|  |  |
| --- | --- |
| **民大校徽** | 中央民族大学学士学位论文 |
| **Bachelor Thesis of Minzu University of China** |

基于TextRank的文本摘要和关键词提取

|  |  |  |  |
| --- | --- | --- | --- |
| 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4月12日

摘 要

当代人们每天制造着大量数据，每天传播信息，每天也接收无数信息。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。在线上线下每天处理的海量文档数据中，只有极少数的学术论文，绝大部分的文章不包含任何的关键词信息，有的甚至连标题都没有。我们需要一种高效的方法实现中文文本关键词提取和文本摘要，这就是本文将讨论的问题。

本文依据TextRank算法设计了一个中文文本关键词提取和自动文本摘要的系统。主要包含以下方面的内容：

（一）依据TextRank算法的关键词提取技术。

在该算法中默认窗口为5，以窗口内词频（Term Frequency）为初始值，选取图运算后TR值最大的几位作为关键词。

（二）融合BM25的TextRank文本摘要技术。

该技术选取BM25算法作为句间相似度的计算方法，并以IDF值为词特征，最终依据TextRank图运算获取句子的TR值。选中TR值最的几句，以原文顺序拼接得到文本摘要。

本文还涉及引入ICTCLAS中文分词、词性过滤、去停用词、相似度计算方法等内容，并最终在关键词提取和文本摘要上取得了良好的效果。

关键词： 关键词提取；文本摘要；TextRank；BM25

**Abstract**

Contemporary people make a lot of data every day, spread information every day, and receive countless information every day. People are now in an era of information explosion. Rapidly developing network technologies and endless new media are creating countless information every day. While information is becoming richer and more diverse, how to quickly search and accurately identify the required content in massive information has become the biggest problem faced by people and the biggest problem for contemporary mass browsing activities. Among the massive document data processed daily online and offline, there are only a few academic papers, most of which do not contain any keyword information, and some do not even have titles. Therefore, we need an efficient method to achieve Chinese text keyword extraction and text summary, which is the issue that this article will discuss.

This paper designs a system for Chinese text keyword extraction and automatic text summarization based on the TextRank algorithm. It mainly includes the following aspects:

(1) Keyword extraction technology based on TextRank algorithm.

In the algorithm, the default window is 5, the term frequency in the window is used as the initial value, and the digits with the largest TR value after graph operation are selected as keywords.

(2) TextRank text summarization technology integrated with BM25.

This technology selects the BM25 algorithm as the calculation method for the similarity between sentences, and uses the IDF value as the word feature, and finally obtains the TR value of the sentence according to the TextRank graph operation. Select the few sentences with the highest TR value and splice them in the original text order to get a text summary.

This article also involves the introduction of ICTCLAS Chinese word segmentation, part-of-speech filtering, removal of stop words, similarity calculation methods, etc., and finally achieved good results in keyword extraction and text summary.

**Key Words:** Keyword Extraction; Text Summarization; TextRank；BM25

目录

[**一 绪论** 6](#_Toc38040568)

[（一）研究背景 6](#_Toc38040569)

[（二）研究意义 7](#_Toc38040570)

[（三）国内外研究现状 7](#_Toc38040571)

[1、关键词提取 7](#_Toc38040572)

[2、文本摘要 8](#_Toc38040573)

[**二 相关研究与技术** 10](#_Toc38040574)

[（一）中文分词技术 10](#_Toc38040575)

[（二）去停用词 11](#_Toc38040576)

[（三）关键词提取概要 12](#_Toc38040577)

[1. 关键词定义 12](#_Toc38040578)

[2. 关键词提取 12](#_Toc38040579)

[3. 关键词提取的过程 13](#_Toc38040580)

[（五）相关算法与模型 13](#_Toc38040581)

[1. 基于统计的TF-IDF算法 13](#_Toc38040582)

[2. 基于图论的TextRank图算法 14](#_Toc38040583)

[3. BM25算法 15](#_Toc38040584)

[**三 基于TextRank关键词提取技术的设计** 16](#_Toc38040585)

[（一）总体架构图 16](#_Toc38040586)

[（二）预处理文本 16](#_Toc38040587)

[1. ICTCLAS分词 16](#_Toc38040588)

[2. 去停用词 17](#_Toc38040589)

[3. 预处理部分流程图 18](#_Toc38040590)

[（三）TextRank图算法 18](#_Toc38040591)

[1. 创建词汇窗口 18](#_Toc38040592)

[2. 创建图模型 19](#_Toc38040593)

[3. 进行图运算 20](#_Toc38040594)

[**三 融合BM25的TextRank文本摘要方法与设计** 20](#_Toc38040595)

[（一）总体架构图 20](#_Toc38040596)

