Atitit 文档资料的摘要与压缩技术总结

目录

[1. 1. 摘要的作用 1](#_Toc19389)

[1.1. .书写摘要的基本规范和原则 1](#_Toc4900)

[2. 2](#_Toc23206)

[3. 2](#_Toc13483)

[4. 三大层次 纯文本内容摘要 目录 标题摘要 2](#_Toc14313)

[4.1. 格式转换为txt 摘要 2](#_Toc20861)

[4.2. 目录提取 2](#_Toc21837)

[4.3. 标题摘要 2](#_Toc26855)

[4.4. Keyword提取 2](#_Toc24298)

[5. 摘要就是整篇文章和浓缩预览，四个要素 2](#_Toc160)

[6. 主流的文本摘要方式抽取式（extractive），另一种是生成式（abstractive）。 3](#_Toc3499)

[6.1. Keyword 散列化摘要 4](#_Toc22449)

[7. 其他技术 4](#_Toc3872)

[7.1. 分词 4](#_Toc11651)

[7.2. Stopword体系 4](#_Toc13493)

# 1. 摘要的作用

摘要也就是内容提要，是论文中不可缺少的一部分。论文摘要是一篇具有独立性的短文，有其特别的地方。它是建立在对论文进行总结的基础之上，用简单、明确、易懂、精辟的语言对全文内容加以概括，留主干去枝叶，提取论文的主要信息。作者的观点、论文的主要内容、研究成果、独到的见解，这些都应该在摘要中体现出来。好的摘要便于索引与查找，易于收录到大型资料库中并为他人提供信息。因此摘要在资料交流方面承担着至关重要的作用。  
2

## .书写摘要的基本规范和原则

（1）论文摘要分为中文摘要和外文（一般为英文）摘要。摘要在篇幅方面的限定，不同的学校和机构有不同的要求，通常中文摘要不超过300字，英文摘要不超过250个实词，中英文摘要应一致。毕业论文摘要可适当增加篇幅。  
（2）摘要是完整的短文，具有独立性，可以单独使用。即使不看论文全文的内容，仍然可以理解论文的主要内容、作者的新观点和想法、课题所要实现的目的、采取的方法、研究的结果与结论。  
（3）叙述完整，突出逻辑性，短文结构要合理。  
（4）要求文字简明扼要，不容赘言，提取重要内容，不含前言、背景等细节部分，去掉旧结论、原始数据，不加评论和注释。采用直接表述的方法，删除不必要的文学修饰。摘要中不应包括作者将来的计划以及与此课题无关的内容，做到用最少的文字提供最大的信息量。  
（5）摘要中不使用特殊字符，也不使用图表和化学结构式，以及由特殊字符组成的数学表达式，不列举例证。

# 

# 

# 三大层次 纯文本内容摘要 目录 标题摘要

## 格式转换为txt 摘要

## 目录提取

## 标题摘要

## Keyword提取

# 摘要就是整篇文章和浓缩预览，四个要素

* 它被排放在论文的首要位置。
* 摘要主要要包括论文的四个要素，即：

1. **题目**
2. **目的**
3. **方法**
4. **结果**
5. **结论**
6. **关键字**

反映论文的实质性内容，展示论文内容足够的信息，体现论文的创新性，展现论文的重要梗概，一般由具体研究的对象、方法、结果、结论四要素组成。

对象——是论文研究、研制、调查等所涉及的具体的主题范围，体现论文的研究内容、要解决的主要问题，是问题的提出，研究方向的确立与目标的定位。

方法——是论文对研究对象进行研究的过程中所运用的原理、理论、条件、材料、工艺、结构、手段、程序，是完成研究对象的必要手段。

结果——是作者运用研究方法对研究对象进行实验、研究所得到的结果、效果、数据，被确定的关系等，是进行科研所得的成果。

结论——是作者对结果的分析、研究、比较、评价、应用、提出的问题等，是结果的总结，

[发布于 2016-12-20](/question/20028088/answer/136751326" \t "_blank)

# **主流的文本摘要方式**抽取式（extractive），另一种是生成式（abstractive）。

目前主流的文本摘要自动生成有两种方式，一种是抽取式（extractive），另一种是生成式（abstractive）。

抽取式顾名思义，就是按照一定权重，从原文中寻找跟中心思想最接近的一条或几条句子。而生成式是计算机通读原文，在理解整篇文章意思的基础上，重新生成概要。

抽取式摘要目前已经相对成熟，但抽取质量及内容流畅度均差强人意。伴随着深度学习的研究，生成式摘要对质量和流畅度都有很大的提升，但目前也涉及到原文本长度过长、抽取内容不佳等问题的限制。

**文本摘要的发展概况**

抽取式摘要是一种比较成熟的方案，其中Text rank排序算法以其简洁、高效的特点被工业界广泛运用。大体思想就是先去除文章中的一些停用词，之后对句子的相似度进行度量，计算每一句相对另一句的相似度得分，迭代传播，直到误差小于0.0001，再对上述方法得到的关键语句进行排序，即可获得摘要。抽取式摘要主要考虑单词词频，并没有过多的语义信息，像“猪八戒”、“孙悟空”这样的词汇都会被独立对待，无法建立文本段落中完整的语义信息。

入门\_ 文本摘要自动生成技术的前世今生 - 简书.html

## Keyword 散列化摘要

分词然后去重

# 其他技术

## 分词

## Stopword体系