|  |  |
| --- | --- |
| **民大校徽** | 软件工程综合实践 |
|  |

基于TF-IDF的中文文本关键词提取技术

|  |  |  |  |
| --- | --- | --- | --- |
| 姓 名： | 薛志豪 | | |
| 学 号： | 16046023 | 年 级： | 2016级 |
| 院 系： | 信息工程学院 | | |
| 专 业： | 软件工程 | | |
| 指导教师： | 邱丽榕 | | |

2020年3月9日

摘 要

人是信息的创造者，更是信息的传播者和接收者。现如今的人们处于信息爆炸的时代，飞速发展的网络技术、层出不穷的新媒体，每天制造着无法计数的信息。在信息越加丰富多元的同时，如何在海量信息中快速搜寻和准确辨识所需要的内容，成为人们面对的最大难题，也是当代群众浏览活动的最大困扰。

在信息检索面对的海量文本内容时，除学术领域文本和少数结构化的短文本外，绝大部分文本并不标注关键词，甚至没有标题。要在海量文本中进行信息检索，首先要面对的就是关键词的提取。而如何准确高效地实现文档关键词的提取，就是本文即将研究和解决的问题。

关键词提取技术是数据挖掘领域的一个分支，也是自然语言处理领域的一个基础性研究内容。关键词提取技术致力于帮助人们快速捕捉文本特征，了解文本的基本内容，提取文档摘要，以达到快速检索、准确识别感兴趣内容的目的。同时，关键词提取技术还是文本检索技术、网络搜索技术、文档摘要技术和文本聚类技术的基石。

本文研究的主要内容就是从一般性文本中提取出其关键词。在信息检索领域，如果通过人工的方法进行关键词的提取，可以很准确地提取出文本的关键词。但人工提取关键词的效率和成本，在当代互联网海量的信息内容面前显然不可接受。相反，计算机实现的关键词提取技术效率高，但准确性很难提高。本文将从统计类的一般算法入手，提出一种准确、高效的关键词提取方法。

关键词：信息检索；关键词提取；分词；TF-IDF

**Abstract**

People are the creators of information, but also the communicators and receivers of information. Nowadays, people are in the era of information explosion. The rapid development of network technology and the emergence of endless new media are creating countless information every day. As the information becomes richer and more diverse, how to quickly search for and accurately identify the required content in the mass of information has become the biggest problem that people face and the biggest problem for contemporary mass browsing activities.

In the massive text content faced by information retrieval, with the exception of academic field texts and a few structured short texts, most of the texts are not tagged with keywords or even titles. To search for information in massive texts, the first thing to do is to extract keywords. How to accurately and efficiently extract the keywords of the document is the problem to be studied and solved in this article.

Keyword extraction is a branch of data mining and a basic research content in the field of natural language processing. Keyword extraction technology is dedicated to helping people quickly capture text features, understand the basic content of text, and extract document abstracts to achieve the purpose of rapid retrieval and accurate identification of content of interest. At the same time, keyword extraction technology is the cornerstone of text retrieval technology, web search technology, document abstraction technology and text clustering technology.

The main content of this paper is to extract its keywords from general text. In the field of information retrieval, if keywords are extracted manually, the keywords of the text can be accurately extracted. However, the efficiency and cost of manually extracting keywords are obviously unacceptable in the face of the vast amount of information content on the contemporary Internet. In contrast, computer-implemented keyword extraction technology is efficient, but its accuracy is difficult to improve. This paper starts with general statistical algorithms and proposes an accurate and efficient keyword extraction method.

**Key Words:** Information retrieval; Chinese word segmentation; TF-IDF; Keyword extraction

目录

[摘 要 1](#_Toc35109075)

[**Abstract** 2](#_Toc35109076)

[一 绪论 6](#_Toc35109077)

[（一）研究背景及研究意义 6](#_Toc35109078)

[1. 研究背景 6](#_Toc35109079)

[2. 研究意义 6](#_Toc35109080)

[（二）研究现状 7](#_Toc35109081)

[二 相关技术和资源 9](#_Toc35109082)

[（一） 相关技术 9](#_Toc35109083)

[1. 中文分词技术 9](#_Toc35109084)

[（二）相关资源 12](#_Toc35109085)

[三 关键词提取技术 13](#_Toc35109086)

[（一） 关键词提取 13](#_Toc35109087)

[1. 关键词定义 13](#_Toc35109088)

[2. 关键词提取 13](#_Toc35109089)

[3. 关键词提取的过程 14](#_Toc35109090)

[（二） 关键词提取算法、模型 14](#_Toc35109091)

[1. 基于统计的TF-IDF算法[30] 14](#_Toc35109092)

[2. 基于图论的TextRank词图模型[32] 15](#_Toc35109093)

[四 基于TF-IDF的关键词提取设计 17](#_Toc35109094)

[（一） 文本预处理 17](#_Toc35109095)

[1. ICTCLAS分词 17](#_Toc35109096)

[2. 去停用词 17](#_Toc35109097)

[（二） 关键词提取 18](#_Toc35109098)

[1. 语料库IDF值的训练 18](#_Toc35109099)

[2. 权值计算与排序 18](#_Toc35109100)

[（三） 实验结果 18](#_Toc35109101)

[五 总结 20](#_Toc35109102)

[参考文献 21](#_Toc35109103)