[（二）文本预处理 21](#_Toc38040597)

[1. 分句 21](#_Toc38040598)

[2. 分词与去停用词 21](#_Toc38040599)

[（三）相似度计算 22](#_Toc38040600)

[1. 计算IDF词特征 22](#_Toc38040601)

[2. 计算句间BM25值，构建相似度矩阵 22](#_Toc38040602)

[（四）图迭代计算 23](#_Toc38040603)

[（五）提取摘要句与重排序 23](#_Toc38040604)

[**三 实验结果与总结** 24](#_Toc38040605)

[（一）实验结果 24](#_Toc38040606)

[（二）总结与展望 25](#_Toc38040607)

[**参考文献** 26](#_Toc38040608)

[**附 录** 27](#_Toc38040609)

[**致 谢** 28](#_Toc38040610)

**一 绪论**

本部分是论文第一章内容，主要讲述论文研究的背景、研究意义以及国内外研究现状。

（一）研究背景

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。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。每天都有企业推出越来越多的网络服务，同时每天都制造着巨大的信息流。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。

在这个信息爆炸的时代，人们被形形色色的信息淹没，真正需要的信息常会被数不清的无关内容掩盖，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。

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。丰富的信息为我们提供巨大的便利，也增加了人们检索数据和利用信息的复杂度。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。本文依据TextRank算法提出的关键词提取和文本摘要技术，就是基于该背景。

（二）研究意义

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？从计算机技术大发展之后，人们便开始使用计算机进行信息的存储和处理。在传统信息传播中，人们往往标注出文章的关键词或提取一个文本摘要来概括文本信息。关键词和文本摘要帮助用户快速把握文章内容，捕捉关键信息，是如今信息检索的关键技术。

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。

文本摘要技术比关键词提取技术更加复杂。文本摘要是对一个文本进行概括、提炼，得到的一段简介的描述性文字。本文中即将讨论的自动摘要技术，是通过计算机技术从文本中提炼出高度概括的语句，拼接形成有代表意义的摘要文字的技术。随着信息领域的扩张，文本摘要技术也越来越多地表现出不可替代的作用。

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]。目前主流的文本摘要技术大多从原文中选取代表性语句，再按原文顺序拼接为摘要文本。通过研究各种成熟模型，，本文依据TextRank算法设计了一个关键词提取技术，以及融合BM25的TextRank文本摘要技术。。

（三）国内外研究现状

1、关键词提取

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。

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[12]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。近年来，受PageRank算法启发，陆续有学者将图运算的方法引入关键词提取领域，产生了著名的TextRank算法。

国内关键词提取技术大约开始于上世纪八十年代末，同时期国际上相关研究成果已经比较多了。国内很多学者另辟蹊径，专注于中文文本关键词提取相关技术的研究，获得了很多有价值的研究成果。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。2008年，杨[16]将TextRank算法运用到中文关键词提取领域。基于TextRank算法的关键词提取算法成为，中文关键词提取的优秀解决方案之一。

2、文本摘要

与关键词提取技术相类似的，文本摘要技术同样随着信息的膨胀变得火热。文本摘要技术对于用户把握文章基本信息，筛选重要内容至关重要。现代搜索技术是为用户检索、筛选、压缩信息，快速提取关键内容。精准高效的文本摘要技术，有利于帮助节省用户浏览成本，提升用户体验。

文本摘要技术从上世纪五十年代开始获得学者关注，几十年来出现了许多自动文本摘要技术。传统的自动文本摘要方法大多依据统计学方法，依据语料库构建。上世纪90年代，互联网技术的极速膨胀，带来了迫切的文本摘要技术的需求。新世纪机器学习、人工智能技术的蓬勃发展，也为文本摘要技术带来了许多新思路和方法。

1995年，Kupiee基于朴素贝叶斯模型设计了一种文本摘要方法，该模型将句子视为相互独立的事件，通过朴素贝叶斯方法获取权重。1999年，Aone在此基础上提出了新的解决方案——Dimsim系统，考虑到更多的文本特征。2004年后，图算法被引入文本摘要领域，形成了新的研究方向。2004年，Mihalcca和Tarau已经PageRank算法，提出了运用于文本摘要的TextRank算法。目前，基于图算法的比较通用的模型，就是TextRank。

80初，国内学者逐渐注意到文本摘要相关的研究。王永根开始了中文文本摘要技术的研究。90年代末，王开铸提出了基于理解文摘的文本摘要模型。在这之后，中文文本摘要技术的研究开始吸引越来越多的学者涌入。随着2004年图模型引入文本摘要领域，涌现了许多中文摘要解决方案。

