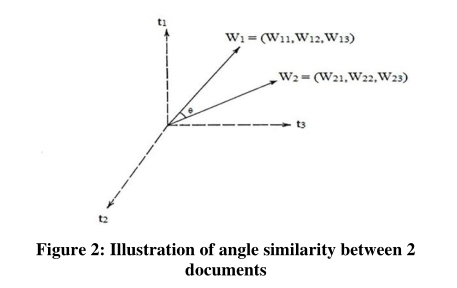
1. **相关工作**

术语权重方案是文本相关信息检索的关键，因此得到了很好的研究。TF\/IDF是最受欢迎的方案之一，其次是Okapi BM25。这一方案的有效性促使研究人员开发了同一方案的修改版本。在下文中，我们将讨论三种这样的修改方案。引入了一种新的项加权方案TF-ATO[6]，即平均项出现次数的项频率，结果表明，它是对TF\/IDF加权方案的改进，TF\/IDF加权方案是向量空间模型的重要组成部分。该方案计算术语在文档中的平均出现次数，并采用区分性方法去除文档中不太重要的权重。当TF-ATO与停止词移除和研究人员建议的鉴别方法结合使用时，其性能优于TF\/IDF方案。提出了另一种改进的TF-IDF方案[7]，该方案结合了相对TF加权和TF归一化，具体取决于文档长度。现有的大多数模型都采用单项频率标准化，这在不同长度的查询中无法很好地平衡短文档和长文档。

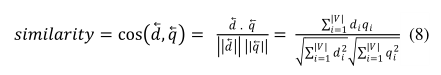
该加权方案采用术语频率归一化的两个方面来确定术语的重要性；一个喜欢短文档，一个喜欢长文档。然后使用查询长度信息组合这两个术语，当排序函数面临不同长度的查询时，查询长度信息在检索长文档和短文档时保持平衡。M.Shirakawa等人[8]已经证明，一个项的IDF等于该项与信息距离空间中的空字符串之间的距离，其中Kolmogorov复杂性是近似的。基于这一发现，他们提出了一种全局术语加权方案N-gram IDF，它是IDF的理论扩展，用于处理任意长度的单词和短语。该方案在不使用任何NLP技术的情况下，在重叠的N-gram中确定主导N-gram，并从文本中提取任意长度的关键项。它使用两种字符串处理技术计算所有可能N-gram的权重，即最大子字符串提取和文档列表。该方案通过设计用于关键词提取和web搜索查询分割的最新方法，取得了具有竞争力的性能。

1. **算法**

**在VSM中，文档集和查询集被视为向量。使用该模型计算向量间相似值的常用方法是向量余弦测度。文档和查询被表示为向量，相似性表示两个向量之间的接近程度（图2）。**



通过欧几里德距离计算这种接近度并不能给出实际的结果。考虑2个文档D1和D2完全相同。在d2后面加上它自己，就得到了一个新的文档d2'。将d1和d2'表示为向量，它们之间的距离很大，尽管它们实际上是相同的。因此，在计算两个向量之间的接近度时，欧几里德距离不是一个很好的度量。为了应对上述情况，将相似度计算为向量角度的函数似乎是最好的选择。如果两个向量很近，它们之间形成的角度就会很小，同样，如果两个向量很远，它们之间形成的角度就会很大。不采用矢量形成的角度，而是考虑角度的余弦值。对于0到180度的角度，余弦值从1到-1不等，这使其成为满足这些要求的理想选择。分数为1表示角度为0o，这意味着文档是相似的。分数为0表示角度为90度，这意味着文档完全不同。余弦加权测度在长度归一化向量上实现，以使其权重具有可比性。方程（8）给出了余弦相似性的公式。



Jaccard相似性度量是计算查询和文档中相似性的另一种度量。在此度量中，索引从最小值0（完全不同）开始，到最大值1（完全相似）。该值由方程式（9）计算得出。



1. **讨论**

向量空间模型是一种基于线性代数的简单模型，旨在克服布尔模型的局限性。与布尔模型相比，VSM的一个主要优点是，分配给术语的权重不是二进制的。通过计算一系列相似性值，从而消除使用布尔模型获得的太少或太多结果，从而实现更好的匹配。另一方面，VSM也有其自身的局限性。主要的局限性在于对语义的敏感性较低。例如，单词“car”和“automobile”不匹配，即两个相同的短语，一个使用“car”，另一个使用“automobile”不会匹配。此外，它不会根据组成词的顺序来区分短语。例如，使用这种方法，“玛丽比约翰快”和“约翰比玛丽快”是无法区分的。余弦和Jaccard是与TF\/IDF加权方案结合使用的两种基本且有效的相似性度量。

1. **总结**

本文简要介绍了一种基本的信息检索模型VSM，它采用TF\/IDF加权方案和余弦和Jaccard相似性度量。在过去的十年里，这个领域进行了大量的研究。这项研究的重点是开发更好的模型，一些是VSM的扩展，如LSA[18]或随机索引[19]，而另一些则基于完全不同的原则，如概率相关性模型[20]。信息的爆炸式增长要求更好的信息检索技术。因此，该领域未来的工作应该集中在开发新的模型、加权方案和相似性度量上，这些模型、加权方案和度量可以有效地利用同一数据集上的语义信息对大型数据集执行。