# 一 绪论

## （一）研究背景及研究意义

### 1. 研究背景

随着时代信息化和数字化进程，网络联入千千万万的设备,连接起千千万万的人。越来越多的网络服务每天都制造着海量的信息。据估计，2020年网络数据将超过350亿TB[1]。数据量的急速膨胀显示了网络内信息、资源量的庞大。

互联网的快速发展连接起万事万物，人们足不出户就可知悉天下事，坐居家中就可静等外卖快递，遥居天涯也能实时面晤。要将网络资源化作人们实实在在的便利，面临着层出不穷的问题。丰富的信息为人们提供便利的同时，也增加了人们检索和利用信息的难度。如何有效地组织、检索信息便成为了网络技术发展过程中最重要的议题。

从万维网诞生之日起，便诞生了数不清的信息检索技术，谷歌更是作为先行者成长为如今首屈一指的巨无霸企业[2]。搜索引擎是依据用户所提供的关键词，在全网找到最有价值的网页反馈给用户。这时，关键词提取技术就对这种信息检索技术至关重要了。关键词提取的效率关系到检索反馈的速度，而关键词提取的准确度则直接关系到信息检索的质量

关键词往往关乎文本的中心思想，可作为文章主旨的代表，是形成文章摘要的关键。准确的关键词提取能提高信息检索的效率，可帮助用户快速把握文章内容，是连接网络资源和用户需求的纽带[3]。

在这个信息爆炸的时代，人们被形形色色的信息淹没，有价值的信息常会被数不清的垃圾信息掩盖，总让人们感到无所适从。曾经在信息检索领域，只能依靠人工方式整理堆积如山的尺牍文件。传统的人工整理手段具有较高的准确度，但效率和成本往往不可接受。信息时代庞大的信息量更使得人工方式不再可能实现，于是人们尝试让计算机来进行这一工作。利用计算机的关键词提取技术具有效率高但准确度低下的特点，于是寻找更加准确高效的关键词提取方法成为了信息检索领域的核心议题之一。

关键词提取技术应用广泛，是文本分类、信息检索、话题检测、自动摘要等众多自然语言处理和信息检索技术的基础性工程[4]。对于很多文本处理相关的操作，如搜索、分类、情感分析，都可以依托关键词提取技术实现[5]。因此，选中关键词题技术作为本文主题具有深刻的意义。

### 2. 研究意义

数字化和信息化的发展给人们提供了无限的可能，也带来了数之不尽的便利。人们安坐而知天下事，与天涯异处的好友犹如对坐，很多人力不可即的东西都可在家中轻松完成。但我们获得这些便利的同时，快速增长的网络资源也带来了巨大的麻烦。面对海量的数据资源，用户如何从中搜寻、识别所需要的资源呢？自计算机能够承载信息开始，便产生了基于计算机的数据处理技术。在传统文本传播中，人们习惯于标注出每一篇文章的主题、关键词。简略、准确的关键词能帮助用户把握文章内容，辨识所需信息。

我们所说的关键词，是指与文章主题密切关联的词语集合，常常作为文章摘要的依据。文本关键词可以帮助用户从大量文本中迅速把握文章的中心内容，协助用户准确检索得到自己所需要的文档。但在现实中，除少数学术类论文会包含关键词的标注外，绝大部分资源并不包含类似标识，甚至连标题都没有。曾经信息传播行业，通过人工方式手动提取关键词。这种方式往往具有极高的准确度，但效率低下，尤其面对海量文本时，其成本显得不可接受。因此，自人们利用计算机处理数据以来，便产生了依托计算机的数据处理和信息检索技术。计算机实现关键词提取技术，具有效率高但准确度低的特点。数十年来，人们一直致力于开发更高效的算法和跟准确的数据处理技术。

近二十年以来，人们积极钻研自然语言处理技术，提出了众多计算机自动提取关键词的算法。关键词提取技术是自然语言处理和数据挖掘技术的基础性工程之一，探索一个更加准确高效的关键词提取方法意义非凡[6]。

通过对现有关键词提取方法的研究，本文选取TF-IDF作为关键词提取算法的核心，并融入多种其他特征。考虑到词性和词长的区别，对于不同词性、词长设置不同权重来改进TF-IDF算法，提高的关键词提取的准确性。

主要工作如下：

1）深入研究TF-IDF算法，实现一个高效的关键词提取方法，并不断修正权值，达到更高的准确度。

2）需要一个足够大的语料库，用于训练IDF值。

3）研究词语相似度计算方法，对相似度高的词汇，进行词频合并。

4）采用中科院ICTCLAS分词软件进行分词和词频统计。

## （二）研究现状

上世纪五十年代，Luhn[7]首次提出了自助标引相关的技术。此后，人们在该技术领域持续研究五十余年。这项技术的研究一直延续到上世纪九十年代，出现关于继承自动标引技术的自动关键词提取技术[8]。

二十世纪末，随着全文索引技术广泛应用，同时传统自助标引方法发效率问题也遇到了天花板。自此，自助标引相关的技术便走上了下坡路，逐步为世人所抛弃[9]。

二十世纪九十年代末到现在，全文索引技术也越来越难适应新的时代。人们需要更加精确的标注结果，这一时期涌现的自动摘要、文本聚类、文档分类和主题检测等新兴互联网服务也对信息检索技术提出了更高的要求。而这些服务都极度依赖于一项准确而高效的关键词提取技术，解决好关键词提取相关的技术才能实现高质量的网络信息服务。自此，关键词提取相关技术成为了技术领域的风口，这一时期得到了快速发展。