主流的文本检索技术大致可分为两种，有监督方法和无监督方法。有监督的文本摘要方法大致有朴素贝叶斯、向量机方法等机器学习方法。无监督的文本摘要方法有着更广泛的应用，TextRank以其简洁高效等特点广受青睐。

**二 相关研究与技术**

这一部分将会介绍本文研究过程中会涉及到的技术与概念，例如关键词提取和文本摘要技术的描述，以及IDF、TextRank和BM25等算法和模型。

（一）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。中文以字为基础，而且单个字往往无法表示具体含义，而是有多个字构成有实际含义的短语。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。在进行中文信息检索时，通常需要先进行分词，其他对于内容、语义等的理解都与分词的质量关联。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。

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如今各种功能详备，性能可靠。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， 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。

（2） 统一的计算理论框架

ICTCLAS根据层叠隐马尔可夫模型构建，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。

（3） 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。

（4） 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。用户可以根据需求，组装最适合自身的分词系统。

（二）去停用词

对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。

表 1 部分停用词表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 停用词表（节选） | | | | | |
| a  an  and  are  as  at  be  but  by  for  if  in  into  is  it  no  not  of  on  or | such  that  the  their  then  there  these  they  this  to  was  will  with  !  "  #  $  %  &  ' | ::  ;  <  =  >  >>  ?  @  A  Lex  [  \  ]  ^  \_  `  exp  sub  sup  | | 一个  一些  一何  一切  一则  一则通过  一天  一定  一方面  一旦  一时  一来  一样  一次  一片  一番  一直  一致  一般  一起 | 上下  上升  上去  上来  上述  上面  下  下列  下去  下来  下面  不  不一  不下  不久  不了  不亦乐乎  不仅  不仅...而且  不仅仅 | 不妨  不定  不对  不少  不尽  不尽然  不巧  不已  不常  不得  不得不  不得了  不得已  不必  不怎么  不怕  不惟  不成  不拘  不择手段 |

（三）关键词提取概要

1. 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。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。

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”。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。

2. 关键词提取

关键词提取也被称为关键词标引，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。

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。计算机实现的关键词提取技术，是指通过计算机技术实现主题概括。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。

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。

3. 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。

（2） 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。

（五）相关算法与模型

1. 基于统计的TF-IDF算法[30]

介绍TF-IDF算法是因为后续基于TextRank的关键词提取和文本摘要设计中，涉及到词频、逆文档频率等特征，同时TF-IDF特征也常用来做相似度比较，所以将此算法作为引入。

TF-IDF算法的核心原理是，词的重要性与它的频数成正比，而且与它在语料库中的频数成反比。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如下公式（1）：

|  |  |
| --- | --- |
|  | 公式（1） |

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。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的计算公式如下公式（2）：

|  |  |
| --- | --- |
|  | 公式（2） |

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（3）：

|  |  |
| --- | --- |
|  | 公式（3） |

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图算法[32]

TexRank算法根据PageRank算法改进而来，是PageRank在文本处理中的应用。PageRank的核心思想是：

1. 如果有很多网页链接到同一网页，那就说明该网页价值高，相应的PR值较大；
2. 如果一个网页被另一个PR值很大的网页链接，也说明该网页重要，PR值较大。

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，该计算公式如下公式（4）：

|  |  |
| --- | --- |
|  | 公式（4） |

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与PageRank算法是一脉相承的，思想也很相似：

1. 如果一个单词出现在其他单词后的频数很高，则说明它很重要，TR值大；
2. 如果一个单词出现在另一个TR值很高的词后面，也说明它很重要，TR值提高。

这样的话，我们可以写出公式（5）：

|  |  |
| --- | --- |
|  | 公式（5） |

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。关于运用到关键词提取和文本摘要中的技术细节，将在后续章节介绍。

3. BM25算法

传统的Text Rank算法运用于文本摘要时，通过重叠词计算句间相似度。这种相似度计算方法缺点明显，无法正确反映文本特征。

BM25算法通常用来做搜索相关性，它的基本思想是：将一个查找语句Query分解为语素集合{qi}，计算每一个语素与搜索文本Doc的相关性，最终将所有语速相关性求和，便得到查找语句Query与待检索文本Doc的相关性。一般公式如下公式（6）：

|  |  |
| --- | --- |
|  | 公式（6） |

其中Q表示搜索语句，D表示待检索文本，qi表示语素（一般是Q的分词结果），Wi是语素qi的权值（一般是qi在Q中词频），R(qi,D)是语素qi与D的相关性（本文选择IDF值作为相关性特征）。

