1. **引言**
   1. **概述**

在近十几年来，计算机领域发展迅猛，个人计算机，终端互联设配也越来越普及，几乎是人人随时随地都能上网，在当今互联网这个巨大的生态圈里面，每年互联网用户产生的数据正以惊人的速度增涨。这些逐年增长的数据中绝大多数的数据都是非结构化的数据，这些数据主要存储的形式都是文本、网页、图片、视频、音频等非结构化的形式。结构化的数据它是经过了一定的存储格式，处理规则加以封装，所以它能够被计算机根据其存储的规则来进行识别，而非结构化的数据比如说文本数据，这些数据没有规则，格式，通常情况下它是不能被计算机直接识别的。当人们想对这些数据进行分析，统计，挖掘的时候，由于其不是结构化的数据，很难通过数据挖掘的技术对其进行处理。人工筛选的方法在这个大数据时代显然是天方夜谭，虽然说人工筛选的方法在上世纪80年代之前都是的确存在的，有需求就有市场这句话放在科研的领域也是行的通的，正因为有对非结构化文本识别和分析的需求，所以研究者们也想出了各种相关的方法。文本挖掘（Text Mining），图像识别，模式识别等一些相关技术就对分析和识别这一些非结构化的数据起到很好的作用。在此篇论文中我们主要讨论一下文本挖掘技术。

文本挖掘(Text Mining)是一个从非结构化文本信息中获取用户感兴趣或者有用的模式的过程。首先它会把非结构化的文本数据通过一定的方法转换成为结构化的数据，然后再从这些结构化的数据中分析提取出我们用户所感兴趣的模式和东西。文本挖掘技术是一项非常复杂的技术，它是一个多学科混杂的一门学科，其中包含有信息技术、机器学习、统计学、数据可视化、数据库技术、数据挖掘。

虽说集这么多门学科于一身，但是它实实在在的在我们这个大数据时代，信息化时代发挥了重要的作用，也是在信息学科领域中比较热门的学科之一。

在生物医学领域，借助着计算机科学技术也在快速的发展，每年发表在互联网上的生物医学文献资料和研究者们在实验室产生的实验数据也正以惊人的速度增长，产生的这些文献资料和实验数据大部分都是以文本的形式存储在计算机设配上面。所以也是非结构化的数据，不能被计算机直接识别。那么我在此篇论文里主要讨论针对文本形式存储的非结构化数据的处理以及信息的提取。