国内外众多学者在关键词提取倾注了颇多心血，在该领域提出过许多高价值的研究工作，也取得了众多极其优秀的研究成果。

关键词提取技术的研究最早在国外发起，现今成熟的关键词提取解决方案大多由国外专家提出。其中，产生了众多极具实践价值的算法和模型[10]。

1999年，Witten[11]借助朴素贝叶斯的方法进行关键词特征模型的建立，开发了KEA系统。该系统将关键词的抽取视作有监督学习的过程，并基于此建立了全新的预测模型。此后，Turney还率先将机器学习中遗传算法和决策树的方法引入关键词提取技术。Turney[12]将关键词的提取作为一个文本分类问题解决——将文本中候选词分为关键词和非关键词两类，再依据大量已标注训练集得到训练模型。很显然该方法过于粗糙，也十分依赖训练集的规模和质量，且对于不同类型文本的关键词提取效果差异很大。

基于有监督学习方法打关键词提取技术，在实现以及实现后的通用性、准确度方面并不能提供一个很好的解决方案，此类技术往往并不可靠。于是，出现了如今主流的众多基于无监督训练方案的关键词提取技术。Matsuo和Ishizuka[13]共同提出了一种可应用于单个文本且不需要一个高质量标注数据集的关键词提取技术。他们提出一种数据压缩算法获取一项特征值n-gram，过滤非关键词元素，依据权重排序获取文本关键词，就是N-gram filtrationt算法。Hulth[14提出了从学术论文中提取关键词的算法Rule Induction——在词频统计的基础上加入了句法分析相关的方法。同时，还对词语的词性进行区分，得到一种准确度较高的自动提取算法。后来，Liu和Tonella[15]还依据TF-IDF算法提出了一种关键词自动提取技术，原理简单，且效果不俗。

国内关键词提取技术相对于国外起步较晚，大约开始于上世纪八十年代末。国内很多学者专注于中文文本关键词提取相关技术的研究，也提出了很多具有代表意义的解决方案。如果不考虑中文分词等差异，大部分关键词提取算法都适用于中文。都云程[16]根据字同现频率提出的关键词提取算法，改进TF-IDF加入词位置和文本长度作为特征。通过不同特征加权计算，从候选词中选取权重大的词作为关键词。索红光[17]基于《知网》建构词汇链实现关键词提取。李素建等提出基于最大熵模型的关键词提取算法。方俊[18]将候选词的语义特征加入考虑提出了新的关键词提取算法。黄磊等通过添入一个新权值提高关键词特征权重，提高了关键词提取的准确度。

# 二 相关技术和资源

本章将介绍本文研究内容内，涉及到的几项基础工具，和研究过程中使用到的数据、资源，包括分词工具ICTCLAS、未登录词识别、搜狗语料库、哈工大停用词表等。

## 相关技术

### 1. 中文分词技术

中文是典型的象形文字，与英文等字母文字不同，中文词与词之间没有空格作为间隔。中文以字为基本单位，而且单个字往往无法表示具体含义，词往往由不等的字构成。中文里词与词的分割是基于语句的，没有明确间隔。对于关键词提取，最基本且最重要的就是准确地进行分词。所以，我们在文本预处理阶段便需要完成中文分词操作。

在进行中文自然语言处理时，通常需要先进行分词，分词效果将直接影响词性、句法树等模块的效果。中文分词作为中文自然语言处理的基础性工程，直接影响自然语言处理对语义的把握准确度。目前，中文分词的方法有很多，虽然说法不同，但可以根据原理分为词典切分法、统计分析法、语义语法 规则分词法和人工智能法四类。

1. 词典切分法[19]

将中文字符串与词典中的词条按照一定的规则进行匹配，若在词典中匹配到某一词条，则查找成功。这种根据字符匹配的中文分词方法，就是机械分词方法。基于词典切分的中文分词方法，理论上只要有完备的词典、不限匹配次数，可以实现相当高的分词准确度。但是，随着词典容量的扩大和匹配文本的增长，其效率会急剧降低。

1. 统计分析法[20]

通过统计字符串在语料库中出现的频率，依据频率判断该字符串是否构成一个词。相似的字符串出现频率越高，表示其构成一个词的可能性越大。同时，字与字相邻出现的频率，也可以很好地代表其构成词的可信度。基于统计分析法最典型的模型方法就是N-gram模型[21]。

1. 语义语法规则分词法

将语义分析引入，作为分词系统的补充，如扩充转移网 络法、知识分词语义分析法、邻接约束法等[22]。

1. 人工智能法[23]

基于人工智能方法的分词技术，又称为理解分词法。根据模拟人对语义的理解，实现分词的效果。如今，有很多学者将神经网络技术引入分词领域，但成熟的模型方法还较少。

当然，还有学者提出基于字的中文分词方法。这类方法不再预先分词，无需费尽心思解决中文里字词组合的复杂性，但同样的，成熟的模型和方法还没有出现。因此，如今的中文关键词提取技术依然依赖于准确高效的分词系统，分词结果的质量对关键词提取环节至关重要。

2. 中文分词系统ICTCLAS[24]

中文分词对关键词提取技术至关重要，本文采用中科院计算技术研究所开发的NLPIR-ICTCLAS系统进行中文分词。ICTCLAS主要支持中文分词、词性标注、命名实体识别、新词发现，还支持导入用户词典。目前，ICTCLAS系统最新2016版本已至此中英文混合分词和基于交叉信息熵算法的关键词提取功能（包括新词与已知词）。如今成熟的分词工具有很多，如中科院计算所NLPIR、ansj分词器、哈工大LTP、清华大学THULAC、Hanlp分词器等等，NLPIR-ICTCLAS主要有四点突出优势：