BM25最关键的参数就是R(qi,D)，词与文档的相关性特征。最通用的会选择IDF值作为相关性特征，本文BM25算法中IDF计算公式如下公式（7）：

|  |  |
| --- | --- |
|  | 公式（7） |

其中N为文档总量，n(qi)为含有语素qi的文档数，计算可得逆文档频率IDF值。

而相关性特征R(qi,D)的计算方法仍比较复杂，其一般公式如下公式（8）（9）：

|  |  |
| --- | --- |
|  | 公式（8） |

|  |  |
| --- | --- |
|  | 公式（9） |

其中k1，k2，b都是调节参数，本文中k1设为1.5，b设为0.75。fi是qi在Q中词频，dl为文档D长度，avgdl是文本集平均文本长度。参数b是用于调节文档长度对相关性影响的。

**三 基于TextRank关键词提取技术的设计**

本章节将介绍基于TextRank算法的关键词提取技术的详细设计，包括总体架构、分词技术、去停用词、图算法等。

（一）总体架构图

如下图，从左至右分别是分词与去停用词、构建词汇图模型、TextRank迭代运算等部分，最终输出关键词序列。

待测文本

有序词表

分词与去停用词

划分词窗口

构建图模型

TextRank迭代

排序

关键词

图 1 基于TextRank的关键词提取总体架构图

（二）预处理文本

### 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。

表 2 分词结果展示

|  |
| --- |
| 部分分词结果示例 |
| 据/p 俄罗斯/nsf 卫星网/user 8月/t 11日/t 发布/v 美国/nsf 《/wkz 国家/n 利益/n 》/wky 杂志/n 刊登/v 的/ude1 文章/n 称/v ，/wd 中国/ns 购买/v 俄/b 制/v 苏/b -/wp 27/m 第四/m 代/q 战机/n ，/wd 为/p 本国/rzs 空军/n 翻开/v 了/ule 现代史/n 的/ude1 页/q 章/n 。/wj 从/p 那时/rzt 起/f ，/wd 中国/ns 空军/n 日益/d 强大/a 。/wj 中国/ns 空军/n 长期以来/dl 落后/a 于/p 像/v 美国/nsf 这样/rzv 的/ude1 世界/n 大国/n ，/wd 从/p 2008年/t 起/vf 中国/ns 开始/v 研制/v 堪/vg 舆/n 美国/nsf F/o -/wp 22/m 猛禽/n 战机/n 和/cc F/o -/wp 35/m 闪电/n -/wp II/n 相/d 媲美/vi 的/ude1 第五/m 代/q 战机/n J/n -/wp 20/m 和/cc J/n -/wp 31/m ，/wd 不仅/c 用/p 它们/rr 装备/v 本国/rzs 空军/n ，/wd 而且/c 还/d 在/p 国际/n 市场/n 销售/vn 。/wj 它/rr 让/v 中国/ns 具有/v 了/ule 远程/b 打击/vn 的/ude1 能力/n ，/wd 能/v 达到/v 西/f 太平洋/ns 的/ude1 任何/rz 地点/n 。/wj J/n -/wp 31/m 可能/v 成为/v J/n -/wp 20/m 的/ude1 有力/a 补充/vn ，/wd 是/vshi 理想/a 的/ude1 战机/n ，/wd 能/v 在/p 西/f 太平洋/ns 切断/v 重要/a 地区/n 。/wj J/n -/wp 31/m 升空/vi 后/f ，/wd 完全/ad 能/v 应/v 对/p 美国/nsf 的/ude1 F/o -/wp 35/m 。/wj 这些/rz 战机/n 能/v 从/p 根本/n 上/f 改变/v 中国/ns 同/p 美国/nsf 以及/cc 同/p 台湾/ns 地区/n 冲突/vn 的/ude1 走向/vn 。/wj 如果/c 中国/ns 大陆/n 通过/p 台湾/ns 海峡/n 进攻/v 台湾/ns (/wkz 解放军/n 每年/r 都/d 要/v 进行/vx 这/rzv 方面/n 的/ude1 演练/vn )/wky ，/wd 因为/c 拥有/v 最/d 先进/a 的/ude1 战机/n 而/cc 具有/v 的/ude1 空中/s 优势/n 是/vshi 解放军/n 进攻/v 取胜/vi 的/ude1 关键/n 因素/n 。/wj 这/rzv 无疑/d 应当/v 引起/v 美国/nsf 的/ude1 不安/an ，/wd 无论是/c 从/p 战略/n 上/f ，/wd 还是/c 从/p 战术/n 上/f 以及/cc 从/p 机动性/n 上/f ，/wd 美国/nsf 《/wkz 国家/n 利益/n 》/wky 警告/v 。/wj |

