## **2.2 文本相似度计算**

**文本相似度计算是通过量化的方法计算词语或者文本间的相似度，这项技术被广泛应用在信息检索、文本分类和智能应答等领域。文本相似度计算是一项复杂的任务，目前有很多学者它进行了研究。它可以被分为两个部分：一是如何对文本进行表示，二是选择适当的相似度计算方法。**

### **2.2.1 向量空间模型**

  目前文本表示最常用的是向量空间模型(vector space model, VSM) 。VSM由G. Salton等人在20世纪60年代末期提出，是自然语言处理使用的主流模型。VSM把文本信息用向量空间中的点表示，用向量来表示文本，从而简化了文本中的关键词之间的复杂关系，也使得模型具备了可计算性。

VSM算法的基本原理可表示为：每一篇文档都可以表示成公式(2.1)的形式。

 (2.1)

在这里，表示文档第个特征词条，则表示第个特征词所对应的权重。特征词权重的计算是VSM算法的核心内容，在计算特征权重时，应当考虑特征词条代表文档特征信息的能力，能力越强，赋予的越大的权重值。最简单的文档向量表示方法是使用词袋模型（Bag of Words，BOW），使用BOW表示文档时，文档向量的长度就是文档中出现的词的数量，文档向量每一维的值为该词在文档中出现的频率。

计算文本的相似度时，可以使用欧氏距离、向量夹角余弦公式和向量内积等进行度量，还有一些方法通过对关键词加权重的方式表示文本向量[24]。

**VSM**假设词与词间线性无关，**不考虑文档中词或短语间的位置与顺序关系，这极大简化了计算文档特征权重时的复杂性，但也失去了许多关于文档结构和语义的相关信息，**造成这个模型无法进行语义相关的判断**。**

### **2.2.2 潜在主题模型**

**潜在主题模型(****Latent Topic Model)基于具有相似含义的词汇总是出现在相似的语境中的假设，利用文本中具有相似语境的词汇间的联系，将文本空间映射到具有更低维度的向量空间中。对于语义关系相近的词来说，它们的向量表示在向量空间中距离较接近，可以说它们拥有相同的主题，且可以用余弦距离来衡量它们的相似度。**

潜在语义索引(Latent Semantic Indexing，简称LSI)**LSI是主题模型的一种，，**是T．K．Landauer、S．T．Dumais等人提出的一种将文档组织成语义空间结构的方法。是对传统的向量空间技术的一种改良。LSI使用one-hot方法表示文档，然后把所有文档放在同一矩阵中，由于使用向量表示之后矩阵规模巨大、而且过于稀疏，因此需要通过奇异值分解（Singular Value Decomposition，SVD）对矩阵降维，最后得到文本中主题和主题之间的关系。LSI方法虽然表示和计算过程都比较简单，但是也有如下几个缺点：（1）通过SVD生成的新矩阵解释性较差；（2）无法解决“一词多义”的问题；（3）和词袋模型一样，忽略了文章中单词的先后顺序。