1. 性能最优

ICTCLAS采取PDAT知识库管理技术，在高效率和高准确性方面均获得了提升，能够管理百万容量的词典库，单机最高检索速度能达到100万条/秒，并且内存占用不足字典库容量的3/2。由于该技术的重大突破，ICTCLAS单机分词速度达到996KB/s，分词精度超过98.45%，各种词典压缩容量不足3M，成为了如今世界上性能最好的中文分词工具[25]。

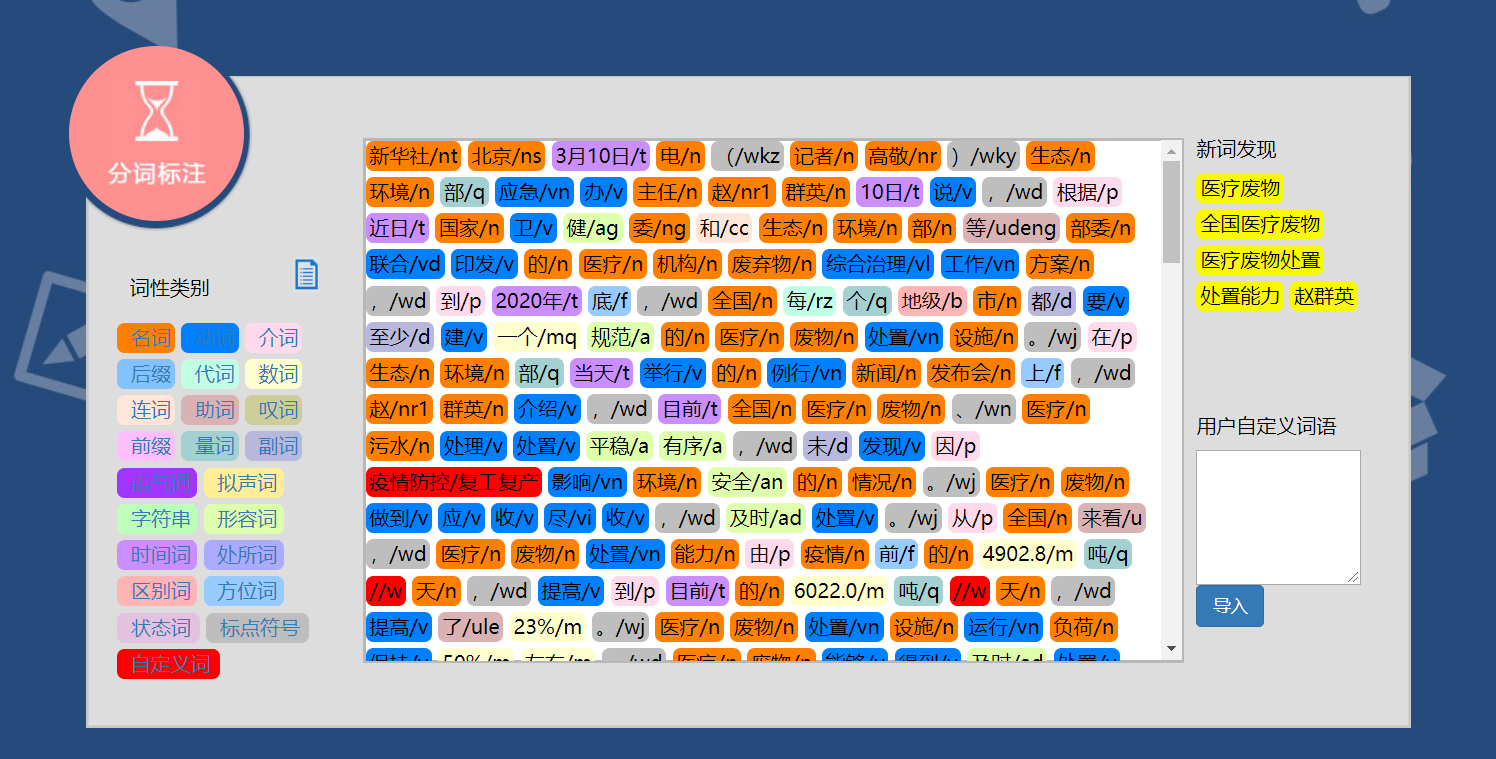
1. 统一的计算理论框架

ICTCLAS根据层叠隐马尔可夫模型，把中文词法分析各种内容统一到一个完整的理论框架里，并最终获得了最好的实践效果。该成果在理论上和实践上都得到了印证，证明了该模型的先进性。

1. 良好的可移植性

ICTCLAS系统使用C/C++完成编写，目前支持包含Linux、FreeBSD、Windows的几乎所有主流系统，并且提供了支持C/C++、C#、Java、Python语言调用方式，适用范围广，使用灵活。

1. 良好的扩展性

ICTCLAS所有功能都是模块化设计的，可以自由组装、自行调用。例如，ICTCLAS支持几乎所有编码方式，支持繁体中文、支持自定义词典。用户可以根据需求，定制最适合自身业务的分词系统。