如上表所示分词结果，以空格为间隔，结果会标注词性。例如“战机/n”，名词；“而且/c”，连词；“。/wj”，标点。

### 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。

本文所描述的去停用词过程，会过滤数次、连词、叹词、拟声词、介词、量词等，还会屏蔽停用词表中包含的词语。部分停用词表请参考第二章表1。

### 预处理部分流程图

导入文本

ICTCLAS分词系统

停用词表

去停用词

有序词表

文档

图 2 分词与去停用词流程图

单文本经过预处理后会形成一个有序的词表，上节分词内容，预处理后结果如下：

表 3 预处理结果展示

|  |
| --- |
| 预处理后结果 |
| [俄罗斯, 8月, 11日, 发布, 美国, 国家, 利益, 杂志, 刊登, 文章, 称, 中国, 购买, 制, 战机, 空军, 翻开, 现代史, 章, 中国, 空军, 强大, 中国, 空军, 落后, 美国, 世界, 大国, 2008年, 中国, 研制, 堪, 舆, 美国, 猛禽, 战机, 闪电, II, 相, 媲美, 战机, J, J, 装备, 空军, 国际, 市场, 销售, 中国, 打击, 能力, 西, 太平洋, 地点, J, J, 补充, 理想, 战机, 西, 太平洋, 切断, 地区, J, 升空, 应, 美国, 战机, 改变, 中国, 美国, 台湾, 地区, 冲突, 走向, 中国, 大陆, 台湾, 海峡, 进攻, 台湾, 解放军, 演练, 拥有, 先进, 战机, 空中, 优势, 解放军, 进攻, 取胜, 关键, 因素, 无疑, 美国, 不安, 战略, 战术, 机动性, 美国, 国家, 利益, 警告] |

很容易看出，已过滤助词、连词和停用词，获得了一个有序表，表元素就是语素。

（三）TextRank图算法

### 创建词汇窗口

TextRank算法是依据PageRank算法建立的，本质是一个图算法。我们设定词汇窗口为5，与词汇q距离5以内的词汇，都会为词汇q提供一个贡献值，贡献值的流通方法参考上一章公式，而词汇TR值的初值依据窗口内词数设定

表 4 词窗口展示

|  |
| --- |
| 部分窗口展示（窗口值为5） |
| 11日=[8月, 俄罗斯, 利益, 发布, 国家, 美国]  2008年=[世界, 中国, 堪, 大国, 研制, 美国, 舆, 落后]  8月=[11日, 俄罗斯, 发布, 国家, 美国]  II=[J, 媲美, 战机, 猛禽, 相, 美国, 闪电]  J=[II, 切断, 升空, 国际, 地区, 地点, 太平洋, 媲美, 市场, 应, 战机, 理想, 相, 空军, 美国, 能力, 补充, 装备, 西]  不安=[关键, 因素, 战术, 战略, 无疑, 机动性, 美国]  世界=[2008年, 中国, 大国, 研制, 空军, 美国, 落后]  中国=[2008年, 世界, 冲突, 刊登, 制, 台湾, 国际, 地区, 堪, 大国, 大陆, 太平洋, 市场, 应, 强大, 战机, 打击, 改变, 文章, 杂志, 海峡, 现代史, 研制, 称, 空军, 章, 美国, 翻开, 能力, 舆, 落后, 西, 购买, 走向, 进攻, 销售]  优势=[先进, 关键, 取胜, 战机, 拥有, 空中, 解放军, 进攻]  俄罗斯=[11日, 8月, 发布, 美国]  先进=[优势, 台湾, 战机, 拥有, 演练, 空中, 解放军]  关键=[不安, 优势, 取胜, 因素, 无疑, 美国, 解放军, 进攻]  冲突=[中国, 台湾, 地区, 大陆, 美国, 走向]  切断=[J, 升空, 地区, 太平洋, 应, 战机, 理想, 西]  刊登=[中国, 利益, 国家, 文章, 杂志, 称, 美国, 购买] |

### 创建图模型

如上表所示，窗口内词都会提供一个贡献值给原词汇，我们据此便可以构建窗口内词汇到原词汇的一条有向边

。。。

图 3 图模型部分示例展示

如上图所示，“11日”收到来自6个词汇的贡献值，并将向右侧6个词汇提供提供贡献值，“11日”窗口内词数为6，所有其TR初值为6。

### 进行图运算

图运算模型建立完成后就是进行迭代运算了。设定迭代运算次数上限为200，最大迭代误差为0.001f。

表 5 TextRank迭代结果展示

|  |
| --- |
| 部分迭代结果展示（top20条） |
| 美国:3.8274937;中国:3.405885;战机:3.3051665;J:1.8571391;空军:1.6860673;进攻:1.407505;台湾:1.3851693;地区:1.3647405;国家:1.34468;解放军:1.2305794;利益:1.226108;西:1.2042682;太平洋:1.2014244;无疑:0.89724874;杂志:0.8872335;因素:0.88611233;文章:0.88280547;关键:0.87925905;猛禽:0.87711215;刊登:0.8757052; |

