|  |  |
| --- | --- |
| **民大校徽** | 计算机应用综合实训项目 |
|  |

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](#_Toc22333)

[Abstract](#_Toc29377)

[目 录](#_Toc906)

[图表目录](#_Toc14495)

[前 言](#_Toc28355)

[一 绪论](#_Toc26910)

[（一）研究背景及研究意义](#_Toc3300)

[1. 研究背景](#_Toc28030)

[2. 研究意义](#_Toc6810)

[（二）研究现状](#_Toc20091)

[二 系统分析](#_Toc18795)

[（一） \*\*\*\*\*\*\*\*\*\*](#_Toc31825)

[1. \*\*\*\*\*](#_Toc556)

[（二）\*\*\*\*\*\*](#_Toc14312)

[（三）\*\*\*\*\*\*\*](#_Toc6142)

[三 \*\*\*\*\*\*\*\*\*\*\*\*\*\*](#_Toc11850)

[（一） \*\*\*\*](#_Toc22641)

[1. \*\*\*\*\*\*](#_Toc8396)

[2. \*\*\*\*\*\*](#_Toc2428)

[（二） \*\*\*\*\*\*\*\*\*\*\*\*\*](#_Toc1946)

[1. \*\*\*\*\*\*\*](#_Toc11365)

[2. \*\*\*\*\*\*\*\*](#_Toc18820)

[四 \*\*\*\*\*\*](#_Toc21720)

[（一）\*\*\*\*\*\*\*](#_Toc15158)

[（二）\*\*\*\*\*](#_Toc7609)

[（三） \*\*\*\*\*\*\*\*\*](#_Toc26956)

[（五）\*\*\*\*](#_Toc20848)

[1. \*\*\*\*\*\*](#_Toc3035)

[2. \*\*\*\*\*\*](#_Toc18888)

[3. \*\*\*\*\*\*](#_Toc14179)

[4. \*\*\*\*\*](#_Toc27546)

[5. \*\*\*\*\*](#_Toc15951)

[五 \*\*\*\*](#_Toc1230)

[（一）\*\*\*\*](#_Toc3946)

[（二）\*\*\*\*\*\*](#_Toc18304)

[（三）\*\*\*\*\*](#_Toc22511)

[（四）\*\*\*\*\*](#_Toc31462)

[1. \*\*\*](#_Toc24810)

[2.\*](#_Toc403)

[3. \*](#_Toc10538)

[4. \*](#_Toc20638)

[5. \*](#_Toc3780)

[（五）\*](#_Toc18750)

[六 总结](#_Toc17345)

[参考文献](#_Toc31830)

# 前 言

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。

# 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。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。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。

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。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。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。

## （二）研究现状

上世纪五十年代，Luhn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。

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。

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。

1999年，Witten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

## 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

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

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。

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。

1. 人工智能法

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

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。

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

3. 未登录词识别

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*。

\*\*\*\*\*\*\*\*\*\*\*\*

## （二）\*\*\*\*\*\*\*\*\*\*\*\*\*

1. \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1. \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# 三 系统设计

本章将说明网站总体结构的设计和数据库的设计。

## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

### \*\*\*\*\*\*\*\*\*\*\*\*\*

本论文系统的描述了\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*，系统流程图如图3-1所示：

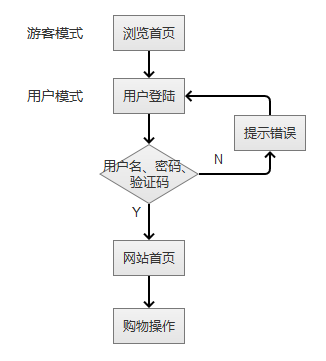


图3-1 系统流程图

### \*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

## \*\*\*\*\*\*\*\*\*\*\*\*\*\*

### 1. 数据库设计原则

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

### 2. 数据库表的设计

下面介绍数据库表的结构

1. 用户表

\*\*\*\*\*表如表3-1所示。

表3-1 用户信息表t\_user

|  |  |  |
| --- | --- | --- |
| 字段 | 说明 | 备注 |
| id | id | 主键，不允许为空 |
| no | 账号 | 不允许为空 |
| pwd | 密码 | 不允许为空 |
| tel | 电话号码 | 不允许为空 |
| description | 个人简介 | 允许为空 |
| role | 角色 | 允许为空 |

# 

# 五 总结

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# 参考文献

1. 朱闯，徐国平，张春龙．校园电子商务的实现[J]. 航空教育研究,2006,7(2):102-103.
2. 马仕海，肇明帅，马刚兴，李根. 校园二手物品交易网建设的研究与实践[J]. 沈阳工程学院学报(自然科学版),2013,(349):03-04.
3. 王志刚. J2EE 在校园二手交易平台中的应用[J]. 湘潭师范学院学报（自然科学版）,2009,(3):011.
4. 卢凯. 交通信号协调控制基础理论与关键技术研究[D]. 华南理工大学,2010:46-48.
5. 徐丽霞. 浅谈电子商务的引用问题. 《中国高新技术企业》,学术期刊,2008年20期.
6. 张磊. 校园二手交易网的设计与实现[D]. 电子科技大学,2013:02-03.
7. 中华人民共和国国家法阵和改革委员会高技术产业司.欧盟电子商务的发展战略分析[R],北京,2007.04.
8. 范亚坤. 校园二手交易网络平台探究.2016.
9. 张亮亮，孟国庆，李瑞英，陈鲁风，杨洪勇. 基于校园网上二手交易平台的研究[J]. 电脑知识与技术,2015,(252):02-16.
10. 李勇军，余雨萍，唐鑫. 大学生二手物品交易平台的设计与实现[J]. 软件与算法,2015,(13):14-17.
11. 刘泽华. 构建大学校园二手物品交易平台可行性分析与策略[J]. 芜湖职业技术学院学报,2015,(93):23-28.
12. Feipeng888. J2EE集成开发环境MyEclipse使用心得.

https://blog.csdn.net/feipeng8848/article/details/69671985 2018年4月17日访问.

1. Katy. 一篇SSH框架的应用介绍.

http://fenghuangwanghui.blog.163.com/blog/static/131807008201245101352761/ 2018年4

月15日访问.

1. Hansion\_xu. Spring是轻量级的J2EE应用程序框架.

https://blog.csdn.net/qq\_22520587/article/details/62454079 2018年4月17日访问

1. Nicolas\_huan. 数据库持久层框架iBatis、myBatis、Hibernate对比

https://blog.csdn.net/nicolas\_huan/article/details/67632852 2018年4月17日访问

1. 吴建勇. 基于java的音乐网站设计与实现[D]. 宿州学院.2011,05-06.