3. 过滤停用词

停用词是指在信息检索过程中，为了节省存储空间和提高处理速度，在进行预处理过程中会自动过滤的字或词，这种词被称为stop\_words（停用词）。在英文文本中，停用词往往是一些冠词、量词之类，如the、a。而在中文文本中，停用词常常表现为虚词、符号和一些无意义字符片段。在进行自然语言处理之前，剔除这些无意义的停用词，对于后续提取关键词的效率和准确度都大有裨益。停用词主要包括虚词这种无实义的词、助词等只为句法结构服务的词。停用词往往与文本主题无关，不但不能代表文本内容，其高频次还会影响关键词的提取，增加语言处理的复杂度。去停用词还能减少词条数、缩小候选词规模，对于提高精度和效率至关重要。

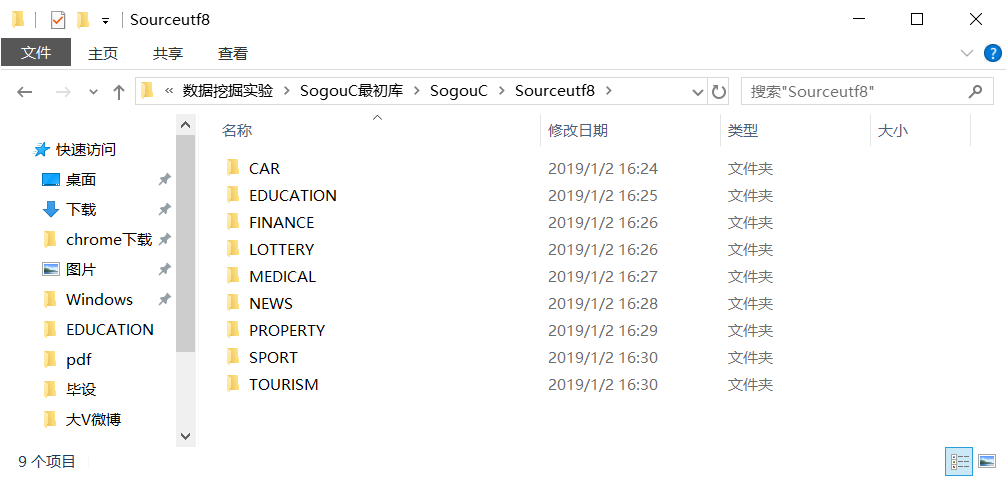
3. 未登录词识别

网络的快速发展，使得网络媒体将天涯海角的人联系在一起。任何一个网络社区建立，都会吸引千千万万志同道合的人聚集旗下，一个社会热点的爆出会引发全国的大规模讨论。在2014年的一次COAE2014评测中，使用分词系统对500条微博进行标准分词，发现给出的1129个新词集中，有881个新词被分成其它词。如“奥利给”会被分为“奥/利/给”，“菊花残”会被分为“菊花/残”。这种不在分词系统词典中，会被错误分割的词，称为未登录词。

这一类连续单字组成的未登录词，在分词过程中往往被错误切分，导致关键词提取的质量降低。所以分词前，需要先发现文本中的未登录词，将未登录提添加进用户词典，再进行分词，才能得到较为准确的结果。

## （二）相关资源

1. 语料库

本文实验中采用的语料数据来自于搜狗语料库，包含9个门类、每个门类均有约8000篇左右文章，共计约6万文档。该语料库主要来自于搜索引擎新闻类文本，文本长度适中，主题鲜明，适于用来做关键词提取的训练。测试时只需从各门类各自抽取几篇样本即可。

1. 停用词表

在网页内容中适当地减少Stop Words出现的频率，可以有效地帮助我们提高关键词密度，而在网页Title中避免出现Stop Words往往能够让我们优化的关键词更突出。去停用词关系到后续关键词提取的效率和准确度，所以一个完备的停用词表对于后续研究至关重要。本文中我选用哈工大停用词表，包含中英文停用词攻击2700余条。

# 三 关键词提取技术

本章将详细介绍关键词提取技术相关的内容，介绍关键词提取技术的分类和方法，以及现今较为成熟的几种模型。

## 关键词提取

### 关键词定义

关键词是指与文章主题直接关联，能够概括文章核心内容的重要词汇，英文称为keyword。我们看关键词字面意思，很容易将关键词理解为单个的词，事实上能够准确概括文本的往往是多个词构成的关键短语。研究表明，大部分关键词实际上都是一个短语结构，即多个词的组合。这主要是因为短语结构包含明确语义，能够准确表达文章主题，最适合成为文本概括的关键词[26]。关键单词的提取难度较小，但是关键短语的提取相对比较复杂。要准确提取关键短语，就需要对短语结构进行明确界定。

关键词按其所表示的文本范围不同，可以分为“局部关键词”和“全文关键词”[27]。局部关键词指能够概括全文某一部分、段落、章节的关键短语，它能够准确表达一定范围内的内容，准确性较高。全文关键字是对整个文本内容的凝练和概括，涉及范围广，与主题联系紧密。

### 关键词提取

关键词提取也被称为关键词标引[28]，即一种将文本中具有代表性、具有辨识性的词汇提取出来的技术，是从文本中提取出能概括文档主题、并与其他文本区分的词语[29]。

关键词提取有两种基本方式：传统的人工方法和如今的计算机关键词提取技术。在计算机技术大规模运用之前，人们要整理大量文档的关键词，需要雇佣人工进行概括、整理。人工方式提取关键词具有准确性高、概括度强的特点，但是主观性强烈，而且效率极其低下。在面对如今网络中难以计数的文本内容时，其成本和效率均难以接受。计算机实现的关键词提取技术，是指通过计算机技术自动地实现关键词提取操作。计算机关键词提取技术具有效率高但是准确度差的特点，所以关键词提取技术所依赖的算法优劣直接关系到关键词提取的效率和质量。但计算机关键词提取技术天然适应大规模文本处理，是如今进行关键词提取的必然选择。