于是我们便可以选取TR值最大的前10条或20条作为关键词序列了。该实验文本经过了11次迭代得到目标结果。

我们可以取[美国;中国;战机;J;空军;进攻;台湾;地区;国家;解放军;利益;西;太平洋;无疑;杂志;因素;文章;关键;猛禽;刊登;]作为关键词。应该说，还是很有概括性的。

**三 融合BM25的TextRank文本摘要方法与设计**

本章节将介绍基于融合BM25的TextRank文本摘要技术的详细设计，包括文本预处理、句间相似度、图运算等内容。

（一）总体架构图

如下图所示，本文摘要方法包含分局与分词、相似度计算、TextRank图运算以及语句重排序等模块。

待测文本

分句词频集合

分句与分词

建立相似度矩阵

构建图模型

IDF算法

BM25算法

相似度计算模块

TextRank迭代

文本摘要

提取摘要句并重排序

图 4 融合BM25的TextRank文本摘要技术架构图

（二）文本预处理

### 分句

基于TextRank的文本摘要方法是要从原文中选取有价值的语句，构成摘要文本。文本中的单句被视为图运算的节点，任两节点之间都有运算边，边的权值为句间相似度，再迭代运算。所以该文本摘要方法的第一步就是分句，本文以[。？?！!；;]为分句间隔。

表 6 分句结果展示

|  |
| --- |
| 分句结果示例 |
| [据俄罗斯卫星网8月11日发布美国《国家利益》杂志刊登的文章称]、[中国购买俄制苏-27第四代战机]、[为本国空军翻开了现代史的页章]、[从那时起]、[中国空军日益强大]、[中国空军长期以来落后于像美国这样的世界大国]、[从2008年起中国开始研制堪舆美国F-22猛禽战机和F-35闪电-II相媲美的第五代战机J-20和J-31]、[不仅用它们装备本国空军]、[而且还在国际市场销售]、[它让中国具有了远程打击的能力]、[能达到西太平洋的任何地点]、[J-31可能成为J-20的有力补充]、[是理想的战机]、[能在西太平洋切断重要地区]、[J-31升空后]、[完全能应对美国的F-35]、[这些战机能从根本上改变中国同美国以及同台湾地区冲突的走向]、[如果中国大陆通过台湾海峡进攻台湾(解放军每年都要进行这方面的演练)]、[因为拥有最先进的战机而具有的空中优势是解放军进攻取胜的关键因素]、[这无疑应当引起美国的不安]、[无论是从战略上]、[还是从战术上以及从机动性上]、[美国《国家利益》警告] |

### 分词与去停用词

该步骤可参考上一章分词和去停用词步骤。区别在于，本章分词的单元是句子，并且输出结果是词频集合，而不是有序表。

导入句子

ICTCLAS分词系统

停用词表

去停用词

词频集合

文档分句

图 5 文本摘要预处理流程图

（三）相似度计算

上一步骤得到的是分句的词频集合，本节将讨论句间相似度的计算方法。本文选择BM25算法作为相似度计算方法，选择IDF作为词特征。这里需要特别解释IDF，我们将单文本分句，然后统计词qi出现于各句的频数，计算逆文本频率

### 计算IDF词特征

我们将IDF值作为词特征，将每一个分句都视为一个文档，总句数为文档总数，并据此统计IDF值。试验IDF值展示如下：

表 7 部分IDF值展示

|  |
| --- |
| 部分IDF值示例 |
| 11日:2.7080502011022096;2008年:2.7080502011022096;8月:2.7080502011022096;II:2.7080502011022096;J:1.7676619176489945;不安:2.7080502011022096;世界:2.7080502011022096;中国:0.7884573603642702;优势:2.7080502011022096;俄罗斯:2.7080502011022096;先进:2.7080502011022096;关键:2.7080502011022096;冲突:2.7080502011022096;切断:2.7080502011022096;刊登:2.7080502011022096;利益:2.151762203259462;制:2.7080502011022096;升空:2.7080502011022096;发布:2.7080502011022096;取胜:2.7080502011022096; |

