与文本挖掘技术关系十分密切的另外一个概念就是文本标准化。语言是非常灵活的一种信息形式,这导致其表现形式通常是杂乱无章的,这为文本内容分析带来诸多不便与困扰。因此,在进行文本分析之前,一般需要对其进行标准化处理,增加文本类型数据的一致性、整洁性。对文本进行标准化处理,可以降低后续文本建模、挖掘、分析等任务的难度,并提升结果的准确度。文本的标准化相当于数据挖掘技术的数据预处理阶段,一般在文本结构化任 务之前。对文本进行标准化处理的具体方法很多,其依赖于所分析的文本集合的特征。很多标准化方法是基于丰富的数据处理经验进行设计与构建的,当前常用的文本信息标准化方法有以下几种。

(1)对文本集合的语言进行统一，去除其他语种文章。

(2)去除无效符号、特殊符号，以及网页标记。

(3)简体字与繁体字的统一(中文文本)。

(4)去除文本显示格式信息(如空格、回行等)。

(5)对文本中的错别字自动修复。

(6)构建同义词或近义词表，减少文本特征。

(7)过滤低频(词汇出现次数很低)的文本特征。