关键词提取是如今自然语言处理的基础领域和核心技术。在信息检索、情感分析、数据挖掘和文本分类集类领域，高质量的关键词提取可以提高整个系统的效率和成效，尤其对于大规模文本处理具有深刻意义。

### 关键词提取的过程

关键词提取过程中，将一个文本视作一个词语集合，该词语集合众多词具有聘书、位置、词长、词性等特征属性。而关键词提取技术，就是依据这些特征判断哪些词可作为关键词，并从中提取出来。所以，关键词提取技术主要有以下两个部分：

1. 对文档进行与处理，获取关键词的候选词集合

这是一个基础性的问题，往往发生在文本预处理阶段，需要进行一系列的词语抽取和过滤 得到候选词集合。首先对文本进行分词操作，将文本转化为一个词汇集合。该集合中会包含很多符号以及虚词、助词等与关键词提取无关的字词，所以还需要提出这一部分字符。目前主流的方法是依据停用词表，过滤这一类无关字符。

1. 依据一定算法分析各项特征，从候选词中选出最适合作为关键词的集合

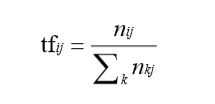
这一部分就是关键词提取技术的核心部分，依据一定算法考虑各项特征值产生关键词权重，并选定关键词。关键词提取算法中可考虑的特征有词频、位置、词性、词长和反文本频率等，依据一定统计学方法，如互信息量、最大熵、交叉信息熵等进行加权操作。最终得到文本中候选词的权重，以及权重排序，选中权值大的候选词作为文本关键词。

## 关键词提取算法、模型

### 1. 基于统计的TF-IDF算法[30]

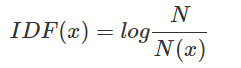
TF-IDF算法的核心原理是，词的重要性与它在文件中出现的次数成正比，但同时会随着它在语料库中出现的频率成反比。即，一个关键词在语料库中的文档频率会很低，而在该文本中词频却会较高。

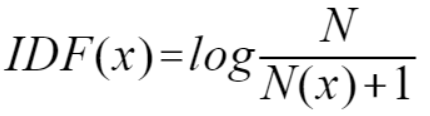
1. 词频TF（term frequency）

词频指的是某一词语在文件中的频率。能够作为关键词的词语，往往在文本中出现得很频繁，即在单个文本中词与的重要程度与其出现次数呈现一定正比关系。而为了削弱长文本与短文本的差异，我们选择词频作为特征值引入算法。词频TF的计算方法如下：

在该公式中，ni代表词语ti在文档dj中出现的频数，Σknij表示文档dj中所有词语的频数之和。

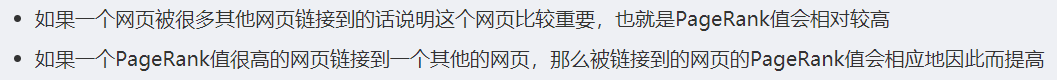
1. 逆文档频率IDF（Inverse Document Frequency）

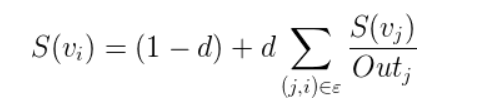
逆文档本质上是一种一直噪音的加权方式，该方法认为文档频率越小的词汇对于单一文本越重要，而文本频率越大则越不能作为代表性词语[31]。在IDF的计算方法中，只在少数文档中出现的词比在很多文档中出现的词语更具有代表性，也因此会获得更大的权重。IDF的计算公式如下：

其中，N代表语料库中文本总数，N(x)代表语料库中包含词x的文档数。因为N/N(x)在语料库训练中往往回事一个远远小于1的小数，很容易下溢出，所以我们通过取对数的方式使得IDF的值更加平滑。但还有一个问题，就算语料库很庞大，还是可能有一些词语在语料库中出现频数为0，这就可能导致计算公式中分母为0。于是IDF有另一个修正后的计算公式，一般也都用这个计算公式：

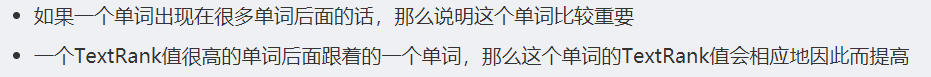
TF-IDF算法的最终值TFIDF(x) = TFi,j \* IDFi ；即词i在文本j中出现的词频乘于词i在语料库中的逆文档频率IDFi得到TD-IDF算法的最终权值。TF-IDF算法原理简单，但其有效性在自然语言处理领域已经得到了印证。TF-IDF作为一种统计类的权重计算方法，在关键词提取领域效用非常。

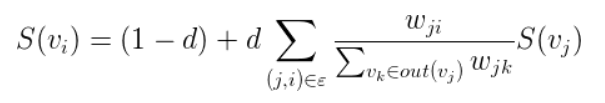
### 2. 基于图论的TextRank词图模型[32]

在介绍TextRank前，我想先给大家介绍下PageRank，实质上个人认为可以把TextRank当做PageRank2.0。谷歌的两位创始人的佩奇和布林，借鉴了学术界评判学术论文重要性的通用方法，“那就是看论文的引用次数”。由此想到网页的重要性也可以根据这种方法来评价。于是PageRank的核心思想就诞生了：