### 计算句间BM25值，构建相似度矩阵

根据第二章公式（6） 可计算出各语句之间的相似度，作为图运算边的权值，可以构建一个相似度矩阵。

BM

BM

BM

BM

BM

BM

BM

BM

BM

图 6 TextRank图模型示例

（四）图迭代计算

分句和相似度矩阵就可以进行迭代运算了。该迭代运算以句子与其他语句的总和相似度为初值。同样设定最大迭代次数200，最大误差0.001f。

表 8 图迭代结果展示

|  |
| --- |
| 迭代结果 |
| {0.7999757788281731/ 1.1667595263673693/ 0.7347247099734802/ 0.15000000000000002/ 1.3240888396721087/ 1.5278160917881811/ 1.6435856347999025/ 0.832871145587601/ 0.15000000000000002/ 0.6626669744911712/ 0.8188513849688095/ 0.9111709266857193/ 0.7602282663197968/ 1.0333947086434558/ 0.8609206045338191/ 0.7580023993295545/ 1.765852055397935/ 0.9610864983346381/ 0.7317756694003548/ 0.702013652382172/ 0.15000000000000002/ 0.15000000000000002/ 1.0042151324957542/} |

上述值与分句结果一一对应，选取其中TR值最大的几句序号，便可以得到摘要候选句。

（五）提取摘要句与重排序

图运算迭代后就可以得到句子TR值排序，将需要长度的摘要句提出并恢复原文顺序，拼接后就可以得到摘要文本了。

表 9 摘要结果展示

|  |
| --- |
| 摘要结果 |
| 中国购买俄制苏-27第四代战机。中国空军日益强大。中国空军长期以来落后于像美国这样的世界大国。从2008年起中国开始研制堪舆美国F-22猛禽战机和F-35闪电-II相媲美的第五代战机J-20和J-31。J-31可能成为J-20的有力补充。能在西太平洋切断重要地区。美国《国家利益》警告。 |

在限定摘要长度150字时，得到上表9所示摘要文字，共6句按原文顺序排列。比较好地概括了实验文本的内容。

**三 实验结果与总结**

本章将展示最终实验成果，并评估所取得的成效，对本文研究内容做出总结和展望。

（一）实验结果

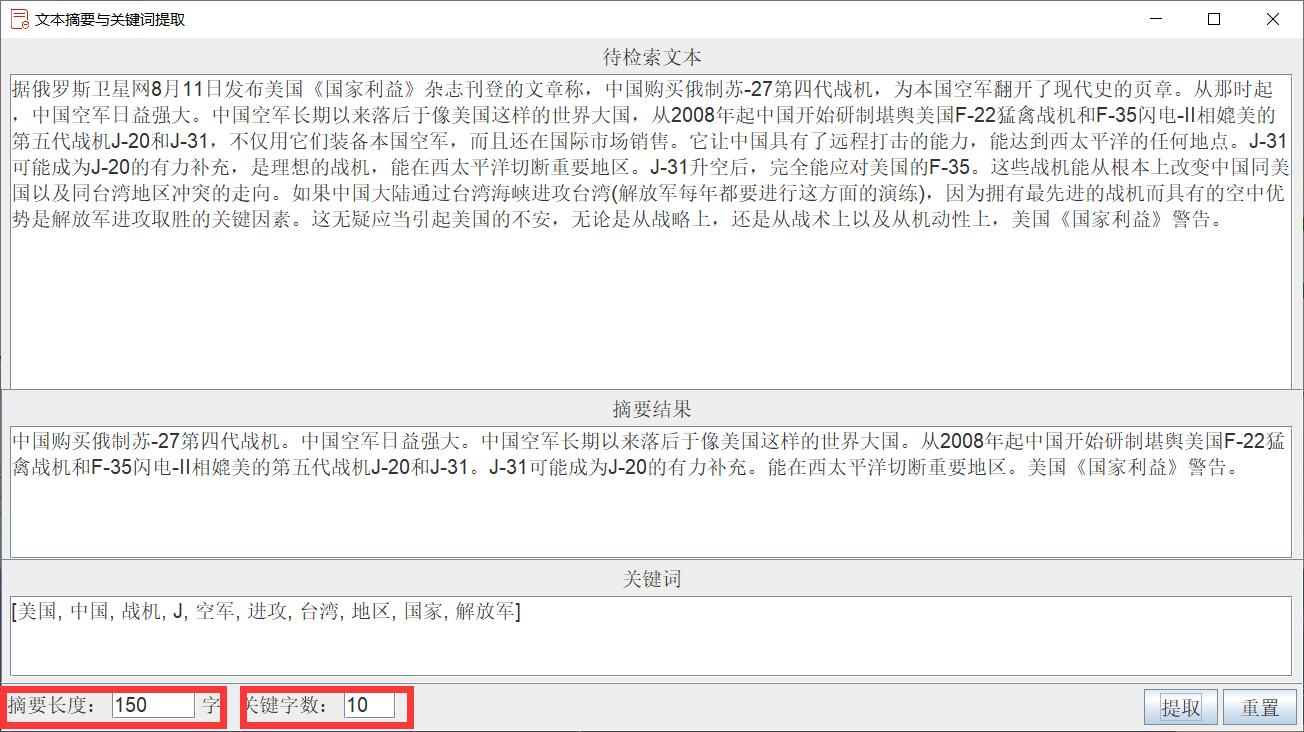
 本文研究内容是基于Java 1.8和IDEA 2020.1环境开发，算法完成后还依据Java swing设计了一个图形界面，展示效果如下：