PageRank算法设计时将整个网络空间看作一个有向图，每一个网页是图上一个节点，网页之间的连接、引用是图上的有向边。算法运行时，为每一结点赋予一个初值，每一个引用会将引用网页的权值的一部分按比例分给被引用网页，该计算公式如下：

该公式中表示页面i被页面j引用；式经过调整使其运算结果能够最终收敛。

而TextRank算法思想就来源于PageRank算法：

这样TextRank的公式就可以由PageRank公式改写为：

该公式表示单词i的权重取决于与在i前面的各个点j组成的(j,i)这条边的权重，以及j这个点到其他其他边的权重之和。该模型也被成为词图模型。

1. 基于贝叶斯的LDA主题模型（Latent Dirichlet Allocation）

LDA主题模型是一种基于贝叶斯的文档主题生成模型。LDA是一个三层的贝叶斯概率模型，包括词、主题、文档三层结构。我们所说的生成模型，指的是我们认定一个文本的每一个词可以通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个逻辑获得。从文档到主题服从多项式分布，而从主题到词服从多项式分布。

LDA是一种非监督式的机器学习技术，可以用来识别大规模文档集或语料库中潜藏的主题信息。它采用了词袋模型的方法，这种词袋模型认为每一篇文档是一个词频向量，从而把文本信息转化为了易于建模的数字信息。然而，词袋方法并没有顾及词与词之间的顺序问题，这一方面降低了解决问题的难度，另一方面也为模型的改进提供了契机。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。

很容易得到这样一个概率公式：

而LDA主题模型要干的事就是：根据给定的一篇文档，反推其主题分布。通俗来说，可以假定认为人类是根据上述文档生成过程写成了各种各样的文章，现在我们想让计算机利用LDA干一件事：你计算机给我推测分析网络上各篇文章分别都写了些啥主题，且各篇文章中各个主题出现的概率大小（主题分布，后面还会提到）是啥。最后依据概率值排序，选择最高的几个作为主题关键词。

基于主题关键词提取算法主要利用的是主题模型中关于主题的分布的性质进行关键词提取。在某些场景，关键词不一定会显示的出现在文档中，这种情况下，前面两种算法显然不能提取出隐含的主题信息，这时就需要用到主题模型。

# 四 基于TF-IDF的关键词提取设计

本章节将具体介绍一个基于TF-IDF算法的关键词提取系统的设计。该系统的实现主要分为两个部分，文本预处理和关键词提取部分。其中文本预处理包括ICTCLAS分词和去停用词；关键词提取部分包括IDF的语料库训练以及TFIDF权重计算。

总体架构图：

ICTCLAS

语料库

IDF算法

训练模块

IDF词库

TFIDF

排序

待测文本

文本词频

关键词

预处理

## 文本预处理

### ICTCLAS分词

不论是待训练语料库，还是测试文本，都必须先进行文本分词。每一个文本都被分词处理为词条组成的集合，这些词条才是后续处理的基本单元。本文中采用中国科学院计算技术研究所开发的ICTCLAS系统实现中英文混合分词，ICTCLAS不光支持分词和词性标注，还支持词频统计，为后续处理提供很大方便。

### 去停用词

本文中根据哈工大停用词表（共计2700余条）进行停用词过滤，提出与文本关键词提取无关的词条，压缩文本处理规模。经过分词和停用词过滤后产生的词条集合就是候选词集合了，可以作为关键词提取阶段的输入信息。

导入文本

ICTCLAS分词系统

停用词表

去停用词

词频统计

文档

## 关键词提取

### 1. 语料库IDF值的训练

基于TF-IDF的关键词提取，需要依赖一项特别的特征权值——逆文档频率IDF。这一项特征值可以通过语料库的预先训练获得。语料库分为9大类，文本规模如下（单位：篇）

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 汽车 | 教育 | 经济 | 博彩 | 医疗 | 快讯 | 不动产 | 体育 | 旅行 |
| 8000 | 6041 | 8000 | 7588 | 8000 | 8000 | 5261 | 8000 | 3564 |

训练流程如下：

输入：经过分词、去停用词和词频统计后的语料库；

Step1：统计语料库中所有候选词的集合以及语料库总规模N；

Step2：统计词语i在语料库中的文档频数N(x);

Step3：计算公式得到词语i的逆文档频率IDF(x);

输出：所有候选词以及其IDF组成的集合。

分词模块

词频集合

语料库

文档频数统计

IDF词库

### 2. 权值计算与排序

语料库训练完成后就可以测试关键词提取了候选词IDF后就可以进行测试了。

测试流程如下：

输入：待测试文本，IDF训练结果；

Step1：将待测文本进行分词、去停用词，并进行词频统计（TF）；

Step2：词频TF\*IDF值，获得每一个候选词的TFIDF值；

Step3：将该文本候选词按TFIDF值排序，可取前五作为关键词；

输出：该文本关键词集合。

## 实验结果

关键词提取系统的质量并不好把握。下面通过一个网络新闻文本进行测试，测试结果如下表所示：

|  |  |
| --- | --- |
| 测试文本 | 人民日报：中西部35万个扶贫项目开工超三成/国家将给52个未摘帽县“吃小灶”  http://paper.people.com.cn/rmrbhwb/html/2020-03/11/content\_1975655.htm |
| 关键词 | 扶贫项目、优先、疫情、农产品 |

TF-IDF算法原理很简单，原始的TF-IDF算法运用于关键词提取事实上表现很差。如果语料库中某一类别文本特别多，那么对于该类别文本，其关键词的IDF（逆文档频率）就会很低，导致对该类别文本的关键词提取效果很不理想。

# 五 总结

TF-IDF算法原理简单，数据挖掘的很多领域都有广泛应用。但是，原始的TF-IDF算法运用于关键词提取是非常粗糙的，主要有以下缺点：

1）受语料库影响大。若某一类别语料比例特别多，则会导致该类别关键词极易被屏蔽，这种关键词提取算法对于语料库要求过高。

2）仅考虑了文本中词频这一种特征。文本中还有词性、词的位置、词义、词长等特征，都对于词语的重要性有很大影响，这些原始TF-IDF都不考虑。

本文中所实现的实际上就是一种原始的TF-IDF关键词提取算法，实验结果很一般，很多文本会提取出过多无关紧要的词语。之后毕设的设计过程中，我希望能尝试将更多的词特征加入考虑，或者尝试以下TextRank算法和LDA模型。

# 参考文献

1. 中国互联网络信息中心(CNNIC).第 25 次中国互联网络发展状况统计报告[EB/OL]. http://www.cnnic.net.cn/uploadfiles/doc/2007/7/18/113843.doc.
2. 李晨.网络搜索引擎与专家检索系统框架和模型研究[D].北京邮电大学,2009.
3. 王毅.基于 web 的信息抽取方法研究[J].科技与生活,2010(13):11.
4. 姜舟.关键短语抽取及相关技术研究[D].哈尔滨工业大学,2010.
5. 王莉,许凯.浅谈文本数据自动标引系统的设计[J].图书馆理论与实践,2013(6):95-97.
6. 杨林. 基于文本的关键词提取方法研究与实现[D].安徽工业大学,2013.
7. H. P. Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literry Information[J]. IBM Journal of Research and Development , 1957, 1(4): 309-317.
8. 张静,自动标引技术的回顾与展望[J].现代情报,2009(4):221-225
9. 罗准辰.关键词抽取的研究与实现[D].国防科学技术大学,2008.
10. 杨凯艳. 基于改进的TFIDF关键词自动提取算法研究[D].湘潭大学,2015.
11. Niraj Kumar, Kannan Srinathan. Automatic keyphrase extraction from scientific documents usingn-gramfiltrationtechnique[C].ProceedingsofDocEng'08Conference,2008:199-208.
12. Turney P. D. Learning Algorithms for Keyphrase Extration[J]. Information Retrieval. 2000, 2(2): 303-336.
13. Matsuo Y., Ishizuka M..Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information[J]. International Journal on Artificial Intelligence Tool, 2003, 13:2004-2008.
14. Hulth. An improved automatic keyword extraction given more linguistic knowledge[A]. In: the 2003 Conference on Empirical Methods in Natural Language Processing[C], 2003:216-233.
15. Liu Y. , Ciliax B. J. ,Borges K., etc. Comparison of Two Schemes for Automatic Keyword Extraction from Medline for Functional Gene Clusteringe[A]. In: the 2004 IEEE Computational Systems Bioinformatics Conference(CSB 2004)[C], 2004.
16. 都云程,周伟,韩艳铧,吕学强.基于字同现频率的关键词自动抽取[J].北京信息科技大学学 报（自然科学版）,2011(6):35-38.
17. 索红光,刘玉树,曹淑英.一种基于词汇链的关键词抽取方法[J].中文信息学报 2006,20(6):25-30.
18. 方俊,郭雷,王晓东.基于语义的关键词提取算法[J].计算机科学,2008,35(6):148-151.
19. 周恕义,徐晓亮.基于字典的中文切词系统在北京市精品课程平台中的设计与实现[J].中国 教育技术装备,2010(9):3-5.
20. 崔虹燕. 一种改进的基于词频统计的中文分词算法研究[J].信息技术,2008(4):124-125.
21. 赵琦.基于 n-gram 模型的中文分词技术研究[D].南开大学,2007.
22. 王广正.基于知网语义相关度计算的汉语自动分词方法的研究[D].云南师范大学,2006.
23. 李华,陈硕,练睿婷.神经网络和匹配融合的中文分词研究[J].心智与计算,2010(2).
24. NLPIR/ICTCLAS 汉语分词系统.http://ictclas.nlpir.org/.
25. 百度百科ICTCLAS：https://baike.baidu.com/item/ICTCLAS/8609504?fr=aladdin
26. 姜舟.关键短语抽取及相关技术研究[D].哈尔滨工业大学,2010.
27. 谢晋.基于词跨度的中文文本关键词提取及在文本分类中的应用[D].浙江工业大学,2011.
28. 百度百科 主题标引：https://baike.baidu.com/item/%E4%B8%BB%E9%A2%98%E6%A0%87%E5%BC%95/11003923?fr=aladdin
29. 章成志.自动标引研究的回顾与展望[J].现代图书情报技术,2007(11):33-39.
30. 徐文海,温有奎.一种基于 TFIDF 方法的中文关键词抽取算法[J].情报理论与实 践,2008(2):298-302.
31. 施聪莺, 徐朝军, 杨晓江.TFIDF 算法研究综述[J].计算机应用,2009,29:167-170.
32. 李娜娜. 基于TextRank的文本自动摘要研究[D].山东师范大学,2019.