图 7 应用界面展示

输入文本后，可选摘要长度和关键字数，并最终输出摘要结果和关键词序列。本文基于TextRank算法的两个不同应用，经多次试验，总体达到了较好效果。

在做这个设计作业之前，我曾依据TF-IDF做过一个关于关键词提取的作业设计。但是TF-IDF的算法需要依赖于一个完备语料库训练idf值，而且对于不同类别文本实现效果差异极大，这类方法运行成本很高，实现成果也差强人意。后来，我就尝试能不能完成一个不依赖于语料库质量的关键词提取作业设计。

TextRank算法是有独特优势的，它实现简单，不依赖于外部参数输入，而且能够取得较好效果。尤其是基于TextRank算法的文本摘要设计，相似度算法有多种选择方案。我看现在成熟的方案有建立词袋模型，依据机器学习相关技术求句间相似度。但我最终选择的BM25算法，因为BM25算法原本就是用于评价搜索相关性的方案，实现简单。

（二）总结与展望

事实上，本文所实现的TextRank关键词提取和文本摘要设计是相当粗糙的，如今有很多针对它的改进方案。

1、句间相似度目前比较好的解决方案是神经网络技术，通过训练大规模数据集实现，可以是其具有更好的实用性，使实验结果呈现更稳定。

2、本文关键词提取算法事实上是一种完全的统计学方法，未考虑词性、词长、语义等诸多特征。很多学者都提出了融合多特征的解决方案，可以取得很大进步。

3、本文设计的两个设计都并不适用于微博等短文本，微博短文本关键词的提取和热点分析近年来倍受学者关注。

**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. 李娜娜. 基于TextRank的文本自动摘要研究[D].山东师范大学,2019.
17. 徐馨韬. 基于Doc2Vec和改进的TextRank的中文单文档摘要研究[D].中国电子科技集团公司电子科学研究院,2019.
18. 陈达. 基于TextRank和Word2Vec的短文本自动摘要算法研究[D].武汉理工大学,2018.
19. 李楠,陶宏才.一种新的融合BM25与文本特征的新闻摘要算法[J].成都信息工程大学学报,2018,33(02):113-118.
20. 李志强,潘苏含,戴娟,胡佳佳.一种改进的TextRank关键词提取算法[J/OL].计算机技术与发展, 2020(03): 1-5 [2020-04-19] .http://kns.cnki.net/kcms/detail/61.1450.TP.20191205.110

4.002.html.

1. 门家乐.基于TextRank的关键词提取算法[J].电子世界,2018(15):31-32.
2. 杜海舟,陈政波,钟孔露.基于上下文关系和TextRank算法的关键词提取方法[J].上海电力学院学报,2017,33(06):607-612.
3. 曹洋. 基于TextRank算法的单文档自动文摘研究[D].南京大学,2016.
4. 李鹏，王斌，石志伟，等． Tag-TextＲank: 一种基于 Tag 的网 页关键词抽取方法［J］ ． 计算机研究与发展，2012( 11) : 2 344-2 351．

**附 录**

附录主要列入正文内容过分冗长、或重复性的内容，或者提供查阅方便所需的辅助性表格、计算程序及说明等，如没有相关内容则本部分可以略去。以下是一个附录实例：

附录1 本文中实验文本

据俄罗斯卫星网8月11日发布美国《国家利益》杂志刊登的文章称，中国购买俄制苏-27第四代战机，为本国空军翻开了现代史的页章。从那时起，中国空军日益强大。中国空军长期以来落后于像美国这样的世界大国，从2008年起中国开始研制堪舆美国F-22猛禽战机和F-35闪电-II相媲美的第五代战机J-20和J-31，不仅用它们装备本国空军，而且还在国际市场销售。它让中国具有了远程打击的能力，能达到西太平洋的任何地点。J-31可能成为J-20的有力补充，是理想的战机，能在西太平洋切断重要地区。J-31升空后，完全能应对美国的F-35。这些战机能从根本上改变中国同美国以及同台湾地区冲突的走向。如果中国大陆通过台湾海峡进攻台湾(解放军每年都要进行这方面的演练)，因为拥有最先进的战机而具有的空中优势是解放军进攻取胜的关键因素。这无疑应当引起美国的不安，无论是从战略上，还是从战术上以及从机动性上，美国《国家利益》警告。

**致 谢**

感谢我的指导老师、我的父母、我的同学，以及所有帮助过我的人们，没有他们我是无法我的毕业论文的，谢谢！